Compact models of multi-scale processes

Rudy Morel

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Compact models of multi-scale processes

Soutenue par
Rudy Morel
Le 29 septembre 2023

École doctorale n°386
Sciences Mathématiques de Paris Centre

Spécialité
Mathématiques Appliquées

Préparée au
Département d'Informatique de l'ENS

Composition du jury :

Stéphane JAFFARD Professeur, Université Paris-Est Créteil	*Rapporteur*
Jean-Luc STARCK Directeur de recherche, CEA	*Rapporteur*
Jean-Philippe BOUCHAUD Directeur de recherche, CFM	*Examinateur*
Shirley HO Professor, Flatiron Institute, CCA	*Examinatrice*
Maarten DE HOOP Professor, Rice University	*Examinateur*
Mathieu ROSENBAUM Professeur, École Polytechnique	*Président du jury*
Stéphane MALLAT Professeur, Collège de France	*Directeur de thèse*

1794
ENS | PSL

Résumé

Les processus multi-échelles, qui présentent des variations sur une large gamme d'échelles, sont présents en physique, finance, biologie, médecine et de nombreux autres domaines. L'objectif principal de cette thèse est de construire des modèles probabilistes de tels processus observés à partir de peu de données et pouvant être échantillonnés numériquement. Ce sujet est crucial pour aborder plusieurs problèmes, notamment la génération, la prédiction et les problèmes inverses tels que la séparation de sources.

Dans cette thèse, nous introduisons les spectres en Scattering («Scattering Spectra»), qui sont basés sur une approximation diagonale de corrélations non linéaires de coefficients d'ondelettes. Ils peuvent être utilisés pour construire des modèles non-Gaussiens de processus multi-échelles, qu'il s'agisse de processus temporels, de processus temporels multi-canaux ou de processus d'image. Nous montrons qu'ils reproduisent des propriétés statistiques importantes de séries temporelles financières, de jets turbulents et de champs physiques.

Nous démontrons que cette représentation en «Scattering Spectra» peut être utilisée pour la séparation de sources à partir de peu de données. Appliquée aux données sismiques sur Mars, ils permettent de séparer avec succès les tremblements de Mars d'événements polluants transitoires appelés «Glitches».

La prédiction sur données limitées peut être abordée en utilisant un modèle précis du processus capable de capturer les dépendances à long terme. Nous introduisons le «Path-Shadowing Monte-Carlo» qui est une méthode à noyau non-locale qui propose de moyenner les quantités futures sur des chemins générés dont l'histoire passée est «proche» de l'histoire réelle (observée). Associée à un modèle basé sur les «Scattering Spectra», cette approche permet d'obtenir des résultats à l'état de l'art pour la prédiction de volatilité en finance et fournit des smiles d'option qui surpassent le marché dans un jeu de trading.

Mots clés : modélisation, multi-échelle, apprentissage non-supervisé, traitement du signal

Abstract

Multi-scale processes, which have variations on a wide-range of scales, are encountered in Physics, Finance, Biology, Medicine and various other fields. The core purpose of this thesis is to construct probabilistic models of such processes observed from limited data and that can be sampled numerically. Such subject is crucial to tackle a number of problems among which generation, prediction and inverse problems such as source separation.

In this thesis we introduce wavelet Scattering Spectra which rely on a diagonal approximation of non-linear correlations of wavelet coefficients. They can be used to construct non-Gaussian models of multi-scale processes, including time-processes, multi-channel time-processes and image processes. Scattering Spectra are shown to capture important statistical properties of financial time-series, turbulent jet and physical fields.

We show that such Scattering Spectra representation can be used to perform source separation on limited data. Applied on Mars seismic data, we are able to successfully separate Marsquakes from transient polluting events called Glitches.

Prediction on limited data can be tackled by utilizing an accurate model of the process which captures long-range dependencies. We introduce Path-Shadowing Monte-Carlo, a non-local kernel method which proposes to predict future quantities by averaging over generated paths whose past history "shadows" the actual (observed) history. When combined with our Scattering Spectra model, this approach yields state-of-the-art volatility prediction in Finance and provides option smiles that outperform the market in a designed trading game.

Keywords : modelling, multi-scale, unsupervised learning, signal processing

Remerciements

Je voudrais remercier en premier lieu Stéphane Mallat, mon encadrant de thèse. Merci de m'avoir appris autant de choses, je ne compte pas le nombre de fois où tu m'as fait découvrir des questions et concepts fascinants. Je resterai à jamais inspiré par ton énergie, ta générosité, ton exigence, ta capacité à débusquer les idées essentielles, et ton humilité. Surtout, merci de m'avoir montré que la recherche est un monde passionnant.

Je voudrais également remercier Jean-Philippe Bouchaud avec qui j'ai eu la chance de travailler étroitement durant cette thèse et qui a énormément influencé mon travail. Ta capacité à proposer des idées inattendues, qui se sont souvent avérées fructueuses, et ta bienveillance sans faille, continueront longtemps à m'inspirer.

Je remercie les membres de mon jury, notamment Stéphane Jaffard et Jean-Luc Starck de m'avoir fait l'honneur d'être rapporteurs de ma thèse. Je leur en suis très reconnaissant.

Je dois beaucoup aux chercheurs que j'ai rencontrés durant ma thèse et qui m'ont montré à quel point la recherche peut être captivante. Je pense à Brice Ménard, et au recul que j'ai su prendre grâce à lui, à Maarten de Hoop, pour la confiance qu'il m'a accordée et les beaux problèmes que j'ai explorés grâce à lui, à Stéphane Jaffard, et son incroyable humilité et générosité, à Erwan Allys, et à ses remarques toujours éclairantes. Merci à Shirley Ho et Michael Eickenberg, pour la facilité d'aborder avec eux des questions prometteuses, j'ai hâte de continuer la discussion. Je remercie tout particulièrement Yves Meyer dont la bienveillance et la capacité à exposer des idées fondamentales m'ont impressionné.

Bien avant d'arriver en thèse j'ai eu la chance de croiser le chemin de professeurs qui m'ont beaucoup fait grandir. À en croire Bourdieu, si certains ont le privilège d'avoir les bonnes stratégies dès le départ, j'ai, moi, eu la chance de rencontrer des professeurs extra-ordinaires. Je remercie tout particulièrement mon professeur de mathématiques au lycée, Luca Spriano, qui m'a montré une voie que je n'aurais pas suivie autrement.

Je voudrais remercier mes amis doctorants qui ont été une famille tout au long de ma thèse. Je pense à mes amis de l'ENS. Merci à Florentin et Gaspar pour leur stimulation, pour les nombreuses discussions, pour tous ces moments si agréables, j'ai été ravi de partager cette aventure avec vous. Merci à Louis, Antoine, John, Étienne, Simon, Tanguy, Samuel, Nathanaël, Roberto. Je ne peux m'empêcher de penser à cette petite aventure qu'a été le Challenge Data et que j'ai organisé avec Tanguy durant deux ans. J'ai aussi eu la chance d'être accueilli dans la chaire EconophysiX, dirigée par Michael Benzaquen que je remercie chaleureusement. Je pense à

Cécilia, Jérôme, Swann, Mehdi, Anirudh, Samy, Pierre, José, Ruben, Léonard, Michele, Fabian, Karl, Salma, Victor, Nirbhay, Antoine (Becharat), Guillaume, Samuel, Natasha, Max, Jutta, Elia, Pierre-Philippe, Armine, Johannes, Riccardo, Théo, Antoine (Fosset).

Je remercie mes amis de classe prépa et de l'ENS Rennes, qui m'ont beaucoup apporté. Je pense à Rémi (G et J), Thibaud, Marie, François, Michel, Corentin, Emeric, Lucien. Merci à Fabien pour ton infinie bienveillance.

Je pense aussi à tous ceux que j'ai croisés durant ma thèse. Merci Maria, quelle joie d'avoir pu te rencontrer au cours de cette aventure, et je suis sûr que l'avenir nous réserve beaucoup de projets passionnants. Merci Ali, de m'avoir beaucoup inspiré tout au long de nos collaborations, j'ai hâte de continuer à travailler avec toi. Merci à Paul, parfait compagnon d'excursion parisienne. Merci à Léa, pour ton énergie impressionnante. Merci Julien pour toutes ces aventures en France et ailleurs.

Je suis accompagné depuis plus de dix ans par mes amis du Sud. Ils m'ont permis de toujours garder la tête froide en m'aérant dans les montagnes à maintes reprises et je les en remercie. Je pense à Iwane, Cécile, Corentin, Vincent, Mathieu, Jeanne.

Merci à ma famille de m'avoir soutenu durant toutes ces années. Surtout, merci à toutes les personnes que j'aurais oubliées de ne pas m'en vouloir.

Contents

Notations

— A stochastic process, or random vector in a discrete setting, x is indexed by u in general $x(u)$. In this thesis we will use the following notations in three specific cases
  — (univariate) $x(t)$ where $t \in \mathbb{R}$ refers to a time index
  — (multivariate) $x(c,t)$ where $c \in \{1,\dots,C\}$ refers to a channel index
  — (multivariate) $x(u_1,\dots,u_d)$ where $u_1,\dots,u_d \in \mathbb{R}$ refer to space variables

  For simplicity we will use the notation x for both the stochastic process (random object) and for a realization e.g. $x \in \mathbb{R}^T, x \in \mathbb{R}^{C\times T}, x \in \mathbb{R}^{L^d}$.

— Given a function $\Phi : \mathbb{R}^N \to \mathbb{R}^M$, instead of writing $\Phi(x)(\lambda)$ for the coordinate λ of Φ applied to x We will utilize brackets when it is convenient $\Phi(x)[\lambda]$.

— For $z \in \mathbb{C}$ we write $|z|$ its modulus and $\varphi(z)$ its argument : $z = |z|e^{i\varphi(z)}$.

— For $z \in \mathbb{C}$ we write z^* its complex conjugate.

— For $x \in \mathbb{C}^{n\times m}$ identified as a matrix we write $x^* \in \mathbb{C}^{m\times n}$ its conjugate transpose.

— For $x \in \mathbb{C}^n$ we note :
  — for $p \geq 1$, $\|x\|_p = (\sum_{k=1}^n |x_k|^p)^{\frac{1}{p}}$ the ℓ^p-norm.
  — $\|x\|_0 = \mathrm{Card}\{1 \leq k \leq n \mid x_k \neq 0\}$ the 0-pseudonorm.

— For $x \in \mathbb{C}^n$ we write $\langle x_k \rangle_k = \frac{1}{n}\sum_{k=1}^n x_k$ the average.

— For $x, y \in \mathbb{C}^n$ we write $\langle x, y\rangle = \sum_{k=1}^n x_k y_k^*$ the Hermitian inner product.

— We write $\mathbb{1}_A$ the indicator function of a set A, $\mathbb{1}_A(x) = 1$ if $x \in A$, $\mathbb{1}_A(x) = 0$ if $x \notin A$.

— For $p > 0$ we write $\mathbf{L}^p(\mathbb{R}^d)$ the space of functions $f : \mathbb{R}^d \to \mathbb{C}$ verifying

$$\|f\|_p \coloneqq \Big(\int_{\mathbb{R}^d} |f(u)|^p \,\mathrm{d}u\Big)^{\frac{1}{p}} < +\infty$$

  where integration is performed with respect to the Lebesgue measure on $\mathbb{R}^d$.

— For two functions $f, g : \mathbb{R}^d \to \mathbb{C}$, the convolution of f, g, is written

$$f \star g(u) = \int_{\mathbb{R}^d} f(v)g(u-v)\,\mathrm{d}v$$

  and is well defined for example when $f \in \mathbf{L}^p(\mathbb{R}^d), g \in \mathbf{L}^\infty(\mathbb{R}^d)$, $1 \leq p \leq +\infty$. We use the same notation for its discrete analogous $x \star y(u) = \sum_{v=1}^n x(v)y(u-v)$ for $x, y \in \mathbb{C}^n$ where the indices are considered modulo n.

— The Fourier transform of $f \in \mathbf{L}^1(\mathbb{R}^d)$ is noted $\widehat{f}$ and written

$$\widehat{f}(\omega) = \int_{\mathbb{R}^d} f(u) e^{-i\langle \omega, u \rangle} \, \mathrm{d}u.$$

The Fourier transform is uniquely extended by density on functions $f \in \mathbf{L}^2(\mathbb{R}^d)$. The discrete Fourier transform of a vector $x = (x_0, \ldots, x_{n-1}) \in \mathbb{C}^n$ is written

$$\widehat{x}(\omega) = \sum_{\ell=0}^{n-1} x_\ell e^{-i\frac{2\pi}{n}\ell\omega}$$

— For two sequences $u, v \in \mathbb{R}^{\mathbb{N}}$ we write $u = o(v)$ if $\forall \epsilon > 0, \exists n_0, \forall n \geq n_0, |u_n| \leq \epsilon |v_n|$. We write $u = \mathcal{O}(v)$ if $\exists C, \exists n_0, \forall n \geq n_0, |u_n| \leq C|v_n|$.

— For a random vector $x \in \mathbb{R}^N$ with probability distribution $p : \mathbb{R}^N \to \mathbb{R}$ and $\Phi : \mathbb{R}^N \to \mathbb{C}$ we write

$$\mathbb{E}_p\{\Phi(x)\} = \int_{\mathbb{R}^N} \Phi(x) p(x) \, \mathrm{d}x$$

provided that $\mathbb{E}_p\{|\Phi(x)|\} < +\infty$. We will omit the notation p when there is no ambiguity on the probability distribution : $\mathbb{E}\{\Phi(x)\}$.
For a random stationary vector $x(t)$ we will often make use of the following property

$$\mathbb{E}\{\langle \Phi(x(t)) \rangle_t\} = \mathbb{E}\{\Phi(x(t))\}$$

the right-hand term being independent on t.

— For a random vector $x \in \mathbb{C}^n$ and a random variable $y \in \mathbb{C}$ we write $\mathbb{E}\{x|y\}$ the conditional expectation of x given y.

Chapter 1

Introduction

Processes encountered in many fields are multi-scale i.e. they have variations on a wide range of scales. For example, this is the case of a financial price time-series, or a seismic recording. In Physics, this is the case of many two or three dimensional observations, such as the velocity of a turbulent fluid, or the large-scale distribution of dark matter in the Universe.

The core purpose of this thesis is to build probabilistic models of multi-scale processes, from limited observations, and that can be sampled numerically. This is an unsupervised learning problem that can be formulated as building a distribution p_θ, that can be sampled numerically, and that approximates the distribution p of the underlying process x, observed from a single realization $\widetilde{x} \in \mathbb{R}^N$ of limited size, where $\theta \in \mathbb{R}^M$ are parameters with M being the dimension of the model. The process x that we observe is assumed to be stationary, or with stationary increments, and ergodic.

This subject is crucial to tackle a number of problems formulated in a limited data regime, such as generation (drawing new realizations of x), but can also be used for prediction (determining an unknown set of values of x), in particular when there is no access to enough labeled data to train a supervised algorithm. It is also crucial to tackle inverse problems such as unsupervised source separation (unraveling an observed mixture of source signals) that can be made well-posed if we assume a prior model on the sources.

The data constraint is a strong restriction that emerges from fundamental principles. We often possess a single realization of the process x under study. In Finance, there exists a single realization of the price process of a certain index: the historical realization. In Astrophysics, we observe a single map of the Universe. The process x may contain long-range dependencies at the scale of the realization size, meaning that $x(u)$ and $x(u')$ are still dependent for the largest values of $u-u'$. Such dependencies are hard to estimate precisely because the number of samples of the joint distribution $(x(u), x(u'))$, from a single realization, is very small.

The main challenge in building models of multi-scale processes from limited data resides in a bias-variance tradeoff that we explain below.

Model bias. Processes of interest in many domains are often non-Gaussian. For example, this is evidenced by intermittency or time-asymmetry in certain time processes, or by the presence

of transient structures in a two-dimensional physical field. In particular, well understood Gaussian models, which are characterized by their average and covariance, fail to capture essential properties of the process. These models are biased because they rely on a poor description of the process under study.

Model variance. One way of improving the accuracy of a parameterized model p_θ in approximating the underlying distribution p is to increase the number of parameters $\theta \in \mathbb{R}^M$, so as to reproduce an enriched set of statistics $\Phi(x)$. Calibrating the model then consists in finding a $\theta \in \mathbb{R}^M$ from limited data $\widetilde{x}$ such that p_θ reproduces these statistics. However, enlarging the vector of statistics Φ increases the variance of $\Phi(x)$ which makes its estimation on limited data $\widetilde{x}$ harder. This means that the model p_θ of the same process p may differ significantly when estimated from one realization $\widetilde{x}$ to the other.

Thus the number of parameters must be chosen carefully. In this thesis, we call *compact model* a model p_θ where the number of parameters M, referred as the model dimension, grows as $o(N)$ in the size N of a single realization $\widetilde{x}$.

One of the challenge in this thesis is to leverage the multi-scale nature of observed data so as to navigate this bias-variance tradeoff by defining a prior on the underlying distribution p. In order to highlight the main contributions of the thesis, we present some key concepts and tools in the literature to construct models of multi-scale processes.

1.1 Maximum entropy models

1.1.1 Promoting model diversity

In his seminal paper Jaynes [Jaynes, 1957] proposes to set up a model of p from partial observations by maximizing its entropy. The entropy of a process distribution p is given by $H(p) = -\int p(x) \log p(x) dx$. A macrocanonical model p_θ of process x can be defined as a maximum entropy distribution conditioned by the exact value of a vector of moments $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \mu$ where $\Phi : \mathbb{R}^N \mapsto \mathbb{R}^M$ and one can chose $\mu = \mathbb{E}_p\{\Phi(x)\}$ for now. If they exist, they have an exponential probability distribution

$$p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \theta, \Phi(x) \rangle}, \tag{1.1}$$

for a given $\theta \in \mathbb{R}^M$, where M the number of parameters is also the number of statistics Φ. Such model is the least biased, given moment constraint $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \mu$ in the sense that it is maximally noncommittal with regard to missing information [Jaynes, 1957]. For example, a stationary Gaussian process p is a maximum entropy model conditioned by first and second order moments $\mathbb{E}_p\{\Phi(x)\}$, with $\Phi(x) = \left(\langle x(t) \rangle_t \,, \langle x(t-\tau) x(t) \rangle_t \right)$.

The parameters θ of the model (1.14) can be estimated through Markov Chain Monte-Carlo methods [Lustig, 1998; Betancourt, 2017] which yield an exact, but computationally expensive algorithm when the number of statistics Φ is large, due to the mixing time of the Markov Chain [Levin, 2017; Bruna, 2019].

In this thesis, to avoid this computational issue, we consider microcanonical maximum en-

tropy models which have a maximum entropy distribution on a set

$$\Omega_\epsilon = \{x \in \mathbb{R}^N \mid \|\Phi(x) - \Phi(\widetilde{x})\|_2 \leq \epsilon\}.$$

for a certain error ϵ which is adjusted with the variance of $\Phi(x)$. We usually define Φ so that Ω_ϵ is a compact set of strictly positive Lebesgue measure $\int_{\Omega_\epsilon} dx$, in this case, the microcanonical model has a uniform distribution on Ω_ϵ.

If $\Phi(x)$ concentrates around $\mathbb{E}\{\Phi(x)\}$ then the microcanonical model converges to the macrocanonical model (1.14) when the size N of $\widetilde{x}$ goes to ∞ and ϵ goes to 0. This is the Boltzmann equivalence principle [Lanford, 1975; Gallagher, 2013]. The concentration of $\Phi(x)$ generally imposes that its dimension M is small relatively to the dimension N of x.

Sampling a microcanonical model can be performed through a gradient descent on the loss $x \mapsto \|\Phi(x) - \Phi(\widetilde{x})\|^2$ from an initial realization of a Gaussian noise, which has a maximum entropy distribution [Bruna, 2019].

Typical model failure can be formulated as a drop in the entropy of the process, for example when the model concentrates around a single realization. Given moment constrain $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \mu$ in a macrocanonical model or statistical constraint $\|\Phi(x) - \Phi(\widetilde{x})\|_2 \leq \epsilon$ in a microcanonical model, the entropy maximization is a way of promoting diversity of the realizations i.e. increasing the volume of sets of high probability [Shannon, 1948].

1.1.2 Revisiting the bias-variance tradeoff

The bias-variance tradeoff in building a model of p is made explicit in a maximum entropy model. A maximum entropy model is maximally uncommittal with regard to missing information [Jaynes, 1957], thus a too small number of statistics $\Phi(x)$ may fail to characterize important properties of process x leading to a large model bias.

On the other hand, with no access to p, the moments $\mu = \mathbb{E}_p\{\Phi(x)\}$ need to be estimated by $\Phi(\widetilde{x})$ on a single realization $\widetilde{x}$ of limited size. For the model to be accurate, one needs $\Phi(\widetilde{x})$ to be close to $\mathbb{E}\{\Phi(x)\}$ which can be ensured by choosing low-variance statistics $\Phi(x)$, thus constraining the model variance.

The main challenge in a maximum entropy model is to define statistics Φ which specify important properties of x, so as to yield an accurate model of p, while remaining of low-variance so that $\Phi(\widetilde{x})$ is a good estimation of $\mathbb{E}_p\{\Phi(x)\}$.

High-order moments are an example of candidate where we consider Φ to be the average over time of polynomials in the coordinates of x. For $r \in \mathbb{N}^*$, provided $\mathbb{E}\{|x(u)|^r\} < +\infty$ the moments $\mathbb{E}\{\Phi(x)\}$ read

$$\mathbb{E}\{x(u_1) \dots x(u_r)\} \tag{1.2}$$

Under certain conditions [Billingsley, 2013] the infinite expansion $r \in \mathbb{N}$ provides an exact description of the process distribution. However, they are difficult to estimate from limited data. Indeed, high-order polynomials amplify large events which result in a large variance of estimation. This problem is typically amplified for processes with fat-tailed distributions. Next section focuses on candidates for Φ from the literature.

1.2 Scale dependencies

Processes of interest in real-world data are often non-Gaussian. Designing moments $\mathbb{E}\{\Phi(x)\}$ that characterize non-Gaussian properties of the process and can be estimated on limited data $\widetilde{x}$ through $\Phi(\widetilde{x})$, has been an important research subject. We focus in this section on univariate time processes and will study multivariate extensions in section 1.11.

Separating scales in a multi-scale process x can be done with a wavelet transform, that is presented in section 1.9.1. We show that structure functions can be used to track the evolution of the process distribution at different scales through high-order moments.

In section 1.9.2 we present scattering networks which track the distribution of wavelet coefficients by rather cascading wavelet operators and modulus nonlinearity, inspired by the use of convolutional neural networks in Machine Learning.

In section 1.9.3 we present a different approach, still relying on the wavelet coefficients, but that now looks for scale dependencies in the joint distribution of wavelet coefficients at different times and scales.

1.2.1 Wavelet transform and structure functions

A wavelet transform separates variations at multiple scales. It is computed with a zero-average $\int \psi(t)\mathrm{d}t = 0$ complex filter ψ which is localized both in the time domain and in the Fourier domain [Y Meyer, 1992; Mallat, 1999]. The wavelet transform operator W is then

$$Wx(t,j) = x \star \psi_j(t) \quad \text{where} \quad \psi_j(t) = 2^{-j}\psi(2^{-j}t).$$

More specifically we can chose a wavelet ψ whose Fourier transform $\widehat{\psi}(\omega) = \int \psi(v)e^{-i\omega t}\mathrm{d}t$ is mostly concentrated at frequencies $\omega \in [\pi, 2\pi]$. It results that $\widehat{\psi}_j(\omega) = \widehat{\psi}(2^j\omega)$ is non-negligible mostly in $\omega \in [2^{-j}\omega, 2^{-j+1}\omega]$. This provides a separation of the frequency axis into different bins that constitutes our notion of scales. For a stationary process x, or with stationary increments, the joint process $Wx(t,j)$ is stationary, under certain condition on the wavelet filter [Pipiras, 2017].

Structure functions track the evolution of the distribution of wavelet coefficients at different scales through its high-order moments

$$S(q,j) = \mathbb{E}\{|x \star \psi_j(t)|^q\}. \tag{1.3}$$

While $S(2,j)$ does not contain more information than a Gaussian model can capture, the $S(q,j)$ for $q \neq 2$ may differ from Gaussian statistics.

For multi-scale processes, self-similarity refers to scale invariance properties of the distribution of the process. Definition of self-similarity will be discussed in the following. At the level of structure functions, self-similarity is characterized by a power-law on the range of scales under study

$$S(q,j) = c_q\, 2^{j\zeta_q}. \tag{1.4}$$

Early works in multifractal analysis made use of these exponents ζ_q to determine the singularity spectrum of a signal x which characterizes the variability of pointwise Hölder exponents of x [Bacry, 1993; Muzy, 1994; Jaffard, 2004]. Estimation issues, inherent to high-order moments, can be addressed by introducing modulus powers of wavelet coefficients maxima that are called wavelet leaders [Jaffard, 2006; Wendt, 2009]. These wavelet leaders offer the advantage of not requiring strong stationarity or ergodicity assumptions and can be estimated for both real positive and negative exponents. These multiscale quantities and their scaling have been successfully used to detect and discriminate properties of non-Gaussian processes, such as intermittency, with applications to Medicine for example [Abry, 2010; Saës, 2022].

Building models of multi-scale processes from moments (1.16), or even from recent multiscale quantities used in multifractal analysis, raises an important issue. Such moments do not pick-up important non-Gaussian properties such as time-asymmetry, changing $x \star \psi_j(u)$ into $x \star \psi_j(-u)$ leaves (1.16) unaffected, which is crucial to build accurate models of time-processes.

This issue also affects the fundamental question of defining self-similarity. It admits a strong definition that states that the joint distribution of wavelet coefficients is invariant to dilation, up to random multiplicative factors [Mandelbrot, 1997]. However, as a definition in distribution, it cannot be tested numerically on a single realization. Structure functions provide a numerically tractable (at least for small exponents) definition. However, as mentioned above, it provides a weak description of the process distribution.

An important challenge is to find a notion of self-similarity based on a richer description of the process and that can still be tested numerically on a single realization. This problem is tackled in section 1.10.2.

1.2.2 Scattering transform

Instead of considering high-order statistics, a scattering transform Sx proposes to analyze the time structure of wavelet coefficients $Wx(t, j)$ at a fixed scale 2^j through a cascade of wavelet transforms and modulus non-linearities [Mallat, 2012; Bruna, 2013]. Defined up to a largest scale 2^J, it concatenates scattering coefficients S_m of different orders $0 \leq m \leq J$

$$S_m x(t, j_1, \ldots, j_m) = |\ldots|x \star \psi_{j_1}| \star \psi_{j_2}| \ldots | \star \psi_{j_m}(t) \tag{1.5}$$

for $1 \leq j_1 < \ldots < j_m \leq J$. Scattering moments are estimated through an empirical average $\Phi(x) = \langle S_m x(t, j_1, \ldots, j_m) \rangle_t$.

Unlike structure functions, scattering moments are 1-Lipschitz in x. These coefficients can also be used to analyze intermittency in multi-scale processes [Bruna, 2015]. They can be used also in audio classification [Andén, 2018] or seismic event detection and clustering [Seydoux, 2020; Rodrıguez, 2021]. However, similarly to structure functions, they do not provide a rich enough statistical description of the process to yield accurate models, in particular they don't pick up time-asymmetry.

Scattering coefficients, as well as structure functions, delete the phase of wavelet coefficients through a modulus, in order to obtain non-zero coefficients after time-average. Time-asymmetry

can be picked up by the phase of wavelet coefficients. Indeed, in the case of an analytical wavelet ($\widehat{\psi}$ is real), the filter $\mathrm{Im}\,\psi$ is an odd function and the sign of $x \star \mathrm{Im}\,\psi(t)$ can detect asymmetry. An important question is to retrieve these phase dependencies in order to capture non-Gaussian properties of x.

1.2.3 Non-linear correlations of wavelet coefficients

Another approach to build a non-Gaussian representation of time-processes is to consider moments $\mathbb{E}\{\Phi(x)\}$ in the form of correlations on a 1-Lipshitz representation $\mathcal{R}$

$$\mathbb{E}\{\mathcal{R}x(t,\lambda)\,\mathcal{R}x(t',\lambda')^*\} \tag{1.6}$$

where t, t' are time indices and λ, λ' are indices of the representation $\mathcal{R}$. For example, Machine Learning literature provides representations $\mathcal{R}$ in the form of cascades of linear convolutional operators and pointwise non-linearity that are called convolutional neural networks [Gatys, 2015; Ustyuzhaninov, 2017]. However, this leads to a lot of correlation features M, much larger than the size of the data N, with the risk of having large model variance. Besides, the interpretation of the coefficients is difficult.

Dependencies across separate scales $2^j \neq 2^{j'}$ were shown to be crucial to characterize the distribution of a multi-scale process, in particular non-Gaussian properties. For example, the presence of a burst in a time-series or structures in an image gives rise to large coefficients around this location [Portilla, 2000]. Authors in [Gatys, 2015; Ustyuzhaninov, 2017] actually show that correlating feature maps obtained with filters of different size is key to obtain the best perceptual results of texture syntheses.

Setting $\mathcal{R} = W$ in moments (1.19) builds a linear model in the wavelet coefficients of x. As such, this is a Gaussian model which is not an accurate model of many processes of interest. Such failure is explained by the fact that correlation of wavelet coefficients do not capture scale dependencies. Indeed, for processes x with a regular power-spectrum, wavelet correlation

$$\mathbb{E}\{x \star \psi_j(t)\, x \star \psi_{j'}(t')^*\}$$

has a fast decay away from $t = t'$ and $j = j'$ [Wornell, 1993]. Indeed, for separate scales $2^j \neq 2^{j'}$, the frequency supports of $x \star \psi_j(t)$ and $x \star \psi_{j'}(t')$ barely overlap, due to the wavelet scale separation. The two processes $x \star \psi_j(t)$ and $x \star \psi_{j'}(t')$ thus oscillate at different frequencies and their correlation is canceled by phase oscillations. An important question is thus to retrieve scale dependencies through non-linear correlations.

In the context of texture generation, authors in [Portilla, 2000] propose to capture joint time-scale dependencies through the correlation of wavelet coefficients and their modulus. This amounts to consider the representation $\mathcal{R} = \rho W$ in (1.19) where $\rho(z) = (z, |z|)$ and $\rho W x(t,j) = (x \star \psi_j(t), |x \star \psi_j(t)|)$. We write $C_{\rho W}(t,t',j,j')$ the resulting correlation matrix that contains the three correlation matrices

$$\mathbb{E}\{Wx\, Wx\}\ ,\ \mathbb{E}\{Wx\, |Wx|^T\}\ ,\ \mathbb{E}\{|Wx|\, |Wx|^T\}. \tag{1.7}$$

Taking a modulus prevents the phase cancellation effect mentioned above. Indeed, a modulus eliminates the phase of coefficients $x \star \psi_{j'}$ responsible for their oscillation. The process $|x \star \psi_{j'}|$ has a frequency support around $\omega = 0$ which now overlaps with the frequency support of $x \star \psi_j$ for $j > j'$.

More generally, phase harmonics go a step further in the analysis of the phase of wavelet coefficients [Leonarduzzi, 2019; Mallat, 2020]. The phase harmonic $[z]^k$ of a complex number $z \in \mathbb{C}$ are defined by multiplying its phase $\phi(z)$ by an integer $k \in \mathbb{N}$ while keeping its modulus unchanged $[z]^k = |z|^k e^{ik\phi(z)}$ where $z = |z|e^{i\varphi(z)}$. Phase harmonic correlations are obtained from (1.19) by setting $\mathcal{R} = \rho W$ with $\rho(z) = ([z]^0, [z]^1, [z]^2, \dots)$ being an extension of the previous phase-modulus operator that was restricted to $k = 0$ and $k = 1$. It yields the correlations

$$\mathbb{E}\{[x \star \psi_j(t)]^k [x \star \psi_{j'}(t')]^{k'*}\}. \tag{1.8}$$

For $k \geq 1$ the phase-harmonic $[\cdot]^k$ accelerates the oscillation of wavelet coefficients $x \star \psi_j(t)$ and can thus be used to realign Fourier supports of wavelet coefficients $x \star \psi_j(t)$ and $x \star \psi_{j'}(t')$ so as to prevent the above mentioned phase cancellation effect. This enables capturing dependencies across scales $2^j \neq 2^{j'}$.

Now that such models consider non-linear correlation across times and scales, the number of coefficients becomes larger than the size of a single realization. As shown in [Brochard, 2022] there is an important risk of reconstructing part of the observed realization, because the estimation of the moments from limited data is too difficult. Phase-harmonic correlations 1.8, estimated through time-average, showcase an imbalanced bias-variance tradeoff towards large variance representations Φ. Compared to structure functions (1.16), they do not do not leverage any self-similarity property of the field.

An important challenge is to characterize scale dependencies in x from a reduced set of coefficients by leveraging the scale regularity of the process.

1.3 Scattering Spectra

Building up on the previous works we reviewed in last sections, chapter 2 introduces a correlation-representation called Scattering Spectra, it is a compact representation of scale dependencies that can be estimated on limited data.

For that we further exploit the multi-scale prior on process x in two different ways that are exposed in the following sections.

1.3.1 Diagonal scattering covariance

We start from the phase-modulus correlations (1.7). The time structure of the envelopes $|Wx|$ is captured through correlations across all t, t'. However, such envelopes generally have long-range dependencies with a regular cross-spectrum and we know that such process covariance can be compressed through wavelet transform [Wornell, 1993]. Cascading a second wavelet transform yields a scattering transform $Sx = W|Wx|$ with $Sx(t, j_1, j_2) = |x \star \psi_{j_1}| \star \psi_{j_2}(t)$

(see section 1.9.2). The auto-correlation of scattering transform coefficients is $\mathbb{E}\{Sx\, Sx^T\} = W\, \mathbb{E}\{|Wx|\, |Wx|^T\}\, W^T$. This matrix considers correlations of scattering coefficients across separate channels and are an extension of the standard scattering coefficients (1.18) of order $m = 2$. For processes with a regular envelope cross-spectra, such as the ones encountered across Finance or Physics, owing to wavelet correlation compression properties [Wornell, 1993], the matrix $\mathbb{E}\{Sx(t, j_1, j_2)\, Sx(t', j_1', j_2')\}$ has a sparse structure and is concentrated along its diagonal $t = t', j_2' = j_2$. We write Diag such diagonal projection.

One of the main contribution of chapter 2 is to introduce the Scattering Spectra which are a diagonal approximation of the non-linear correlations (1.7)

$$\Big(\mathrm{Diag}\,\mathbb{E}\{Wx, Wx^T\}\ ,\ \mathrm{Diag}\,\mathbb{E}\{Wx, |Wx|^T\}\ ,\ \mathrm{Diag}\,\mathbb{E}\{W|Wx|, W|Wx|^T\}\Big). \tag{1.9}$$

They extend the standard wavelet power spectrum $\mathrm{Diag}\,\mathbb{E}\{Wx, Wx^T\}$. Sign-asymmetry, often called skewness, is picked up by the second correlation matrix $\mathrm{Diag}\,\mathbb{E}\{Wx, |Wx|^T\}$. Intermittency in the process is characterized by the third matrix $\mathrm{Diag}\,\mathbb{E}\{W|Wx|, W|Wx|^T\}$. These are complex coefficients and we prove that their imaginary part captures time-asymmetry.

They are estimated by replacing $\mathbb{E}$ by an average over time $\langle . \rangle_t$. For a realization $x \in \mathbb{R}^N$ of size $N = T$ the number of time-steps, the Scattering Spectra $\Phi(\widetilde{x})$ consist of $\mathcal{O}(\log_2^3 T)$ order 2 coefficients, much lower than T, and can thus be estimated on limited data.

We show that they provide accurate models of Financial price time-series and univariate turbulent jet, and capture main non-Gaussian properties such as fat-tails distributions, intermittency, sign-asymmetry and time-asymmetry. Interestingly, a model based on such order 2 moments is shown to reproduce higher order statistics of order up to 5.

1.3.2 Wide-sense self-similarity

As explained in section 1.9.3, the non-linear correlations (1.7) characterize scale dependencies and capture non-Gaussian properties such as sign-asymmetry and time-asymmetry that where not captured by structure functions (1.16).

In chapter 2 we prove that the strong definition of self-similarity defined on the distribution of the process [Mandelbrot, 1997] implies a scaling invariance of the matrices (1.7) up to normalization factors. This definition is said to be wide-sense as an analogy with the wide-sense time-stationarity.

Wide-sense definition. Commonly used in signal processing, wide-sense time-stationarity considers the correlation matrix across time $C(t, \tau) = \mathbb{E}\{x(t)\, x(t + \tau)\}$. Let us assume a x has a zero-mean $\mathbb{E}\{x(t)\} = 0$. Process x is said to be wide-sense time-stationary if the correlation $C(t, \tau)$ is independent on t. This thus states that matrix C is invariant to time-shift. The phase-modulus correlation matrix $C_{\rho W}(t, t', j, j')$ characterizes time-scale dependencies through the three matrices $\mathbb{E}\{Wx\, Wx\}, \mathbb{E}\{Wx\, |Wx|\}, \mathbb{E}\{|Wx|\, |Wx|\}$ (1.7). Let us reindex this matrix as $C_{\rho W}(t, \tau, j, a)$, with $\tau = t - t', a = j - j'$. Due to time-stationarity, this correlation does not depend on t. Wide-sense self-similarity shows that the same property holds for the log-scales

j, j'. In order to assert that, we need to account for the fact that the wavelet coefficients variance, the power-spectrum, may not be constant across scales. We thus normalize the correlation $C_{\rho W}$ by the wavelet power spectrum of x. Chapter 2 shows that for a self-similar process (in the strong sense), the normalized phase-modulus matrix $C_{\rho W}(t, \tau, j, a)$ does not depend on j, it is invariant to scale shift

$$C_{\rho W}(t, \tau, j, a) = C_{\rho W}(0, \tau, 0, a). \tag{1.10}$$

This is the wide-sense self-similarity definition introduced in this thesis. Fig. 1.1 (right) illustrates this invariance to scale shift on the matrix $\mathbb{E}\{|Wx|\,|Wx|\}$.

This definition relies on a set of statistics that characterize important non-Gaussianity such as intermittency, sign-asymmetry and time-asymmetry and that can be used to build accurate models of p, through their Scattering Spectra reduction (1.20). Under wide-sense self-similarity, the scattering Spectra (1.20), which compress the non-linear correlations (1.7), are proved to be invariant to scale shift, up to renormalization,

Numerically, one can show that imposing the scale invariant Scattering Spectra recovers the power-law scaling of structure functions (1.16) up to order 4.

Scale regularity. While self-similarity seems to be satisfied for a financial price process from the scale of a few minutes to the scale of a decade, it is not satisfied in general e.g. for turbulent flows. The dilation of the process x acts on the matrix $C_{\rho W}(j, j' - j)$ as a translation of its first variable j. Scale regularity can be defined as the regularity of $C_{\rho W}$ as a function of j. In chapter 3 we compute a Fourier transform along j which yields $\widehat{C}_{\rho W}(\omega, j' - j)$. For a process with scale regularity the coefficients $\widehat{C}_{\rho W}(\omega, j' - j)$ have a fast decay away from $\omega = 0$. This is used to provide a model with adaptive number of coefficients by thresholding Fourier harmonics. This enables to build further reduced models of processes with scale regularity and helps reducing the model variance.

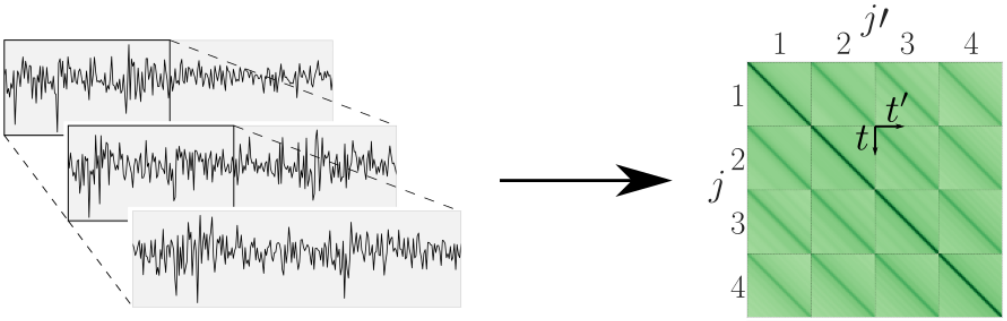


Figure 1.1 – (Left) different zooms in a financial log-return time-series illustrating self-similarity. (Right) Wide-sense self-similarity definition states that the normalized phase-modulus correlation matrix $C_{\rho W}(t, t', j, j')$ across times t, t' and scales j, j' depends only on $j - j'$. It provides a rich definition of self-similarity that can be tested on limited data.

1.4 Multivariate processes

Multi-scale processes $x(u)$ encountered in many domains are often multivariate in the sense that they are indexed by a collection of variables $u = (u_1, \dots, u_d)$ each belonging to a certain space. We will consider two types of multivariate processes that provide two different extensions of univariate processes.

The first type considers $u = (u_1, \dots, u_d) \in \mathbb{R}^d$ belonging to a d-dimensional lattice. Particular processes of interest include 2-dimensional, e.g. fracture surfaces [Lakhal, 2023], or 3-dimensional fields in Physics, e.g. dark matter fields [Villaescusa-Navarro, 2020]. In this case, the space $\mathbb{R}^d$ is naturally equipped with the Euclidean norm and the signal processing tools such as Fourier basis and wavelets can be extended. This case is addressed in section 1.11.1.

The second possible extension towards multivariate processes is to consider a process x described as a collection of time-processes, $x(t) = (x_1(t), \dots, x_C(t))^T$ with $t \in \mathbb{R}$ a time variable, but the channel c in $x_c(t)$ is arbitrary. For example, a financial index is composed of multiple stocks c having their own price evolving over time t. While the closest neighbors of a discrete time t are $t-1$ and $t+1$, what are the closest neighbors of a channel index c? This case is addressed in section 1.11.2.

In order to build compact models, we discuss and briefly state the notion of multivariate regularity that we rely on to build compact models of x.

1.4.1 Physical fields

Turbulent flows are important examples of physical fields, ruled by the Navier–Stokes equations. In his pioneering work in 1941, Kolmogorov [Kolmogorov, 1941a; Kolmogorov, 1 941; Kolmogorov, 1941b] introduces a self-similar Gaussian model of turbulence which predicts that the projection of the velocity field on a line is a stationary process whose power spectrum has a power-law decay with exponent 2/3.

Intermittency in turbulent flows. Turbulence flows are highly non-Gaussian, and Kolmogorov's initial theory was then refined to take into account intermittency, that is evidenced by the multifractality of the field [Kolmogorov, 1962; Frisch, 1991]. One of the main question was to interpret and include intermittency of turbulent flows in a model. The importance of scale dependencies for explaining intermittency goes back to turbulence models called "shell models" [Lorenz, 1963; Desnianskii, 1974; Siggia, 1978]. They consist of modeling a turbulence by a Navier-Stokes-like equation in each "octave-shell", which are dyadic regions in the Fourier domain, including interaction terms between neighbor shells [Parisi, 1985].

Two-dimensional wavelets and angle dependencies. We are interested here in physical fields as multivariate processes indexed by u belonging to $\mathbb{R}^d$, with $d = 2$. In this case, the univariate wavelet filters $\psi_\lambda(t)$ mentioned above can be extended to multivariate wavelets $\psi_{j,\theta}(u)$ that are also localized in both space and Fourier domain. The wavelet transform $Wx(u, j, \theta) = x \star \psi_{j,\theta}(u)$ extracts variations of x around u at scale 2^j and in the direction $e_\theta = (\cos\theta, \sin\theta)$.

The angle dependencies are crucial to characterize a number of non-Gaussian properties such as the presence of vortices in turbulent fields or filament in cosmological fields. For example, a filament in a field typically produces wavelet coefficients whose amplitudes are large across several scales in the orthogonal direction of the filament but are small in the direction of the filament. Building statistical description Φ that characterize the angle dependencies has been studied in Physics [Allys, 2020; Brochard, 2022; Zhang, 2021]. Authors consider an extension of the phase harmonic covariances (1.8) reviewed in section 1.9.3 that now correlates different phase harmonics at different positions and different oriented scales (different scales and angles). However, this representation contains an even larger number of coefficients, in particular the number of coefficients M may exceed the number of sample of a single field realization, with the risk of reconstructing parts of the observed signal as noticed in [Brochard, 2022], because the estimation is too difficult. Such models showcase an imbalanced bias-variance tradeoff towards large model variance.

One of the main question is to build a low-dimensional model of physical fields that captures intermittency, the presence of structures like vortices, and multifractal properties of a process observed from a single realization.

Our contribution. In order to capture scale and angle dependencies from limited data, we introduce in chapter 3 an extension of the Scattering Spectra to physical fields in Turbulence or Cosmology. Similarly to (1.20) they propose a low-dimensional approximation of the non-linear correlation matrices $\mathbb{E}\{Wx\, Wx\}, \mathbb{E}\{Wx\, |Wx|\}, \mathbb{E}\{|Wx|\, |Wx|\}$ that now correlates different space positions u, u' and different oriented scales λ, λ'. Again, this is performed through a diagonal approximation in a second wavelet operator.

One of the main contribution of this chapter is to provide maximum entropy models of a number of physical fields from Scattering Spectra, that produces very convincing field realizations and that captures high-order moments studied in Cosmology such as bi-spectrum, tri-spectrum, but also structure functions up to order 4. Maximum entropy models from a single realization are shown on figure 1.5.

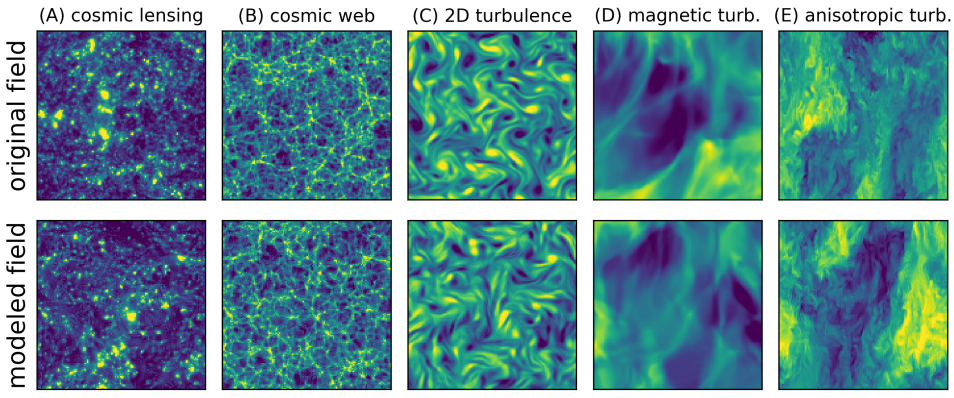


Figure 1.2 – Scattering Spectra models of physical fields. (Top) Original field. (Bottom) Sample from our model estimated on a single realization.

1.4.2 Multi-channel time-processes

A multi-channel time-process $x(t) = (x_1(t), \dots, x_C(t))^T$ is of a different nature than the processes discussed in the previous section. The prominent example of such process in this thesis is the price of different stocks c of the same financial index which cannot be disposed on a line that can be traversed from left to right. It means that $x(t) \in \mathbb{R}^C$ where $\mathbb{R}^C$ is not equipped with the euclidean norm, the true notion of distance, if it exists, is hardly accessible.

Correlation matrices. Building models of such processes requires capturing the dependencies across different channels. The correlation matrix across channels $\Sigma = \mathbb{E}\{x(t)^T x(t)\}$ does contain important information such as the financial sectors e.g. industrial, pharmaceutical. However, building a model relying solely on the correlation matrix to model cross-channel dependencies presents two issues. First, the correlation matrix does not characterize non-Gaussian properties across channels such as the presence of bursts localized at the same time, for example when the market as a whole enters a crisis. More generally, non-Gaussian properties across channels have been evidenced through non-linear statistics by studying deviation of copulas of stock pairs to Gaussian copulas [Chicheportiche, 2014b]. Second, estimation the correlation matrix on limited data is a challenge [Potters, 2005; Tumminello, 2007]. Unlike for a time-correlation matrix of a stationary process, we don't know a predefined basis such as the Fourier basis that would diagonalize the cross-channel correlation matrix nor do we know a wavelet transform across channels to quasi-diagonalize it. In particular, methods based on non-linear wavelet correlations such as [Régaldo-Saint Blancard, 2023] that were performed for 3 or 5 channels cannot be extended on a process observed from a single realization of around 250 stocks because they would yield much more coefficients than the size N of a single realization.

Factor models. Factor models identify a few directions $w^1, \dots, w^r \in \mathbb{R}^C$ along channels and focus on modeling the time structure of the univariate process $\langle w, x\rangle$ projected along these directions $\langle w, x\rangle(t) = \sum_{c=1}^{C} w_c x_c(t)$ called a factor. In a certain sense, it looks at the few factors whose stochastic structures rule the joint stochastic structure. The implicit prior of such models is that the process x can be well described by a reasonable number of factors. While capturing important non-linear dependencies across stocks, they often make simplifying assumptions on the stochastic structure of the factors, for example they may not capture the joint time-asymmetry of the stocks x [Reigneron, 2011].

The main challenge is to find a small number of factors, whose stochastic structure should be modeled accurately, so as to accurately model the joint process x.

Our contribution. In chapter 4 we introduce a maximum entropy factor model that models the stochastic structure of selected factors through the Scattering Spectra introduced in chapter 2. It consists in choosing a vector of statistics $\Phi(x) = (\Phi_{\text{single}}(x), \Phi_{\text{cross}}(x))$ where $\Phi_{\text{single}}(x) = \langle \Phi(x_c) \rangle_c$ characterizes the average stochastic structure of stocks taken individually,

while $\Phi_{\text{cross}}(x)$ characterizes dependencies across channels through selected factors

$$\Phi_{\text{cross}}(x) = \Big(\Phi(\langle w^1, x\rangle), \dots, \Phi(\langle w^r, x\rangle)\Big). \tag{1.11}$$

We show that taking the first $r = 10$ first sparse directions obtained via dictionary learning provides a model that reproduces the main non-Gaussian properties across stocks, including time-asymmetry revealed through a moment of order 3. The vector of statistics (1.21) contains only $r + 1 = 11$ times more statistics than in the univariate case, while the process is $C = 253$ larger in its size. Our model thus strongly relies on the implicit regularity that only a few factors drive the process, at least up to the validation statistics presented in the literature.

1.5 Source separation on Mars

Unsupervised source separation is an example of inverse problem. In a simplified setting, it aims at retrieving source signals $s, n \in \mathbb{R}^N$ from the observation of a mixture signal $x = s + n$ with no access to separated training examples. This is an ill-posed problem that requires prior knowledge on the sources. In certain cases n is assumed to be a noisy signal i.e. a multi-scale signal with certain self-similar properties, and this problem can be regarded as denoising.

Classical signal-processing based source separation methods [Cardoso, 1989; Jutten, 1991; Nandi, 1996; Cardoso, 1998; Starck, 2004; Jutten, 2004; Bobin, 2007] while being extensively studied and well understood, often rely on overly restrictive assumptions regarding the sources, e.g., sources being distributed according to Gaussian or Laplace distributions, which might negatively bias the outcome of source separation [Cardoso, 1998; Parra, 2003].

On the other hand, unsupervised deep learning source separation methods [Févotte, 2009; Drude, 2019; Wisdom, 2020; Liu, 2022; Denton, 2022; Neri, 2021] do not rely on the existence of labeled training data and instead attempt to infer the sources based on the properties of the observed signals. These methods make minimal assumptions about the underlying sources, which make them a suitable choice for realistic source separation problems. Despite their success, unsupervised source separation methods often require tremendous amount of data during training [Wisdom, 2020], which is often infeasible in certain applications such as problem arising in planetary space missions, e.g. due to challenges associated with data acquisition. Moreover, generalization concerns preclude the use of data-driven methods trained on synthetic data in real-world applications due to the discrepancies between synthetic and real data.

Recent works leverage the ability of wavelet representations Φ to accurately describe statistical properties of non-Gaussian multi-scale signals [Regaldo-Saint Blancard, 2021]. Assuming they know the process n and are able to generate as many independent realizations $n_1, \dots, n_K$, their idea is to define a candidate $\bar{n}$ that solves statistical constraints specified by Φ. The candidate signal $\widetilde{n}$ is obtained through a gradient descent that is initiated at x, the mixed signal which contains precious signal-dependent information.

In many cases of interest, such as extraterrestrial seismology, the non-stationarity of the data prevents drawing enough independent realizations of clean signals $n_1, \dots, n_K$, and the challenge

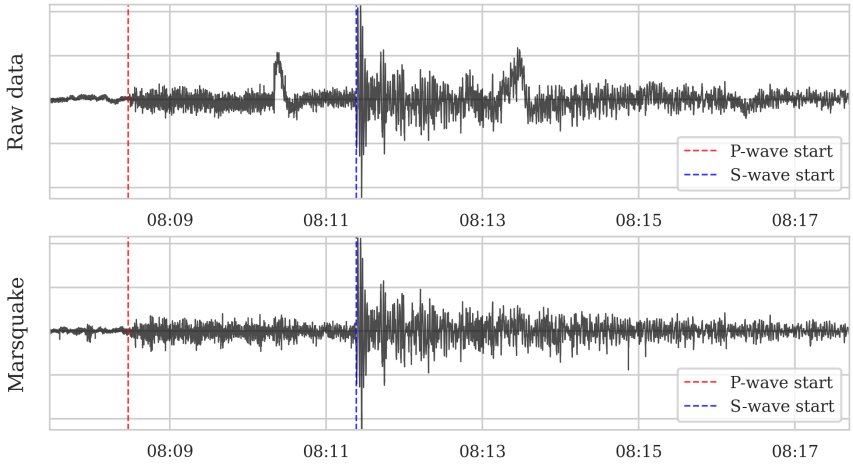


Figure 1.3 – Marsquake separation from limited data through Scattering Spectra.

is to build a prior knowledge of n from limited data.

In chapter 5 we consider seismic data recorded during NASA's InSight mission on Mars. We propose to tackle this challenge by using our Scattering Spectra representation Φ introduced in chapter 2 which can be estimated on limited data. Plugging such statistical description in the same framework [Regaldo-Saint Blancard, 2021] we have been able to use only 50 realizations of the background Mars seismic noise to remove transient events such as glitch from Marsquake data, see Fig. 1.6.

1.6 Prediction on limited data

Prediction, in the context of time-series, is the task of determining quantities $Q(\widetilde{x}_{\text{future}}) \in \mathbb{R}^{M_{\text{future}}}$ of the unknown future $\widetilde{x}_{\text{future}}$ from a given observed past $\widetilde{x}_{\text{past}}$. With a mean-square error objective, this amounts to estimating the following conditional expectation

$$\mathbb{E}\{Q(x_{\text{future}}) \mid x_{\text{past}} = \widetilde{x}_{\text{past}}\}. \tag{1.12}$$

This problem arises especially in Finance where the price process x is multi-scale with long-range dependencies. One can think of variance prediction where Q is an average of squares, or option pricing where Q is the payoff of a call option [1].

Linear regression models propose to identify predictors $h(x_{\text{past}}) \in \mathbb{R}^{M_{\text{past}}}$ that "correlate" the most with $Q(x_{\text{future}})$. Setting aside the search for best predictors $h(x_{\text{past}})$ [Ghysels, 2006; Christiansen, 2012], this requires estimating $M_{\text{past}} \times M_{\text{future}}$ correlation coefficients on limited data. In order to avoid overfitting, best linear methods focus on predicting few quantities, e.g. $M_{\text{future}} = 1$, with a few well identified predictors $h(x_{\text{past}})$ e.g. $M_{\text{past}} \leq 4$ [Guyon, 2022].

Non-local kernel methods by-pass the estimation of correlation coefficients by averaging over data points. It assigns weights to observed data points $(x^i_{\text{past}}, x^i_{\text{future}}), 1 \leq n$ based on their

1. In this case, the expectation $\mathbb{E}$ is under the risk-neutral measure (see chapter 6).

proximity to $\widetilde{x}_{\text{past}}$ defined by a kernel k, e.g. a Gaussian kernel, and perform a weighted average

$$\bar{Q}(\widetilde{x}_{\text{future}}) = \sum_{i=1}^{n} w_i k(\widetilde{x}_{\text{past}}, x^i_{\text{past}}) Q(x^i_{\text{future}}). \tag{1.13}$$

This is called a Nadaraya–Watson estimator [Nadaraya, 1964; Watson, 1964], under certain hypotheses on process x it is an unbiased estimator of (1.22) when the number of data $n \to +\infty$ and the kernel concentrates around $\widetilde{x}_{\text{past}}$ [Hansen, 2008].

These methods are non-local in the sense that there are no reasons for the most similar data x^i to be near $\widetilde{x}_{\text{past}}$ in time. For image denoising, non-local means [Buades, 2011] exploit the same idea. In order to denoise the central pixel of a patch, their algorithm compares the patch with all the patches in the image based on their Euclidien norm. An estimate is obtained by a weighted average over the central pixels of the collected patches. This algorithm achieves good denoising results, showing that the algorithms succeeds in finding similar patches, among the few patches available in a single image, that are informative enough to denoise the current pixel.

Such method would fail for processes in Finance, which have a lot of noise. Indeed, there is very few chances that different short realizations of financial log-returns are close to each other, from limited data. This is to say that the volume of set of trajectories with high probability is very large i.e. a financial price process is a process with high entropy.

Opting for a non-local kernel method, the goal is to exploit the regularity of the process in order to be able to find enough closest paths x^i with predictive power on $\widetilde{x}_{\text{past}}$ despite the high entropy of the process.

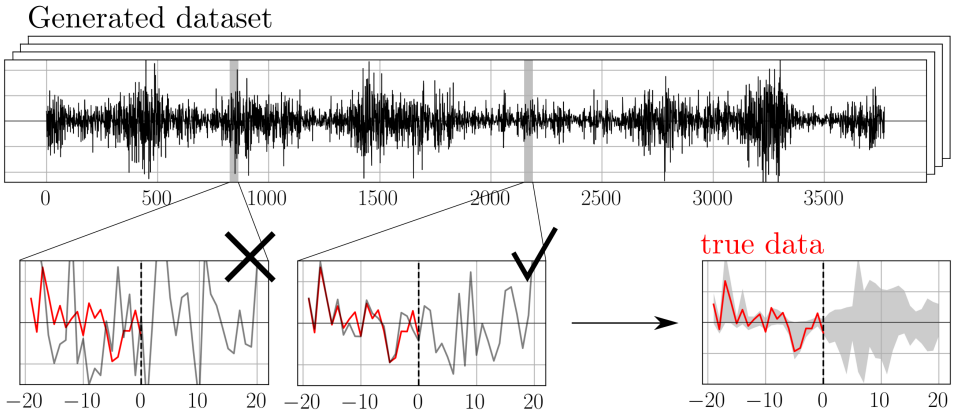


Figure 1.4 – Path shadowing illustration for prediction on limited data. Given an observed past path (red) we average future quantities on paths with similar past, called shadowing paths.

In chapter 6 we propose to scan for paths x^i in a dataset of generated paths from a Scattering Spectra model of x, we call it Path Shadowing, illustrated in Fig. 1.7. This is inspired from the study of chaotic processes. Intuitively, the shadowing property [Hammel, 1987], states that a path which is uniformly close to a true orbit will stay close (shadow) a true path for all time.

In our case, the high entropy of the process is both a blessing and a curse. It helps a Scattering Spectra maximum entropy model to approximate accurately the true distribution

p. However, we need to scan a lot of paths in order to find shadowing paths. We propose to leverage the scale invariance in the process to make this step feasible. For that we choose a kernel that is based on a multi-scale causal representation h of the past x_{past} that considers short-range and long-range past data with only a few parameters [Renaud, 2003; Renaud, 2005; Andreux, 2018]. We also choose a kernel that is invariant to scaling $x \mapsto \lambda x$ for $\lambda > 0$ and dilation $x(t) \mapsto x(\lambda^{-1}t)$.

Path Shadowing Monte-Carlo provides state-of-the-art volatility prediction results and can be used to obtain option smiles whose quality is assessed through a trading game.

1.7 Outline of the thesis

The three first chapters 2,3 and 4 are devoted to building maximum entropy models of multi-scale processes that can be estimated on limited data, by specifying the vector of statistics $\Phi(x)$ that should be imposed.

Models of univariate non-Gaussian processes require characterizing the dependencies across different scales. In chapter 2 we propose to start from the non-linear time-scale correlation matrix (1.7). First, we investigate self-similar properties of this matrix and show that it can be used to define a wide-sense definition of self-similarity. Second, we show how to reduce the number of coefficients by performing a diagonal approximation after a second wavelet operator which yields the Scattering Spectra. A maximum entropy model based on these moments is evaluated for univariate turbulence and financial time-series.

Models of physical fields – two or three dimensional processes – require to characterize the dependencies across oriented scales. In chapter 3 we propose an extended vector of statistics $\Phi(x)$, still called Scattering Spectra, that now characterize dependencies between oriented scales. We show in this chapter that regular dependencies in scales or angles can be leveraged by Fourier thresholding to reduce the number of coefficients of these Scattering Spectra. We propose low-dimensional models of two-dimensional physical fields in Turbulence and Cosmology.

Chapter 4 focuses on multivariate time-series in a specific case that is the different stocks of a financial index. This chapter tackles the problem of characterizing dependencies across time-series with as few coefficients as possible through modeling the time structure along selected directions.

In the remaining chapters 5 and 6, we propose two applications that are made possible by such models. Chapter 5 falls into the realm of inverse problems. We tackle unsupervised source separation on limited data by plugging our model in an existing method. Results are shown on seismic data from a space mission on Mars. Chapter 6 tackles the problem of prediction on limited data. We introduce a non-local kernel method called Path Shadowing Monte-Carlo. It relies heavily on a generative model of the underlying process to produce a diverse set of paths with the correct time-dependencies.

The work in this thesis resulted in seven papers: four submitted journal papers [Morel, 2022; Cheng, 2024; Morel, 2023b; Aubrun, 2024], one published conference paper [Siahkoohi, 2023b], one submitted conference paper [Siahkoohi, 2023a] and one paper in preparation [Morel, 2023a].

— **Rudy Morel**, Gaspar Rochette, Roberto Leonarduzzi, Jean-Philippe Bouchaud, Stéphane Mallat. "*Scale Dependencies and Self-Similar Models with Wavelet Scattering Spectra*". Submitted to a journal, 2022.

— Sihao Cheng, **Rudy Morel**, Erwan Allys, Brice Ménard, Stéphane Mallat. "*Scattering Spectra Models for Physics*". Submitted to a journal, 2023.

— Ali Siahkoohi, **Rudy Morel**, Maarten de Hoop, Erwan Allys, Grégory Sainton, Taichi Kawamura. "*Unearthing InSights into Mars: Unsupervised Source Separation with Limited Data*". International Conference on Machine Learning, 2023.

— Ali Siahkoohi, **Rudy Morel**, Randall Balestriero, Erwan Allys, Grégory Sainton, Taichi Kawamura, Maarten de Hoop. "*Martian time-series unraveled: A multi-scale nested approach with factorial variational autoencoders*". Submitted to a conference, 2023.

— **Rudy Morel**, Stéphane Mallat, Jean-Philippe Bouchaud. "*Path Shadowing Monte-Carlo*". Submitted to a conference, 2023.

— **Rudy Morel**, Stéphane Mallat, Jean-Philippe Bouchaud. "*A maximum entropy factor model of financial stocks*". In preparation, 2023.

— Cécilia Aubrun*, **Rudy Morel***, Michael Benzaquen, Jean-Philippe Bouchaud. "*Riding wavelets: A method to discover new classes of price jumps, 2024*".

.

Résumé détaillé de la thèse

Les processus rencontrés dans de nombreux domaines sont multi-échelles, c'est-à-dire qu'ils présentent des variations sur une large gamme d'échelles. C'est le cas par exemple d'une série temporelle de prix financiers, ou d'un enregistrement sismique. En physique, c'est le cas de nombreuses observations en deux ou trois dimensions, telles que la vélocité d'un fluide turbulent, ou la distribution à grande échelle de la matière noire dans l'univers.

Le but principal de cette thèse est de construire des modèles probabilistes de processus multi-échelles, à partir d'observations limitées, et pouvant être échantillonnés numériquement. Il s'agit d'un problème d'apprentissage non-supervisé qui peut être formulé comme le fait de construire d'une distribution p_θ, pouvant être échantillonnée numériquement, et qui approche la distribution p du processus sous-jacent x, dont on observe une seule réalisation $\widetilde{x} \in \mathbb{R}^N$ de taille limitée, où $\theta \in \mathbb{R}^M$ sont des paramètres avec M la dimension du modèle. Dans cette thèse, on suppose le processus x stationnaire, ou à incréments stationnaires, et ergodique.

Ce sujet est crucial pour attaquer de nombreux problèmes formulés dans un régime de données limitées, tels que la génération (tirer de nouvelles réalisations de x), mais est également utile pour la prédiction (déterminer un ensemble de valeurs inconnues de x), en particulier lorsqu'il n'y a pas suffisamment de données étiquetées pour entraîner un algorithme supervisé. Il est également crucial pour aborder des problèmes inverses tels que la séparation de source non-supervisée (séparer deux signaux dont seule la somme est observée) qui peuvent être rendus bien-définis si l'on suppose un modèle a priori de l'une des sources.

La contrainte des données est une restriction forte qui émerge de principes fondamentaux. Nous possédons souvent une seule réalisation du processus x étudié. En finance, il existe une seule réalisation du processus de prix d'un certain indice : la réalisation historique. En astrophysique, nous observons une seule carte de l'univers. Le processus x peut contenir des dépendances à longue portée à l'échelle de la réalisation, ce qui signifie que $x(u)$ et $x(u')$ restent dépendants pour les plus grandes valeurs de $u-u'$. De telles dépendances sont difficiles à estimer précisément car le nombre d'échantillons de la distribution jointe $(x(u), x(u'))$, à partir d'une seule réalisation, est très petit.

Le principal défi dans la construction de modèles de processus multi-échelles à partir de données limitées réside dans un compromis biais-variance que nous expliquons ci-dessous.

Biais du modèle. Les processus d'intérêt dans de nombreux domaines sont souvent non-gaussiens. Cela peut se manifester par l'intermittence ou l'asymétrie temporelle pour processus

temporels, ou par la présence de structures transitoires dans un champ physique bidimensionnel. En particulier, les modèles gaussiens, qui sont bien compris, caractérisés par leur moyenne et leur covariance, échouent à capturer des propriétés essentielles du processus. Ces modèles sont biaisés car ils reposent sur une description assez pauvre du processus étudié.

Variance du modèle. Un moyen d'améliorer la précision d'un modèle paramétrique p_θ dans l'approximation de la distribution sous-jacente p est d'augmenter le nombre de paramètres $\theta \in \mathbb{R}^M$, de manière à reproduire un ensemble plus riche de statistiques $\Phi(x)$. Calibrer le modèle consiste alors à trouver un $\theta \in \mathbb{R}^M$ à partir de données limitées $\widetilde{x}$ telles que p_θ reproduit ces statistiques. Cependant, élargir le vecteur de statistiques Φ augmente sa variance, ce qui rend son estimation sur des données limitées $\widetilde{x}$ plus difficile. Cela signifie que le modèle p_θ du même processus p peut différer significativement lorsqu'il est estimé sur des réalisations $\widetilde{x}$ différentes.

Ainsi, le nombre de paramètres doit être choisi avec soin. Dans cette thèse, nous appelons *modèle compact* un modèle p_θ dont le nombre de paramètres M, désigné comme la dimension du modèle, croît en $o(N)$ avec la taille N d'une seule réalisation $\widetilde{x}$.

L'un des défis de cette thèse est de tirer parti de la nature multi-échelles des données observées afin d'arbitrer au mieux le compromis biais-variance en définissant un a priori sur la distribution sous-jacente p. Afin de mettre en évidence les principales contributions de la thèse, nous présentons quelques concepts clés et outils de la littérature pour construire des modèles de processus multi-échelles.

1.8 Modèles à maximum d'entropie

1.8.1 Promouvoir la diversité des modèles

Dans son article fondateur, Jaynes [Jaynes, 1957] propose d'établir un modèle de p à partir d'observations partielles en maximisant son entropie. L'entropie d'une distribution de processus p est donnée par $H(p) = -\int p(x) \log p(x) dx$. Un modèle macro-canonique p_θ du processus x peut être défini comme une distribution d'entropie maximale conditionnée par la valeur exacte d'un vecteur de moments $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \mu$, où $\Phi : \mathbb{R}^N \mapsto \mathbb{R}^M$ et où l'on choisit $\mu = \mathbb{E}_p\{\Phi(x)\}$ pour l'instant. S'ils existent, ces modèles ont une distribution de probabilité exponentielle

$$p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \theta, \Phi(x) \rangle}, \tag{1.14}$$

pour un $\theta \in \mathbb{R}^M$ donné, où M est le nombre de paramètres qui est également le nombre de statistiques Φ. Un tel modèle est le moins biaisé, étant donné une contrainte de moment $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \mu$, dans le sens où il est le plus neutre possible en ce qui concerne les informations manquantes [Jaynes, 1957]. Par exemple, un processus gaussien stationnaire p est un modèle d'entropie maximale conditionné par les moments d'ordre un et deux $\mathbb{E}_p\{\Phi(x)\}$, avec $\Phi(x) = \big(\langle x(t) \rangle_t \,, \langle x(t-\tau) x(t) \rangle_t \big)$.

Les paramètres θ du modèle (1.14) peuvent être estimés par des méthodes de Monte-Carlo à chaînes de Markov [Lustig, 1998 ; Betancourt, 2017] qui fournissent un algorithme exact, mais computationnellement coûteux lorsque le nombre de statistiques Φ est grand, en raison du temps

de mélange de la chaîne de Markov [Levin, 2017 ; Bruna, 2019].

Dans cette thèse, pour éviter ce problème computationnel, nous considérons des modèles à maximum d'entropie micro-canoniques qui ont une distribution d'entropie maximale sur un ensemble

$$\Omega_\epsilon = \{x \in \mathbb{R}^N \mid \|\Phi(x) - \Phi(\widetilde{x})\|_2 \leq \epsilon\}.$$

pour une certaine erreur ϵ qui est ajustée à la variance de $\Phi(x)$. En général, Φ est de sorte que Ω_ϵ est un ensemble compact de mesure de Lebesgue $\int_{\Omega_\epsilon} dx$ strictement positive. Dans ce cas, le modèle micro-canonique a une distribution uniforme sur Ω_ϵ.

Si $\Phi(x)$ se concentre autour de $\mathbb{E}\{\Phi(x)\}$ alors le modèle micro-canonique converge vers le modèle macro-canonique (1.14) lorsque la taille N de $\widetilde{x}$ tend vers ∞ et ϵ tend vers 0. C'est le principe d'équivalence de Boltzmann [Lanford, 1975 ; Gallagher, 2013]. La concentration de $\Phi(x)$ impose généralement que sa dimension M soit petite par rapport à la dimension N de x.

Échantillonner un modèle micro-canonique peut être réalisé par une descente de gradient sur l'erreur $x \mapsto \|\Phi(x) - \Phi(\widetilde{x})\|^2$ à partir d'une réalisation initiale d'un bruit gaussien, qui a une distribution d'entropie maximale [Bruna, 2019].

Un cas typique d'échec dans la modélisation est un modèle de faible entropie, par exemple lorsque le modèle se concentre autour d'une seule réalisation. Étant donné la contrainte de moment $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \mu$ dans un modèle macro-canonique, ou la contrainte statistique $\|\Phi(x) - \Phi(\widetilde{x})\|_2 \leq \epsilon$ dans un modèle micro-canonique, maximiser l'entropie est un moyen de promouvoir la diversité des réalisations, c'est-à-dire d'augmenter le volume des ensembles de probabilité élevée [Shannon, 1948].

1.8.2 Revisiter le compromis biais-variance

Le compromis biais-variance dans la construction d'un modèle de p apparaît explicitement dans un modèle d'entropie maximale. Un modèle d'entropie maximale est maximallement neutre concernant les informations manquantes [Jaynes, 1957]. Ainsi un nombre trop petit de statistiques $\Phi(x)$ peut être insuffisant pour caractériser les propriétés importantes du processus x conduisant à un biais de modèle élevé.

D'autre part, sans accès à p, les moments $\mu = \mathbb{E}_p\{\Phi(x)\}$ sont estimés par $\Phi(\widetilde{x})$ sur une seule réalisation $\widetilde{x}$ de taille limitée. Pour que le modèle soit précis, il est nécessaire que $\Phi(\widetilde{x})$ soit proche de $\mathbb{E}\{\Phi(x)\}$, ce qui peut être assuré en choisissant des statistiques $\Phi(x)$ de faible variance, limitant ainsi la variance du modèle.

Le principal défi dans un modèle d'entropie maximale est de définir des statistiques Φ qui spécifient les propriétés importantes de x, de manière à produire un modèle précis de p, tout en restant à faible variance pour que $\Phi(\widetilde{x})$ soit une bonne estimation de $\mathbb{E}_p\{\Phi(x)\}$.

Les moments d'ordre élevé sont un candidat qui revient à définit Φ comme la moyenne temporelle de polynômes en les coordonnées de x. Pour $r \in \mathbb{N}^*$, en supposant que $\mathbb{E}\{|x(u)|^r\} < +\infty$, les moments $\mathbb{E}\{\Phi(x)\}$ s'écrivent

$$\mathbb{E}\{x(u_1) \dots x(u_r)\} \tag{1.15}$$

Sous certaines conditions [Billingsley, 2013], l'expansion infinie $r \in \mathbb{N}$ fournit une description exacte de la distribution du processus. Cependant, ils sont difficiles à estimer à partir de données limitées. En effet, les polynômes d'ordre élevé amplifient les événements rares, ce qui entraîne une grande variance d'estimation. Ce problème est typiquement amplifié pour les processus avec des distributions à queue épaisse. La prochaine section se concentre sur des candidats pour Φ issus de la littérature.

1.9 Dépendances d'échelle

Les processus d'intérêt dans le monde réel sont souvent non-gaussiens. Concevoir des moments $\mathbb{E}\{\Phi(x)\}$ qui caractérisent les propriétés non-gaussiennes du processus et qui peuvent être estimés sur des données limitées $\widetilde{x}$ par $\Phi(\widetilde{x})$, est un sujet de recherche important. Nous nous concentrons dans cette section sur les processus temporels univariés et étudierons les extensions multivariées dans la section 1.11.

Séparer les échelles dans un processus multi-échelle x peut être fait avec une transformation en ondelettes, qui est présentée dans la section 1.9.1. Nous montrons que les fonctions de structure peuvent être utilisées pour suivre l'évolution de la distribution du processus à différentes échelles via des moments d'ordre élevé.

Dans la section 1.9.2, nous présentons les réseau en scattering qui charactérisent la distribution des coefficients d'ondelettes en cascadant des opérateurs d'ondelettes et des non-linéarités, inspirés par l'utilisation de réseaux neuronaux convolutionnels en apprentissage automatique.

Dans la section 1.9.3, nous présentons une approche différente, reposant toujours sur les coefficients d'ondelettes, mais qui recherche maintenant des dépendances d'échelle dans la distribution jointe des coefficients d'ondelettes à travers le temps et les échelles.

1.9.1 Transformée en ondelettes et fonctions de structure

Une transformation en ondelettes sépare les variations à plusieurs échelles. Elle est calculée avec un filtre complexe ψ à moyenne nulle $\int \psi(t)\mathrm{d}t = 0$ qui est localisé à la fois dans le domaine temporel et dans le domaine fréquentiel [Y Meyer, 1992 ; Mallat, 1999]. L'opérateur de transformation en ondelettes W est alors

$$Wx(t, j) = x \star \psi_j(t) \quad \text{où} \quad \psi_j(t) = 2^{-j}\psi(2^{-j}t).$$

Plus spécifiquement, nous pouvons choisir une ondelette ψ dont la transformée de Fourier $\widehat{\psi}(\omega) = \int \psi(v)e^{-i\omega t}\mathrm{d}t$ est principalement concentrée à des fréquences $\omega \in [\pi, 2\pi]$. Il résulte que $\widehat{\psi}_j(\omega) = \widehat{\psi}(2^j\omega)$ est non-négligeable principalement dans $\omega \in [2^{-j}\omega, 2^{-j+1}\omega]$. Cela fournit une séparation de l'axe des fréquences en différents fenêtres qui constituent notre notion d'échelles. Pour un processus stationnaire x, ou avec des incréments stationnaires, le processus joint $Wx(t, j)$ est stationnaire, sous certaines conditions sur le filtre en ondelettes [Pipiras, 2017].

Les fonctions de structure suivent l'évolution de la distribution des coefficients d'ondelettes

à différentes échelles à travers ses moments d'ordre élevé

$$S(q, j) = \mathbb{E}\{|x \star \psi_j(t)|^q\}. \tag{1.16}$$

Bien que $S(2, j)$ ne contienne pas plus d'informations qu'un modèle gaussien ne puisse capturer, les $S(q, j)$ pour $q \neq 2$ peuvent différer des statistiques gaussiennes.

Pour les processus multi-échelles, l'auto-similarité fait référence aux propriétés d'invariance d'échelle de la distribution du processus. Une définition de l'auto-similarité sera discutée dans la suite. Au niveau des fonctions de structure, l'auto-similarité est caractérisée par une loi de puissance sur la plage des échelles étudiée

$$S(q, j) = c_q \, 2^{j\zeta_q}. \tag{1.17}$$

Les premiers travaux en analyse multifractale ont utilisé ces exposants ζ_q pour déterminer le spectre de singularité d'un signal x qui caractérise la variabilité des exposants de Hölder ponctuels de x [Bacry, 1993 ; Muzy, 1994 ; Jaffard, 2004]. Les problèmes d'estimation, inhérents aux moments d'ordre élevé, peuvent être résolus en introduisant des puissances de modules de maxima de coefficients d'ondelettes appelées *wavelet leaders* [Jaffard, 2006 ; Wendt, 2009]. Ces *wavelet leaders* offrent l'avantage de ne pas nécessiter de fortes hypothèses de stationnarité ou d'ergodicité et peuvent être estimés pour des exposants réels positifs et négatifs. Ces quantités multi-échelles et leur évolution à travers les échelles ont été utilisées avec succès pour détecter et discriminer les propriétés des processus non-gaussiens, tels que l'intermittence, avec des applications en médecine par exemple [Abry, 2010 ; Saës, 2022].

La construction de modèles de processus multi-échelles à partir de moments (1.16), voire même à partir de quantités multi-échelles récentes utilisées en analyse multifractale, possède un désavantage de taille. De tels moments ne capturent pas le propriétés non-gaussiennes importantes telles que l'asymétrie temporelle, en modifiant $x \star \psi_j(u)$ en $x \star \psi_j(-u)$ ne modifie pas (1.16), ce qui est crucial pour construire des modèles précis de processus temporels.

Ce problème affecte également la question fondamentale de la définition de l'auto-similarité. Celle-ci admet une définition forte qui stipule que la distribution jointe des coefficients d'ondelettes est invariante par dilatation, à des facteurs multiplicatifs aléatoires près [Mandelbrot, 1997]. Cependant, en tant que définition en distribution, elle ne peut pas être testée numériquement sur une seule réalisation. Les fonctions de structure fournissent une définition numériquement vérifiable (du moins pour de petits exposants). Cependant, comme mentionné ci-dessus, elle fournit une description faible de la distribution du processus.

Un défi important est de trouver une notion d'auto-similarité basée sur une description plus riche du processus et qui puisse être testée numériquement sur une seule réalisation. Ce problème est abordé dans la section 1.10.2.

1.9.2 Transformée en scattering

Au lieu de considérer des statistiques d'ordre élevé, une transformée de scattering Sx propose d'analyser la structure temporelle des coefficients d'ondelettes $Wx(t, j)$ à une échelle fixe 2^j à

travers une cascade de transformations en ondelettes et de non-linéarités [Mallat, 2012 ; Bruna, 2013]. Définis jusqu'à une échelle maximale 2^J, cettre transformée concatène les coefficients de scattering S_m à différents ordres $0 \leq m \leq J$

$$S_m x(t, j_1, \ldots, j_m) = |\ldots|x \star \psi_{j_1}| \star \psi_{j_2}| \ldots | \star \psi_{j_m}(t) \tag{1.18}$$

pour $1 \leq j_1 < \ldots < j_m \leq J$. Les moments de scattering sont estimés par une moyenne empirique $\Phi(x) = \langle S_m x(t, j_1, \ldots, j_m) \rangle_t$.

Contrairement aux fonctions de structure, les moments de scattering sont 1-Lipschitz en x. Ces coefficients peuvent être utilisés pour analyser l'intermittence dans les processus multi-échelles [Bruna, 2015]. Ils peuvent également être utilisés dans la classification audio [Andén, 2018] ou la détection et le regroupement d'événements sismiques [Seydoux, 2020 ; Rodrıguez, 2021]. Cependant, de la même manière que les fonctions de structure, ils ne fournissent pas une description statistique assez riche du processus pour obtenir des modèles précis, en particulier ils ne capturent pas l'asymétrie temporelle.

Les coefficients de scattering, tout comme les fonctions de structure, suppriment la phase des coefficients d'ondelettes via un module complexe, afin d'obtenir des coefficients non nuls après moyennage temporel. L'asymétrie temporelle peut être détectée par la phase des coefficients d'ondelettes. En effet, dans le cas d'une ondelette analytique ($\widehat{\psi}$ est réelle), le filtre $\operatorname{Im} \psi$ est une fonction impaire et le signe de $x \star \operatorname{Im} \psi(t)$ détecte l'asymétrie du signal. Une question importante est de capturer ces dépendances de phase afin de capturer les propriétés non-gaussiennes de x.

1.9.3 Corrélations non linéaires des coefficients d'ondelettes

Une autre approche pour construire une représentation non-gaussienne des processus temporels consiste à considérer des moments $\mathbb{E}\{\Phi(x)\}$ sous la forme de corrélations sur une représentation 1-Lipshitz $\mathcal{R}$

$$\mathbb{E}\{\mathcal{R}x(t, \lambda)\, \mathcal{R}x(t', \lambda')^*\} \tag{1.19}$$

où t, t' sont des indices temporels et λ, λ' sont des indices de la représentation $\mathcal{R}$. Par exemple, la littérature sur l'apprentissage automatique fournit des représentations $\mathcal{R}$ sous la forme de cascades d'opérateurs de convolution linéaires et de non-linéarités ponctuelles appelées réseaux neuronaux convolutionnels [Gatys, 2015 ; Ustyuzhaninov, 2017]. Cependant, cela conduit à un nombre énorme de coefficients corrélations M, bien plus grand que la taille N des données, avec le risque d'avoir un modèle à grande variance. De plus, l'interprétation des coefficients est très difficile.

Les dépendances entre différentes échelles $2^j \neq 2^{j'}$ se sont avérées cruciales pour caractériser la distribution d'un processus multi-échelle, en particulier ses propriétés non-gaussiennes. Par exemple, la présence d'une crise dans une série temporelle ou de structures dans une image donne lieu à de grands coefficients au même emplacement [Portilla, 2000]. Les auteurs de [Gatys, 2015 ; Ustyuzhaninov, 2017] montrent en fait que la corrélation de cartes à des échelles différentes est essentielle pour obtenir les meilleurs résultats perceptuels pour la synthèse de textures.

Fixer $\mathcal{R} = W$ dans les moments (1.19) conduit à un modèle linéaire en les coefficients

d'ondelettes de x. En tant que tel, il s'agit d'un modèle gaussien qui n'est pas précis pour de nombreux processus d'intérêt.

Cette défaillance s'explique par le fait que la corrélation des coefficients d'ondelettes ne capture pas les dépendances entre les échelles. En effet, pour les processus x avec un spectre de puissance régulier, la corrélation d'ondelettes

$$\mathbb{E}\{x \star \psi_j(t)\, x \star \psi_{j'}(t')^*\}$$

décroît rapidement autour de $t = t'$ et $j = j'$ [Wornell, 1993]. En effet, pour des échelles séparées $2^j \neq 2^{j'}$, les supports fréquentiels de $x \star \psi_j(t)$ et $x \star \psi_{j'}(t')$ se chevauchent à peine, en raison de la séparation d'échelle dû aux ondelette. Les deux processus $x \star \psi_j(t)$ et $x \star \psi_{j'}(t')$ oscillent donc à des fréquences différentes et leur corrélation est annulée par les oscillations de phase.

Une question importante est donc de retrouver les dépendances d'échelle à travers des corrélations non linéaires.

1.10 Spectres en scattering

En s'appuyant sur les travaux précédents que nous avons examinés dans les dernières sections, le chapitre 2 introduit une représentation de corrélation appelée spectres en scattering, qui est une représentation compacte des dépendances d'échelle pouvant être estimée sur des données limitées.

Pour cela, nous exploitons davantage le caractère multi-échelles du processus x de deux manières différentes qui sont exposées dans les sections suivantes.

1.10.1 Covariance de coefficients de scattering

Nous partons des corrélations module-phase (1.7). La structure temporelle des enveloppes $|Wx|$ est capturée par des corrélations entre tous les t, t'. Cependant, de telles enveloppes ont généralement des dépendances à longue portée avec un spectre croisé régulier et nous savons que cette covariance de processus peut être compressée via la transformée en ondelettes [Wornell, 1993]. Cascader une deuxième transformée en ondelettes produit une transformée en scattering $Sx = W|Wx|$ avec $Sx(t, j_1, j_2) = |x \star \psi_{j_1}| \star \psi_{j_2}(t)$ (voir la section 1.9.2). La matrice de corrélation résultante après une deuxième transformée en ondelettes est $\mathbb{E}\{Sx\, Sx^T\} = W\, \mathbb{E}\{|Wx|\, |Wx|^T\}\, W^T$. Cette matrice considère les corrélations des coefficients de scattering entre des canaux séparés et constitue une extension des coefficients de scattering standard (1.18) d'ordre $m = 2$. Pour les processus avec des spectres croisés d'enveloppe réguliers, tels que ceux rencontrés en finance ou en physique, grâce aux propriétés de compression de corrélation des ondelettes [Wornell, 1993], la matrice $\mathbb{E}\{Sx(t, j_1, j_2)\, Sx(t', j_1', j_2')\}$ a une structure parcimonieuse et est concentrée le long de sa diagonale $t = t', j_2' = j_2$. Nous notons Diag une telle projection diagonale.

Une des principales contributions du chapitre 2 est d'introduire les spectres en scattering qui

sont une approximation diagonale des corrélations non linéaires (1.7).

$$\Big(\mathrm{Diag}\,\mathbb{E}\{Wx, Wx^T\}\ ,\ \mathrm{Diag}\,\mathbb{E}\{Wx, |Wx|^T\}\ ,\ \mathrm{Diag}\,\mathbb{E}\{W|Wx|, W|Wx|^T\}\Big). \tag{1.20}$$

Ils étendent le spectre de puissance standard des coefficients d'ondelettes $\mathrm{Diag}\,\mathbb{E}\{Wx, Wx^T\}$. L'asymétrie de signe, souvent appelée *skewness*, est détectée par la deuxième matrice de corrélation $\mathrm{Diag}\,\mathbb{E}\{Wx, |Wx|^T\}$. L'intermittence du processus est caractérisée par la troisième matrice $\mathrm{Diag}\,\mathbb{E}\{W|Wx|, W|Wx|^T\}$. Ce sont des coefficients complexes et nous prouvons que leur partie imaginaire capture l'asymétrie temporelle.

Ils sont estimés en remplaçant $\mathbb{E}$ par une moyenne temporelle $\langle . \rangle_t$. Pour une réalisation $x \in \mathbb{R}^N$ de taille $N = T$, le nombre de pas de temps, les spectres en scattering $\Phi(\widetilde{x})$ consistent en $\mathcal{O}(\log_2^3 T)$ coefficients d'ordre 2, bien inférieurs à T, et peuvent donc être estimés sur des données limitées.

Nous montrons qu'ils fournissent des modèles précis de séries temporelles financières et de série de turbulence univariés, et capturent les principales propriétés non-gaussiennes telles que les distributions à queues épaisse, l'intermittence, l'asymétrie de signe et l'asymétrie temporelle. Fait intéressant, un modèle basé sur de tels moments d'ordre 2 est capable de reproduire les statistiques d'ordre supérieur (jusqu'à $q = 5$).

1.10.2 Auto-similarité au sens large

Comme expliqué dans la section 1.9.3, les corrélations non linéaires (1.7) caractérisent les dépendances d'échelle et capturent des propriétés non-gaussiennes telles que l'asymétrie de signe et l'asymétrie temporelle qui n'étaient pas capturées par les fonctions de structure (1.16).

Dans le chapitre 2, nous montrons que la définition forte d'auto-similarité définie sur la distribution du processus [Mandelbrot, 1997] implique une invariance d'échelle des matrices (1.7), à un facteur de normalisation près. Cette définition est dite au sens large en analogie avec la stationnarité temporelle au sens large.

Définition au sens large. Communément utilisée en traitement du signal, la stationnarité temporelle au sens large considère la matrice de corrélation à travers le temps $C(t, \tau) = \mathbb{E}\{x(t)\, x(t+\tau)\}$. Supposons que x ait une moyenne nulle $\mathbb{E}\{x(t)\} = 0$. Le processus x est dit stationnaire au sens large si la corrélation $C(t, \tau)$ est indépendante de t. Cela revient à dire que la matrice C est invariante par translation temporelle. La matrice de corrélation module-phase $C_{\rho W}(t, t', j, j')$ caractérise les dépendances à travers le temps et les échelle via les trois matrices $\mathbb{E}\{Wx\, Wx\}, \mathbb{E}\{Wx\, |Wx|\}, \mathbb{E}\{|Wx|\, |Wx|\}$ (1.7). Réindexons cette matrice en $C_{\rho W}(t, \tau, j, a)$, avec $\tau = t - t', a = j - j'$. En raison de la stationnarité au sens large, cette corrélation ne dépend pas de t. L'auto-similarité au sens large implique que la même propriété vaut pour les log-échelles j, j'. Afin de l'affirmer, nous devons prendre en compte le fait que la variance des coefficients d'ondelettes, le spectre de puissance en ondelette, n'est pas nécessairement constant à travers les échelles. Nous normalisons donc la corrélation $C_{\rho W}$ par le spectre de puissance en ondelette de x. Le chapitre 2 montre que pour un processus auto-similaire (au sens fort), la matrice de phase-

module normalisée $C_{\rho W}(t, \tau, j, a)$ ne dépend pas de j, elle est invariante par translation d'échelle. C'est ce que l'on appelle l'auto-similarité au sens large, introduite dans cette thèse. La figure 1.1 (à droite) illustre cette invariance par translation d'échelle sur la matrice $\mathbb{E}\{|Wx|\,|Wx|\}$.

Cette définition repose sur un ensemble de statistiques qui caractérisent des propriétés non-gaussiennes importantes telles que l'intermittence, l'asymétrie de signe et l'asymétrie temporelle, et qui peuvent être utilisées pour construire des modèles précis de p, via leur réduction en spectre en scattering (1.20). En supposant l'auto-similarité au sens large, nous prouvons que les spectres en scattering (1.20), qui compressent les corrélations non linéaires (1.7), sont invariants par translation d'échelle, à renormalisation près.

Numériquement, on peut montrer qu'imposer des spectres en scattering invariants par changement échelle permet de retrouver la décroissance en loi de puissance des fonctions de structure (1.16) jusqu'à l'ordre 4.

Régularité d'échelle. Bien que l'auto-similarité semble être satisfaite pour un processus de prix financier de l'échelle de quelques minutes à l'échelle d'une décennie, elle ne l'est pas en général, pour une série de turbulence par exemple. La dilatation du processus x agit sur la matrice $C_{\rho W}(j, j' - j)$ comme une translation de sa première variable j. La régularité en échelle peut être définie comme la régularité de $C_{\rho W}$ en fonction de j. Dans le chapitre 3, nous calculons une transformée de Fourier le long de j qui donne $\widehat{C}_{\rho W}(\omega, j' - j)$. Pour un processus régulier à travers les échelles, les coefficients $\widehat{C}_{\rho W}(\omega, j' - j)$ décroissent rapidement en ω. Cela est utilisé pour fournir un modèle avec un nombre adaptatif de coefficients en seuillant les harmoniques de Fourier. Cela permet de construire des modèles encore plus réduits de processus avec régularité le long des échelles et aide à réduire la variance du modèle.

1.11 Processus Multivariés

Les processus multi-échelles $x(u)$ rencontrés dans de nombreux domaines sont souvent multivariés dans le sens où ils sont indexés par une collection de variables $u = (u_1, \ldots, u_d)$ appartenant chacune à un certain espace. Nous considérerons deux types de processus multivariés pour fournir deux extensions différentes des processus univariés.

Le premier type considère $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$ appartenant à un réseau d-dimensionnel. Cela inclue par exemple les processus bidimensionnels, comme les surfaces de fracture [Lakhal, 2023], ou les champs tridimensionnels en physique, comme les champs de matière noire [Villaescusa-Navarro, 2020]. Dans ce cas, l'espace $\mathbb{R}^d$ est naturellement équipé de la norme euclidienne et nous disposons des outils de traitement du signal tels que la transformée de Fourier ou la transformée en ondelettes. Ce cas est abordé dans la section 1.11.1.

La deuxième extension possible vers des processus multivariés consiste à considérer un processus x décrit comme une collection de processus temporels, $x(t) = (x_1(t), \ldots, x_C(t))^T$ avec $t \in \mathbb{R}$ une variable temporelle, mais avec c une variable arbitraire servant à indexer les séries temporelles (c pour *canal*). Par exemple, un indice financier est composé de plusieurs actions c ayant chacune leur propre prix évoluant au cours du temps t. Alors que les plus proches voisins

d'un temps discret t sont $t-1$ et $t+1$, quels sont les plus proches voisins d'un indice de canal c ? Ce cas est abordé dans la section 1.11.2.

Afin de construire des modèles compacts, nous discutons et énonçons brièvement une notion de régularité multivariée sur laquelle nous nous appuyons pour construire des modèles compacts de x.

1.11.1 Champs physiques

Les écoulements turbulents sont des exemples importants de champs physiques, régis par les équations de Navier-Stokes. Dans son travail pionnier en 1941, Kolmogorov [Kolmogorov, 1941a ; Kolmogorov, 1 941 ; Kolmogorov, 1941b] introduit un modèle gaussien auto-similaire de turbulence qui prévoit que la projection du champ de vitesse sur une ligne est un processus stationnaire dont le spectre de puissance décroît selon une loi de puissance d'exposant de 2/3.

Intermittence dans les écoulements turbulents. Les écoulements turbulents sont hautement non-gaussiens, et la théorie initiale de Kolmogorov a ensuite été affinée pour rendre compte de l'intermittence, qui est mise en évidence par la multifractalité du champ [Kolmogorov, 1962 ; Frisch, 1991]. L'une des principales questions a été d'interpréter et d'inclure l'intermittence des écoulements turbulents dans un modèle. L'importance des dépendances d'échelle pour expliquer l'intermittence remonte aux modèles de turbulence appelés "shell models" [Lorenz, 1963 ; Desnianskii, 1974 ; Siggia, 1978]. Ils consistent à modéliser une turbulence par une équation similaire à celle de Navier-Stokes dans chaque "octave shell", qui sont des régions dyadiques dans le domaine de Fourier, avec des termes d'interaction entre les "shell" voisines [Parisi, 1985].

Ondelettes bidimensionnelles et dépendances angulaires. On s'intéresse ici aux champs physiques comme des processus multivariés indexés par u appartenant à $\mathbb{R}^d$, avec $d = 2$. Dans ce cas, les filtres d'ondelettes univariées $\psi_\lambda(t)$ mentionnés ci-dessus peuvent être étendus aux ondelettes multivariées $\psi_{j,\theta}(u)$ qui sont également localisées à la fois dans l'espace et dans le domaine de Fourier. La transformée en ondelettes $Wx(u, j, \theta) = x \star \psi_{j,\theta}(u)$ extrait les variations de x autour de u à l'échelle 2^j et dans la direction $e_\theta = (\cos\theta, \sin\theta)$.

Les dépendances angulaires sont cruciales pour caractériser un certain nombre de propriétés non-gaussiennes telles que la présence de tourbillons dans des champs turbulents ou de filaments dans des champs cosmologiques. Par exemple, un filament dans un champ produit généralement des coefficients d'ondelettes dont les amplitudes sont grandes sur plusieurs échelles dans la direction orthogonale au filament mais sont petites dans la direction du filament. La construction d'une description statistique Φ caractérisant les dépendances angulaires a été étudiée en physique [Allys, 2020 ; Brochard, 2022 ; Zhang, 2021]. Les auteurs envisagent une extension des covariances harmoniques de phase (1.8) examinées dans la section 1.9.3 corrélant désormais différents harmoniques de phase à différentes positions et à différentes échelles orientées (différentes échelles et angles). Cependant, cette représentation contient un nombre encore plus grand de coefficients, en particulier le nombre de coefficients M peut dépasser le nombre d'échantillons d'une seule réalisation de champ, avec le risque de reconstruire des parties du signal observé comme

remarqué dans [Brochard, 2022], car l'estimation est trop difficile. De tels modèles mettent en évidence un compromis biais-variance déséquilibré en faveur de la grande variance du modèle.

L'une des questions importantes est de construire un modèle compact de champs physiques qui capture l'intermittence, la présence de structures telles que les tourbillons, et les propriétés multifractales d'un processus observé à partir d'une seule réalisation.

Notre contribution. Afin de capturer les dépendances d'échelle et d'angle à partir de données limitées, nous introduisons dans le chapitre 3 une extension des Spectres de Diffusion aux champs physiques en turbulence ou en cosmologie. De manière similaire à (1.20), ils approchent de manière parcimonieuse les matrices de corrélation non linéaires $\mathbb{E}\{Wx\, Wx\}, \mathbb{E}\{Wx\, |Wx|\}, \mathbb{E}\{|Wx|\, |Wx|\}$ qui corrèlent désormais différentes positions spatiales u, u' et différentes échelles orientées λ, λ'. Encore une fois, cela est possible grâce à une approximation diagonale après une seconde transformée en ondelettes.

L'une des principales contributions de ce chapitre est de fournir des modèles à maximum d'entropie pour une variété de champs physiques à partir des spectres en scattering, qui produisent des réalisations de champ visuellement convaincantes et qui capturent des moments d'ordre élevé étudiés en cosmologie tels que le bi-spectre, le tri-spectre, mais aussi les fonctions de structure jusqu'à l'ordre 4. Des réalisations de ces modèles estimés à partir d'une seule observation sont présentés dans la figure 1.5.

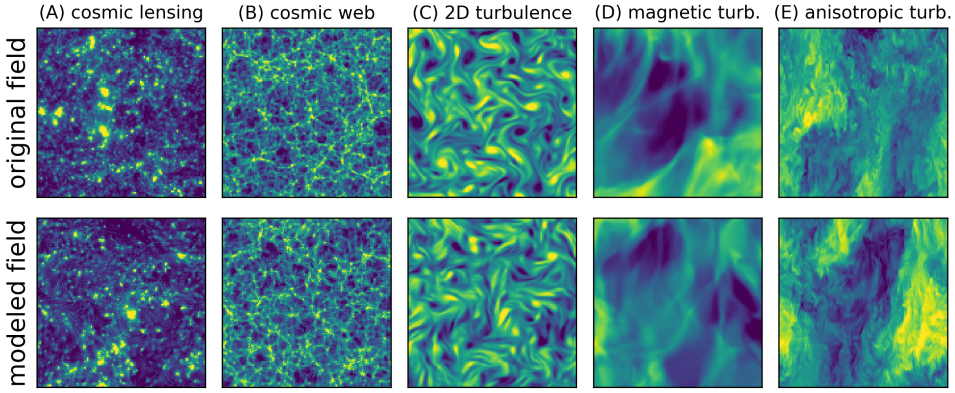


FIGURE 1.5 – Modèles de champs physiques à partir de leurs spectres en scattering. (Haut) Champ observé. (Bas) Échantillon de notre modèle estimé sur une seule réalisation.

1.11.2 Processus temporels multi-canaux

Un processus temporel multi-canal $x(t) = (x_1(t), \dots, x_C(t))^T$ est d'une nature différente des processus discutés dans la section précédente. L'exemple le plus important d'un tel processus dans cette thèse est le prix de différentes actions c du même indice financier. Ces actions ne peut pas être disposées sur une ligne pouvant être traversée de gauche à droite. Cela signifie que $x(t) \in \mathbb{R}^C$ où $\mathbb{R}^C$ n'est pas équipé de la norme euclidienne, la véritable notion de distance, si elle existe, est difficilement accessible.

Matrices de corrélation. La construction de modèles de tels processus nécessite de capturer les dépendances entre différents canaux. La matrice de corrélation entre canaux $\Sigma = \mathbb{E}\{x(t)^T x(t)\}$ contient des informations importantes telles que les secteurs financiers, par exemple industriel, pharmaceutique. Cependant, construire un modèle reposant uniquement sur la matrice de corrélation pour modéliser les dépendances entre canaux présente deux problèmes. Premièrement, la matrice de corrélation ne caractérise pas les propriétés non-gaussiennes entre canaux telles que la présence de crises localisées dans le temps au même moment à travers tous les actifs. Plus généralement, des propriétés non-gaussiennes entre canaux ont été mises en évidence par des statistiques non linéaires en étudiant l'écart des copules de paires d'actions par rapport aux copules gaussiennes [Chicheportiche, 2014b]. Deuxièmement, l'estimation de la matrice de corrélation sur des données limitées reste un défi [Potters, 2005 ; Tumminello, 2007]. Contrairement à une matrice de corrélation temporelle d'un processus stationnaire, nous ne connaissons pas de base prédéfinie, telle que la base de Fourier, qui diagonaliserait la matrice de corrélation entre canaux, ni ne connaissons une transformée en ondelettes entre canaux pour quasi-diagonaliser celle-ci. En particulier, les méthodes basées sur les corrélations d'ondelettes non linéaires telles que [Régaldo-Saint Blancard, 2023] qui ont été réalisées pour 3 ou 5 canaux ne peuvent pas être étendues à un processus avec $C = 253$ actions observé à partir d'une seule réalisation car elles produiraient beaucoup plus de coefficients que la taille N d'une seule réalisation.

Modèles à facteurs. Les modèles à facteurs identifient quelques directions $w^1, \dots, w^r \in \mathbb{R}^C$ le long des canaux et se concentrent sur la modélisation de la structure temporelle du processus univarié $\langle w, x\rangle$ projeté le long de ces directions $\langle w, x\rangle(t) = \sum_{c=1}^{C} w_c x_c(t)$ appelé facteur. Ces modèles se concentrent sur les quelques facteurs dont la structure stochastique régit la structure stochastique jointe. L'hypothèse implicite faites par de tels modèles est que le processus x peut être bien décrit par un nombre raisonnable de facteurs. Bien qu'ils capturent des dépendances non linéaires importantes entre actions, les modèles dans la littérature font souvent des hypothèses simplificatrices sur la structure stochastique des facteurs, par exemple ils ne capturent pas l'asymétrie temporelle jointe des actions x [Reigneron, 2011].

Le principal défi est de trouver un petit nombre de facteurs, dont la structure stochastique se doit d'être modélisée avec précision, afin de modéliser précisément le processus joint x.

Notre contribution. Dans le chapitre 4, nous introduisons un modèle à facteurs à maximum d'entropie qui modélise la structure stochastique de chaque facteur sélectionné par le biais des spectres en scattering introduits dans le chapitre 2. Il consiste à choisir un vecteur de statistiques $\Phi(x) = (\Phi_{\text{single}}(x), \Phi_{\text{cross}}(x))$ où $\Phi_{\text{single}}(x) = \langle \Phi(x_c) \rangle_c$ caractérise la structure stochastique moyenne des actions prises individuellement, tandis que $\Phi_{\text{cross}}(x)$ caractérise les dépendances entre canaux à travers des facteurs sélectionnés

$$\Phi_{\text{cross}}(x) = \Big(\Phi(\langle w^1, x\rangle), \dots, \Phi(\langle w^r, x\rangle) \Big). \tag{1.21}$$

Nous montrons que prendre $r = 10$ directions parcimonieuses obtenues via l'apprentissage de dictionnaires fournit un modèle qui reproduit les principales propriétés non-gaussiennes à travers les actions, y compris l'asymétrie temporelle révélée par un moment d'ordre 3. Le vecteur de statistiques (1.21) contient seulement $r+1 = 11$ fois plus de statistiques que dans le cas univarié, alors que le processus est $C = 253$ fois plus grand en taille. Notre modèle repose donc fortement sur la régularité implicite selon laquelle seuls quelques facteurs régissent le processus, du moins lorsqu'ils s'agit de capturer les statistiques présentes dans la littérature.

1.12 Séparation de sources sur Mars

La séparation de sources non-supervisée est un exemple de problème inverse. Dans un cadre simplifié, cela consiste à récupérer les signaux sources $s, n \in \mathbb{R}^N$ de l'observation du signal mélangé $x = s + n$ sans accès à des exemples d'entraînement séparés. C'est un problème mal défini qui nécessite des connaissances préalables sur les sources. Dans certains cas, n est supposé être un signal bruité, c'est-à-dire un signal multi-échelle avec certaines propriétés auto-similaires, et ce problème peut être considéré comme du débruitage.

Les méthodes classiques de séparation de sources basées sur le traitement du signal [Cardoso, 1989 ; Jutten, 1991 ; Nandi, 1996 ; Cardoso, 1998 ; Starck, 2004 ; Jutten, 2004 ; Bobin, 2007], bien qu'ayant été largement étudiées et comprises, reposent souvent sur des hypothèses excessivement restrictives concernant les sources, par exemple, le fait que les sources ont des distributions gaussiennes ou de Laplace, ce qui peut biaiser négativement le résultat de la séparation de sources [Cardoso, 1998 ; Parra, 2003].

D'autre part, les méthodes non-supervisées de séparation de sources par apprentissage profond [Févotte, 2009 ; Drude, 2019 ; Wisdom, 2020 ; Liu, 2022 ; Denton, 2022 ; Neri, 2021] ne reposent pas sur l'existence de données d'entraînement étiquetées et tentent plutôt d'inférer les sources en se basant sur les propriétés des signaux observés. Ces méthodes font des hypothèses minimales sur les sources sous-jacentes, ce qui en fait un choix adapté pour des problèmes réalistes de séparation de sources. Malgré leur succès, les méthodes non-supervisées de séparation de sources nécessitent souvent une quantité énorme de données pendant l'entraînement [Wisdom, 2020], ce qui est souvent irréalisable dans certaines applications telles que les problèmes liés aux missions spatiales planétaires, par exemple en raison des défis liés à l'acquisition de données. De plus, des préoccupations de généralisation excluent l'utilisation de méthodes basées sur les données entraînées sur des données synthétiques dans les applications réelles en raison des différences entre les données synthétiques et réelles.

Des travaux récents exploitent la capacité des représentations en ondelettes Φ à décrire avec précision les propriétés statistiques des signaux multi-échelles non-gaussiens [Regaldo-Saint Blancard, 2021]. En supposant que l'on a accès à des réalisations indépendantes $n_1, \ldots, n_K$ du processus n, l'idée est de définir un candidat $\widetilde{n}$ qui résout les contraintes statistiques spécifiées par Φ. Le signal candidat $\widetilde{n}$ est obtenu par descente de gradient initialisée à x, le signal mélangé qui contient des informations précieuses sur le signal.

Dans de nombreux cas d'intérêt, tels que la sismologie extra-terrestre, la non-stationnarité

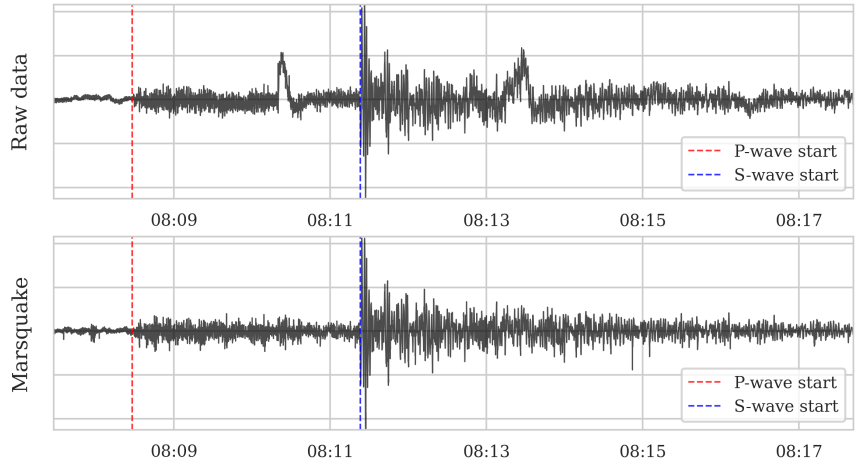


FIGURE 1.6 – Séparation des tremblements de Mars à partir de données limitées grâce aux spectres de diffusion.

des données empêche de tirer suffisamment de réalisations indépendantes de signaux propres $n_1, \ldots, n_K$, et le défi est de construire un modèle préalable de n à partir de données limitées, ce qui est précisément l'objectif de cette thèse.

Dans le chapitre 5 nous considérons des données sismiques enregistrées lors de la mission InSight de la NASA sur Mars. Nous proposons de relever ce défi en utilisant notre représentation en spectre de scattering Φ introduite dans le chapitre 2 qui peut être estimée sur des données limitées. En intégrant une telle description statistique dans le même cadre que [Regaldo-Saint Blancard, 2021], nous avons été en mesure d'utiliser seulement 50 réalisations du bruit sismique de fond sur Mars pour séparer les événements transitoires tels que les glitchs, des données sismiques de Marsquake, voir Fig. 1.6.

1.13 Prédiction sur données limitées

La prédiction, dans le contexte des séries temporelles, consiste à régresser des quantités $Q(\widetilde{x}_{\text{future}}) \in \mathbb{R}^{M_{\text{future}}}$ du futur $\widetilde{x}_{\text{future}}$ à partir d'un passé observé donné $\widetilde{x}_{\text{past}}$. Avec un pour critère la minimisation de l'erreur quadratique moyenne, cela revient à estimer l'espérance conditionnelle suivante

$$\mathbb{E}\{Q(x_{\text{future}}) \mid x_{\text{past}} = \widetilde{x}_{\text{past}}\}. \tag{1.22}$$

Ce problème se pose notamment en finance où le processus de prix x est multi-échelle avec des dépendances à long terme. On peut penser à la prédiction de la variance future, où Q est une moyenne de carrés, ou au prix des options où Q est le paiement d'une option d'achat [2].

Les modèles de régression linéaire proposent d'identifier les prédicteurs $h(x_{\text{past}}) \in \mathbb{R}^{M_{\text{past}}}$ qui "corrèlent" le plus avec $Q(x_{\text{future}})$. En mettant de côté la recherche des meilleurs prédicteurs $h(x_{\text{past}})$ [Ghysels, 2006 ; Christiansen, 2012], cela nécessite d'estimer $M_{\text{past}} \times M_{\text{future}}$ coefficients de corrélation sur des données limitées. Afin d'éviter le sur-apprentissage, les meilleures méthodes

2. Dans ce cas, l'espérance $\mathbb{E}$ est sous la mesure risque-neutre (voir chapitre 6).

linéaires se concentrent sur la prédiction de quelques quantités, par exemple $M_{\text{future}} = 1$, avec quelques prédicteurs bien identifiés $h(x_{\text{past}})$ par exemple $M_{\text{past}} \leq 4$ [Guyon, 2022].

Les méthodes à noyau non locales contournent l'estimation des coefficients de corrélation en moyennant sur des données observées. Elles attribuent des poids aux données observées $(x^i_{\text{past}}, x^i_{\text{future}}), 1 \leq n$ en fonction de leur proximité avec $\widetilde{x}_{\text{past}}$ définie par un noyau k, par exemple un noyau gaussien, et effectuent une moyenne pondérée

$$\bar{Q}(\widetilde{x}_{\text{future}}) = \sum_{i=1}^{n} w_i k(\widetilde{x}_{\text{past}}, x^i_{\text{past}}) Q(x^i_{\text{future}}). \tag{1.23}$$

Cela s'appelle un estimateur de Nadaraya–Watson [Nadaraya, 1964 ; Watson, 1964], sous certaines hypothèses sur le processus x c'est un estimateur non biaisé de (1.22) lorsque le nombre de données $n \to +\infty$ et que le noyau se concentre autour de $\widetilde{x}_{\text{past}}$ [Hansen, 2008].

Ces méthodes sont non locales dans le sens où il n'y a aucune raison pour que les données les plus similaires x^i soient proches de $\widetilde{x}_{\text{past}}$ dans le temps. Pour le débruitage d'image, les méthodes non locales [Buades, 2011] exploitent la même idée. Afin de débruiter le pixel central d'un patch, leur algorithme compare le patch avec tous les autres patches de l'image en fonction de leur norme euclidienne. Une estimation du pixel débruité est obtenue par une moyenne pondérée sur les pixels centraux des patches collectés. Cet algorithme obtient de bons résultats de débruitage, montrant que l'algorithme parvient à trouver des patches similaires, parmi les quelques patches disponibles dans une seule image, qui sont suffisamment informatifs pour débruiter le pixel actuel.

Une telle méthode échouerait pour les processus en finance, qui sont très bruités. En effet, il y a très peu de chances que différentes réalisations courtes de log-rendements financiers soient proches les unes des autres, à partir de données limitées. Cela signifie que le volume de l'ensemble des trajectoires avec une probabilité élevée est très grand, c'est-à-dire qu'un processus de prix financier est un processus avec une entropie élevée.

En optant pour une méthode à noyau non locale, l'objectif est d'exploiter la régularité du processus afin de pouvoir trouver suffisamment de chemins proches x^i avec un pouvoir prédictif sur $\widetilde{x}_{\text{past}}$ malgré l'entropie élevée du processus.

Dans le chapitre 6 nous proposons de rechercher des chemins x^i dans un ensemble de données de chemins générés à partir d'un modèle basé sur les spectres en scattering de x, que nous appelons "path shadowing", illustré dans la Fig. 1.7. Cela est inspiré de l'étude des processus chaotiques. Intuitivement, la propriété de "shadowing" [Hammel, 1987] stipule qu'un chemin qui est uniformément proche d'une vraie orbite d'un système dynamique restera proche (ombré) d'un vrai chemin pour toujours.

Dans notre cas, l'entropie élevée du processus est à la fois une bénédiction et un malédiction. Elle permet de construire des modèles à maximum d'entropie à base de spectres en scattering qui approchent correctement la vraie distribution p. Cependant, nous devons parcourir beaucoup de chemins pour trouver des "shadowing paths". Nous proposons de tirer parti de l'invariance d'échelle du processus pour rendre cette étape réaliste. Pour cela, nous choisissons un noyau basé sur une représentation causale multi-échelle h du passé x_{past} qui prend en compte les données

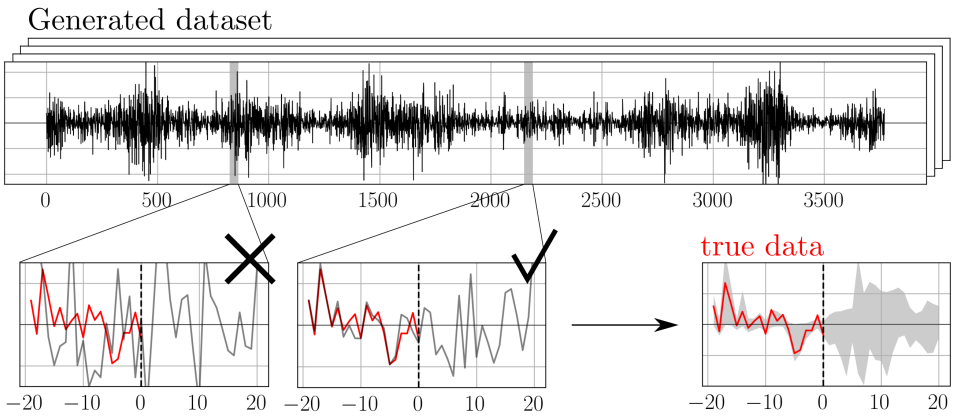


FIGURE 1.7 – Illustration de l'ombrage de chemin pour la prédiction sur données limitées. Étant donné un chemin passé observé (en rouge), nous calculons une moyenne des quantités futures sur des chemins similaires dans le passé, appelés chemins d'ombrage.

passées à courte et longue portée avec seulement quelques paramètres [Renaud, 2003 ; Renaud, 2005 ; Andreux, 2018]. Nous choisissons également un noyau invariant par changement d'échelle $x \mapsto \lambda x$ pour $\lambda > 0$ et par dilatation $x(t) \mapsto x(\lambda^{-1}t)$.

La méthode de "Path Shadowing Monte-Carlo" fournit des résultats de prédiction de volatilité de pointe et peut être utilisé pour obtenir des "smiles" d'options dont la qualité est évaluée à travers un jeu de trading.

1.14 Plan de la thèse

Les trois premiers chapitres 2,3 et 4 sont consacrés à la construction de modèles à maximum d'entropie de processus multi-échelles pouvant être estimés sur des données limitées, en spécifiant le vecteur de statistiques $\Phi(x)$ qui doit être imposé.

Les modèles de processus non-gaussiens univariés nécessitent de caractériser les dépendances à travers différentes échelles. Dans le chapitre 2 nous proposons de partir de la matrice de corrélation non linéaire temps-échelle (1.7). Premièrement, nous étudions les propriétés auto-similaires de cette matrice et montrons qu'elle peut être utilisée pour définir une définition de l'auto-similarité au sens large. Deuxièmement, nous montrons comment réduire le nombre de coefficients en effectuant une approximation diagonale après une deuxième transformée en ondelettes, ce qui donne les spectres en scattering. Un modèle à maximum d'entropie basé sur ces moments est évalué pour la turbulence unidimensionnelle et les séries temporelles financières.

Les modèles de champs physiques – pour des processus bidimensionnels ou tridimensionnels – nécessitent de caractériser les dépendances à travers des échelles orientées. Dans le chapitre 3 nous proposons un vecteur étendu de statistiques $\Phi(x)$, toujours appelé Spectres de Diffusion, qui caractérisent désormais les dépendances entre des échelles orientées. Nous montrons dans ce chapitre que les dépendances régulières en échelles ou en angles peuvent être exploitées par seuillage dans une base de Fourier, ce qui réduit le nombre de coefficients de ces spectres en scattering. Nous proposons des modèles compactes de champs physiques bidimensionnels en

turbulence et en cosmologie.

Le chapitre 4 se concentre sur les séries temporelles multivariées dans le cas spécifique des différentes actions d'un indice financier. Ce chapitre aborde le problème de caractériser les dépendances entre séries temporelles avec le moins de coefficients possible en modélisant la structure temporelle le long de directions sélectionnées.

Dans les chapitres restants 5 et 6, nous proposons deux applications rendues possibles par de tels modèles. Le chapitre 5 aborde les problèmes inverses via la séparation de sources non-supervisée sur des données limitées. Nous intégrant notre modèle dans une méthode existante et des résultats sont présentés sur des données sismiques d'une mission spatiale sur Mars. Le chapitre 6 aborde le problème de la prédiction sur des données limitées. Nous introduisons une méthode à noyau non locale appelée "Path Shadowing Monte-Carlo". Elle repose fortement sur un modèle génératif du processus sous-jacent pour produire un ensemble diversifié de chemins avec les bonnes dépendances temporelles.

Le travail de cette thèse a abouti à sept articles, parmi lesquels quatre articles soumis à des journaux [Morel, 2022 ; Cheng, 2024 ; Morel, 2023b ; Aubrun, 2024], un article de conférence publié [Siahkoohi, 2023b], un article de conférence soumis [Siahkoohi, 2023a] et un article en préparation [Morel, 2023a] :

— **Rudy Morel**, Gaspar Rochette, Roberto Leonarduzzi, Jean-Philippe Bouchaud, Stéphane Mallat. "*Scale Dependencies and Self-Similar Models with Wavelet Scattering Spectra*". Soumis à un journal, 2022.

— Sihao Cheng, **Rudy Morel**, Erwan Allys, Brice Ménard, Stéphane Mallat. "*Scattering Spectra Models for Physics*". Soumis à un journal, 2023.

— Ali Siahkoohi, **Rudy Morel**, Maarten de Hoop, Erwan Allys, Grégory Sainton, Taichi Kawamura. "*Unearthing InSights into Mars : Unsupervised Source Separation with Limited Data*". International Conference on Machine Learning, 2023.

— Ali Siahkoohi, **Rudy Morel**, Randall Balestriero, Erwan Allys, Grégory Sainton, Taichi Kawamura, Maarten de Hoop. "*Martian time-series unraveled : A multi-scale nested approach with factorial variational autoencoders*". Soumis à une conférence, 2023.

— **Rudy Morel**, Stéphane Mallat, Jean-Philippe Bouchaud. "*Path Shadowing Monte-Carlo*". Soumis à une conférence, 2023.

— **Rudy Morel**, Stéphane Mallat, Jean-Philippe Bouchaud. "*A maximum entropy factor model of financial stocks*". En préparation, 2023.

— Cécilia Aubrun*, **Rudy Morel***, Michael Benzaquen, Jean-Philippe Bouchaud. "*Riding wavelets : A method to discover new classes of price jumps, 2024*".

.

Chapitre 2

Models of univariate time-processes : scale dependencies through Scattering Spectra.

Foreword

The first type of processes that we tackle in this chapter are univariate processes, which depend on a single time variable, that could also be a space variable. We introduce the wavelet Scattering Spectra which provide non-Gaussian models of time-processes having stationary increments. A complex wavelet transform computes signal variations at each scale. Dependencies across scales are captured by the joint correlation across time and scales of complex wavelet coefficients and their modulus. This correlation matrix is nearly diagonalized by a second wavelet transform, which defines the Scattering Spectra. We show that this vector of moments characterizes a wide range of non-Gaussian properties of multi-scale processes. This is analyzed for a variety of processes, including fractional Brownian motions, Poisson, multifractal random walks and Hawkes processes. We prove that self-similar processes have Scattering Spectra which are scale invariant. This property can be tested statistically on a single realization and defines a class of wide-sense self-similar processes. We build maximum entropy models conditioned by Scattering Spectra coefficients, and generate new time-series with a microcanonical sampling algorithm. Applications are shown for highly non-Gaussian financial and turbulent time-series.
This chapter is adapted from the following submitted paper. Rudy Morel, Gaspar Rochette, Roberto Leonarduzzi, Jean-Philippe Bouchaud, Stéphane Mallat. Scale Dependencies and Self-Similar Models with Wavelet Scattering Spectra, 2022

Contents

2.1 Introduction

Time-series having stationary increments with variations on a wide range of scales are encountered in physics, Finance, biology, medicine and many other fields. Such multi-scale time-series typically include complex intermittent phenomena with local bursts of activity, and time-asymmetries due to some form of causality. The importance of this topic was first recognized by Mandelbrot [Mandelbrot, 1968 ; Mandelbrot, Mandelbrot, 1982] and led to a considerable body of work on multifractal signals [Bacry, 1993 ; Muzy, 1994 ; Abry, 2000 ; Jaffard, 2004 ; Jaffard, 2006 ; Wendt, 2009 ; Leonarduzzi, 2018]. Among multi-scale processes, self-similar models have a probability distribution which is invariant to scaling, up to multiplicative factors. To validate numerically such models, it is however necessary to introduce weaker forms of self-similarity that can be estimated over limited data.

Simplified multi-scale models have been introduced from marginal distributions of signal increments, by Frisch and Parisi [Frisch, 1985]. They define a weak form of self-similarity from a scale invariance of high order moments of these marginal distributions. This can be sufficient to detect non-Gaussian distributions. Section 2.2 reviews these models together with the multifractal formalism, which replaces increments by wavelet coefficients. Marginal distributions at each scale are simple to estimate, but they do not capture dependencies of signal variations across scales. These dependencies are crucial to specify many properties, in particular, the existence of transient events, which have particular signatures at multiple scales.

Models of multi-scale distributions can be defined as a maximum entropy distribution conditioned by a vector of moments $\mathbb{E}\{\Phi(x)\}$. If they exist, they have an exponential probability distribution

$$p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \theta, \Phi(x)\rangle}.$$

for $\theta \in \mathbb{R}^M$, where M is the number of moments. Maximum entropy models depend only on the energy vector $\Phi(x)$, which needs to be chosen appropriately. Gaussian processes are maximum entropy models conditioned by first and second order moments.

A central result of this chapter is the construction of Φ, so that $\mathbb{E}\{\Phi(x)\}$ specifies scale dependencies, and provide accurate models of multi-scale time-series. The dimension M of $\Phi(x)$ is much smaller than the dimension of x, so that it can define a consistent mean estimator which converges to $\mathbb{E}\{\Phi(x)\}$ when the dimension of x increases. As a result, maximum entropy models can be estimated from a single realization of x. New samples of x are generated by sampling a microcanonical model, which approximates the macrocanonical model [Bruna, 2019].

A wavelet transform computes multi-scale signal variations. Complex wavelet coefficients carry a complex modulus and a complex phase information. Section 2.3 proves that wavelet coefficients are nearly uncorrelated at different scales, because their phases oscillate at different frequencies. To measure non-linear dependencies across scales, it is tempting to move towards higher order moments [Brillinger, 1965]. This requires to compute many moments with high variance estimators, which gives poor numerical results over limited size time-series. Lower variance estimators have been studied by replacing high order moments with phase harmonics [Mallat, 2020] or by eliminating the phase with a modulus non-linearity [Bruna, 2013 ; Portilla,

2000]. We show that scale dependencies can be captured by correlating wavelet coefficients and their modulus. We prove that self-similar processes yield normalized correlation matrices which are invariant to scaling. Section 2.3.3 derives a definition of wide-sense self-similarity, which is analogous to the definition of wide-sense stationarity, where invariance to translation of correlation matrices is replaced by an invariance to scaling.

Wavelet modulus cross-correlation matrices are too large to be estimated accurately from a single time-series realization. Section 2.4 shows that applying a second wavelet transform defines a scattering covariance matrix which is nearly diagonal. Dependencies across scales are captured by diagonal scattering cross-correlation coefficients, also called scattering cross-spectrum, which can be estimated from a single realization. We shall see that the Scattering Spectra provide an interpretable dashboard which captures non-Gaussian properties, including bursts of activity and time-asymmetries, as well as self-similarity.

Fractional Brownian motions, Poisson processes, multifractal random walks and Hawkes processes are often used as models of multi-scale processes which may or may not be self-similar. Section 2.5 shows that the Scattering Spectra reveal their specific properties. By analyzing the Scattering Spectra of S&P financial time-series and turbulent jets, we show that none of the mathematical models presented captures all properties of these complex time-series.

Section 2.6 defines maximum entropy models conditioned by Scattering Spectra values. We generate time-series according to these models with the microcanonical sampling algorithm in [Bruna, 2019]. We show that these generative models can approximate fractional Brownian motions, multifractal random walks and Hawkes processes but also S&P financial time-series or turbulent jets. The code used in numerical experiments is available at `https://github.com/RudyMorel/scattering_spectra`.

2.2 Multi-scale moments

We consider a multi-scale random process $x(t)$ whose increments are stationary. Gaussian models are maximum entropy models conditioned by first and second order moments. In order to capture non-Gaussian and self-similar properties, one can compute higher order moments of increments. The section reviews the scaling properties of these moments.

2.2.1 Self-similarity and power spectrum

If $x(t)$ is stationary then its increments are stationary but the reverse is not always true. For example, a Brownian motion x has stationary increments but $\mathbb{E}\{x(t)\}$ and $\mathbb{E}\{x^2(t)\}$ depend on t [Pipiras, 2017]. The increments of a random process $x(t)$ at intervals $2^j \in \mathbb{R}^+$ for $j \in \mathbb{R}$ are written

$$\delta_j x(t) = x(t) - x(t - 2^j).$$

The lag 2^j can also be interpreted as a scale parameter. We suppose that $\delta_j x$ is stationary for any $j \in \mathbb{R}$.

Mandelbrot et al. [Mandelbrot, 1997] introduced a strong definition of self-similarity from

the joint distribution of increments. A process x is said to be self-similar [Mandelbrot, 1997] up to a maximum scale 2^J if for all $\ell \geq 0$ there exist real random variables A_ℓ which are log infinitely divisible and independent of x such that

$$\left\{\delta_j x(t)\right\}_{j \leq J, t \leq N} \stackrel{d}{=} A_\ell \left\{ \delta_{j-\ell} x(2^{-\ell} t)\right\}_{j \leq J, t \leq N}. \tag{2.1}$$

This equality is in distribution, which means that joint probability distributions of random variables on the left and right hand-sides are equal for any $(j_1, ..., j_n)$ and $(t_1, ..., t_n)$ with $n > 0$. Increments thus have joint distributions which are invariant to dilation, up to random multiplicative factors. The maximum scale 2^J is called the *integrable scale*. It may be fixed, in that case we assume that $x(t)$ and $x(t-\tau)$ are nearly independent for $\tau \gg 2^J$.

If increments are stationary then their auto-correlation $\mathbb{E}\{\delta_j x(t)\, \delta_{j'} x(t-\tau)\}$ only depends on τ. By renormalizing its Fourier transform along τ, one can mathematically define a generalized power spectrum $P_x(\omega)$ of x [Pipiras, 2017]. The non-stationarity of x appears as a singularity of $P_x(\omega)$, which tends to ∞ at $\omega = 0$. This power spectrum specifies second order moments of increments.

With a scaling argument, one can prove [Pipiras, 2017] that self-similar processes have a power spectrum which is also self-similar and thus has a power-law scaling

$$P_x(\omega) = c_2 \, |\omega|^{-\zeta_2 - 1} \tag{2.2}$$

which is singular at $\omega = 0$.

2.2.2 Increment high order moments

First and second order moments define Gaussian maximum entropy models. In order to build non-Gaussian multi-scale models, one can compute q order moments of increments, if they exist :

$$\forall j \in \mathbb{R} \quad , \quad \mathbb{E}\{|\delta_j x(t)|^q\}. \tag{2.3}$$

For self-similar processes, these multi-scale moments have a power-law scaling

$$\mathbb{E}\{|\delta_j x(t)|^q\} = \widetilde{c}_q \, 2^{j\zeta_q}. \tag{2.4}$$

If x is Gaussian and self-similar then one can verify that ζ_q is linear in q [Pipiras, 2017]. It results that any non-linear dependency of ζ_q as a function of q implies that x is not Gaussian. This was initially proposed by Kolmogorov as a test to detect non-Gaussian properties in turbulent flows. Under appropriate hypotheses, the multifractal theory [Jaffard, 2004] proves that (2.3) specifies the pointwise Holder regularity of x, through a spectrum of singularity.

The moment power-law scaling (2.4) is a weak form of self-similarity which can be tested statistically. On the other hand, the strong self-similarity definition (2.1) is highly restrictive, often not satisfied, and impossible to be tested on a single realization. A.2 shows that the strong distribution self-similarity (2.1) implies the weak moment self-similarity (2.4). This scaling is

simple to test numerically but is a relatively weak characterization of self-similarity. The high order increment moments (2.4) remain unchanged when computed on $x(-t)$, and hence do not detect time-asymmetries. They do not either capture dependencies of increments at different scales 2^j. Section 2.3 introduces a stronger wide-sense definition of self-similarity, which relies on multi-scale moments that depend upon joint time-scale dependencies of x.

2.2.3 Estimation and wavelet transform

Defining consistent estimators of moments with fast convergence is necessary to compute maximum entropy models from a single realization of x. It has been proved that a wavelet transform yields nearly optimal estimators of second order moments for self-similar processes [Flandrin, 1992 ; Wornell, 1993 ; Masry, 1993 ; McCoy, 1996]. We thus replace increments by a wavelet transform, whose properties are briefly reviewed.

2.2.3.1 Wavelet transform

A wavelet $\psi(t)$ has a fast decay away from $t = 0$, polynomial or exponential for example, and a zero-average $\int \psi(t)\,\mathrm{d}t = 0$. We normalize $\int |\psi(t)|^2\,dt = 1$. The wavelet transform computes the variations of a signal x at each scale 2^j with

$$Wx(t, j) = x \star \psi_j(t) \quad \text{where} \quad \psi_j(t) = 2^{-j}\psi(2^{-j}t).$$

If $\psi = \delta(t) - \delta(t-1)$ then it computes signal increments $Wx(t, j) = \delta_j x(t)$. To relate regularity properties of signals from their wavelet coefficients, it is necessary to use wavelets having a better frequency localization than a difference of Diracs [Jaffard, 2004]. We use a complex wavelet ψ having a Fourier transform $\widehat{\psi}(\omega) = \int \psi(t)\, e^{-i\omega t}\,\mathrm{d}t$ which is real, and whose energy is mostly concentrated at frequencies $\omega \in [\pi, 2\pi]$. It results that $\widehat{\psi}_j(\omega) = \widehat{\psi}(2^j\omega)$ is non-negligible mostly in $\omega \in [2^{-j}\pi, 2^{-j+1}\pi]$. We suppose that ψ has $m \geq 1$ vanishing moments, which means that $|\widehat{\psi}(\omega)| = O(|\omega|^m)$ in the neighborhood of $\omega = 0$. We will refer to the modulus and complex phase of $Wx(t, j)$ as the *amplitude* and *phase* of the complex wavelet coefficient.

In the following we shall restrict the scales 2^j to dyadic scales, and hence j to integers. The wavelet transform W satisfies an energy conservation law [Mallat, 1999] specified in A.1, which implies that it is invertible.

All numerical calculations below are performed with a complex Battle-Lemarié wavelet [Battle, 1987 ; Lemarié, 1988], restricted to positive frequencies. Figure 2.1 shows the real and imaginary parts of ψ as well as its Fourier transform. It has an exponential decay away from $t = 0$, it has $m = 4$ vanishing moments and satisfies an energy conservation law (A.1). If the input signal is sampled at $t \in \mathbb{Z}$ then we can only compute wavelet coefficients for $2^j > 1$ and hence $j \geq 1$.

2.2.3.2 Wavelet covariance and spectrum

Since $\int \psi_j(t)dt = 0$ it results that $\mathbb{E}\{x \star \psi_j(t)\} = 0$. If x has stationary increments then one can show [Pipiras, 2017] that wavelet coefficients are jointly stationary. Their covariance across

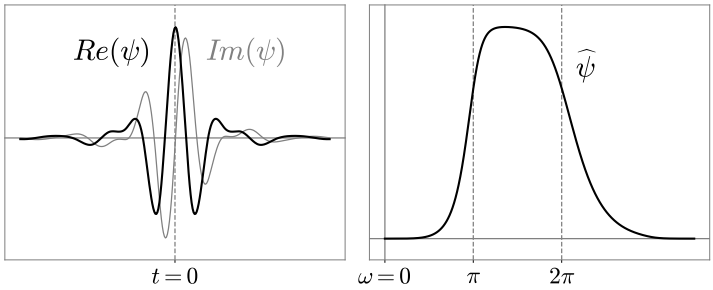


FIGURE 2.1 – Left : complex Battle-Lemarié wavelet $\psi(t)$ as a function of t. Right : Fourier transform $\widehat{\psi}(\omega)$ as a function of ω.

time and scale can be written from the power spectrum of x :

$$\mathbb{E}\{x \star \psi_j(t)\, x \star \psi_{j-a}(t - 2^j\tau)^*\} = \frac{1}{2\pi} \int P_x(\omega)\, \widehat{\psi}(2^j\omega)\, \widehat{\psi}^*(2^{j-a}\omega)\, e^{i\tau 2^j\omega}\, d\omega, \tag{2.5}$$

for time-lag $2^j\tau \in \mathbb{R}$ and scale-lag $a \in \mathbb{Z}$. This covariance becomes negligible when $|a| > 0$ for which the supports of $\widehat{\psi}(\omega)$ and $\widehat{\psi}(2^a\omega)$ barely overlap. Indeed, the phases of $x \star \psi_j$ and $x \star \psi_{j-a}$ vary at different rates, which cancels their correlation. For processes x with a power-law decaying power-spectrum (2.2), one can prove that such covariance has a polynomial decay away from $\tau = 0$ and an exponential decay away from $a = 0$ [Wornell, 1993]. As shown on figure 2.4a, these coefficients are negligible for distant scales $|a| > 1$ and have small non-zero values for $|a| = 1$ because wavelets have a small frequency overlap.

The diagonal covariance values define a *wavelet spectrum* :

$$\sigma_W^2(j) = \mathbb{E}\{|x \star \psi_j(t)|^2\} = \frac{1}{2\pi} \int P_x(\omega)\, |\widehat{\psi}(2^j\omega)|^2\, d\omega. \tag{2.6}$$

It integrates $P_x(\omega)$ over the frequency intervals $[2^{-j}\pi, 2^{-j+1}\pi]$, where $\widehat{\psi}(2^j\omega)$ is mostly supported. It does not depend upon t because of stationarity, and is thus estimated by an empirical average

$$\widetilde{\sigma}_W^2(j) = \left\langle |x \star \psi_j(t)|^2 \right\rangle_t. \tag{2.7}$$

2.2.3.3 Self-similar wavelet coefficients

If x is strongly self-similar according to (2.1) then A.2 derives that

$$\forall \ell \geq 0\ ,\ \left\{x \star \psi_j(t)\right\}_{j \leq J, t \leq N} \overset{d}{=} A_\ell\ \left\{x \star \psi_{j-\ell}(2^{-\ell}t)\right\}_{j \leq J, t \leq N}. \tag{2.8}$$

A.2 also proves that wavelet moments of self-similar processes have the same scaling properties as increments in (2.4). For all q such that the moments are defined, there exists c_q such that

$$\forall j, \quad \mathbb{E}\{|x \star \psi_j|^q\} = c_q\, 2^{j\zeta_q}. \tag{2.9}$$

2.3 Dependencies across scales with phase-modulus wavelet correlations

We saw that wavelet coefficients of stationary processes are nearly uncorrelated across scales. Yet, next section shows that non-Gaussian processes have strong dependencies across scales. Section 2.3.2 captures these dependencies by correlating complex wavelet coefficients and their modulus. Section 2.3.3 shows that self-similar processes have a normalized phase-modulus wavelet correlation matrix which is invariant to scale shift. This invariance defines a wide-sense self-similarity.

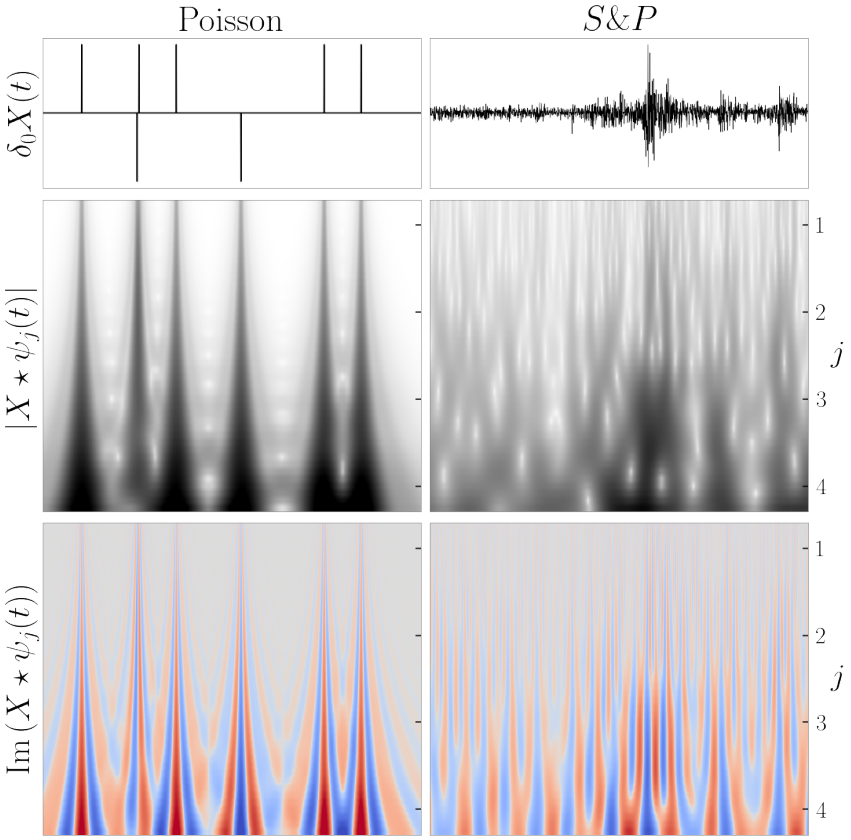


FIGURE 2.2 – Top : increments $\delta_0 x(t)$ of a signed Poisson process and the financial S&P daily increments from 03/01/2000 to 10/10/2018. Middle : wavelet modulus $|x \star \psi_j(t)|$. The vertical axis corresponds to the log-scale index j which is real. Dark color represents large values. These modulus have dependencies across scales produced by Diracs or bursts of activity. Bottom : imaginary part of $x \star \psi_j(t)$. Red and blue colors represent negative and positive values respectively. It shows that localized structures such as Dirac create correlation of phases across octaves, when j increases by 1 or more.

2.3.1 Scale dependencies as a trace of non-Gaussianity

Section 2.2.3 showed that if x has stationary increments then $x \star \psi_j(t)$ and $x \star \psi_{j'}(t')$ are nearly uncorrelated if $|j - j'| > 1$. If x is Gaussian then these coefficients are jointly Gaussian so

it implies that they are independent. On the contrary, we will now see that non-Gaussian time-series exhibits crucial dependency across scales. This dependency provides important information on non-Gaussian properties of x.

Figure 2.2 shows the wavelet transform of S&P financial signal, and of a Poisson process whose increments have a random sign. They are calculated with the complex Battle-Lemarié wavelet. Diracs and bursts of activity in the financial signal create high amplitude wavelet coefficients, which propagate across scales. It induces strong dependencies between wavelet modulus across scales. These dependencies also appear in the wavelet transform phase. Diracs produce high amplitude wavelet coefficients whose phase propagates regularly across octaves, when j increases by 1 or more. On the contrary, financial bursts of activity have a phase that is randomly modified from one octave to the next. Correlations when j increases by less than 1 are due to correlations between the wavelets themselves.

To understand this dependency phenomenon, consider a localized pattern $f(t)$ in the neighborhood of $t = 0$, such as a Dirac. It is randomly translated to define $x(t) = f(t-U)$, where U is a random variable uniformly distributed in $[0, 1]$. Its wavelet coefficients $x \star \psi_j(t) = f \star \psi_j(t-U)$ are centered at $t = U$ at all scales 2^j, and are thus highly dependent. Their amplitude and phase are a signature of the translated pattern f. It illustrates the importance of wavelet coefficient dependencies across scales, for non-Gaussian processes.

2.3.2 Joint phase-modulus correlations across scales

This section introduces a joint wavelet phase-modulus correlation matrix which captures dependencies of wavelet coefficients across scales. Complex wavelet coefficients have a negligible correlation at different scales because they are supported in different frequency bands. Their correlation is thus canceled by phase fluctuations which occur at different rates. We first realign their frequency support with a modulus, and then compute their correlation. Correlations of wavelet coefficient modulus were first studied by Portilla and Simoncelli [Portilla, 2000]. Their properties are analysed in [Mallat, 2020 ; Zhang, 2021]. The joint phase-modulus correlation matrix also preserves phase information, by correlating wavelet coefficients with and without phases. It partly characterizes phase alignments across scales.

2.3.2.1 Non-zero correlations by removing complex phase

Eliminating the phase with a modulus can introduce correlations across scales. Indeed, let us remind that the cross-spectrum $P_{y,z}(\omega)$ of two jointly stationary random processes $y(t)$ and $z(t)$ is the Fourier transform of their cross-correlation

$$P_{y,z}(\omega) = \int \mathbb{E}\{y(t) z(t-\tau)^*\}\, e^{-i\tau\omega}\, d\tau,$$

and the Cauchy-Schwarz inequality proves that

$$|P_{y,z}(\omega)|^2 \leq P_y(\omega)\, P_z(\omega). \tag{2.10}$$

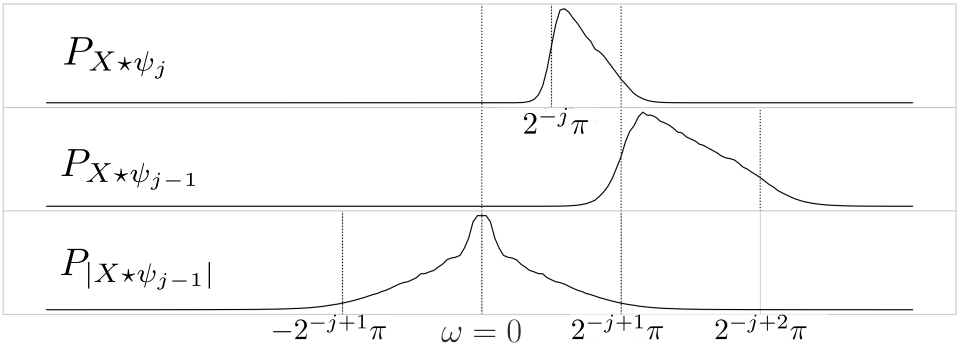


FIGURE 2.3 – Top : power spectrum of $x \star \psi_j$ for the S&P time-series. It is mostly concentrated in $[2^{-j}\pi, 2^{-j+1}\pi]$. Middle : power spectrum of $x \star \psi_{j-1}$. Bottom : power spectrum of $|x \star \psi_{j-1}|$ is mostly concentrated in $[-2^{-j+1}\pi, 2^{-j+1}\pi]$ and strongly overlaps with the power spectrum of $x \star \psi_j$.

The cross-correlation of $y(t)$ and $z(t-\tau)$ is therefore zero if their power spectra do not overlap. Applied to $y = x\star\psi_j$ and $z = x\star\psi_{j-a}$, it verifies once again that they are essentially uncorrelated if $a \neq 0$. Indeed, their power spectrum do not overlap, as illustrated in Figure 2.3. However, we now show that the power spectrum of $y = x \star \psi_j$ and $z = |x \star \psi_{j-a}|$ or of $y = |x \star \psi_j|$ and $z = |x \star \psi_{j-a}|$ can overlap. They can thus have non-zero correlations, after suppressing their mean.

The power spectrum $P_x(\omega)|\widehat{\psi}(2^{j-a}\omega)|^2$ of $x \star \psi_{j-a}$ has a support mostly concentrated in $[2^{-j+a}\pi, 2^{-j+a+1}\pi]$. A modulus eliminates the phase of $x \star \psi_{j-a}$ which oscillates at the center frequency $3 \times 2^{-j+a-1}\pi$. As a consequence, the power spectrum of $|x\star\psi_{j-a}|$ is centered at $\omega = 0$, and its energy is mostly concentrated in $[-2^{-j+a}\pi, 2^{-j+a}\pi]$ [Mallat, 2020 ; Zhang, 2021]. This is shown by Figure 2.3. It results that the spectrum of $x \star \psi_j$ and $|x \star \psi_{j-a}|$ do overlap if $a > 0$, and the spectrum of $|x \star \psi_j|$ and $|x \star \psi_{j-a}|$ overlap for any a since they are both centered at $\omega = 0$.

2.3.2.2 Joint phase-modulus correlation matrix

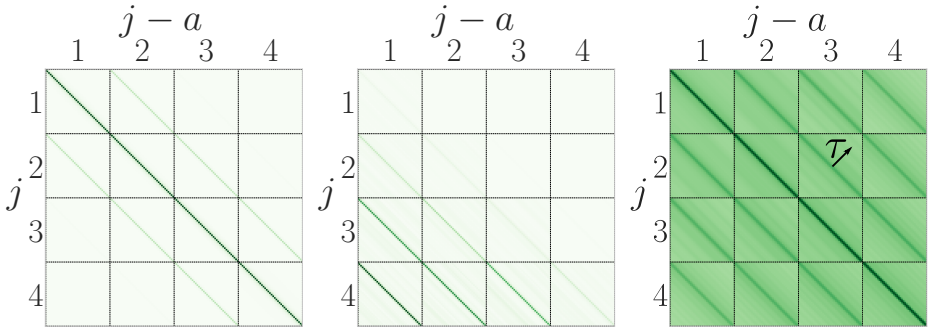
Let us write $\rho(z) = (z, |z|)$ for any $z \in \mathbb{C}$. We now show how to represent the dependencies of wavelet coefficients from joint phase-modulus correlation matrix of

$$\rho W x = (Wx, |Wx|) = \Big(x \star \psi_j(t), |x \star \psi_j(t)|\Big)_{t,j}.$$

If x is sampled at intervals normalized to 1 and is of size T then we compute wavelet coefficients over $\log_2 N$ scales $1 < 2^j \leq N$ corresponding to $1 \leq j \leq \log_2 N$. If x is stationary then $\mathbb{E}\{\rho W x\} = (\mathbb{E}\{Wx\}, \mathbb{E}\{|Wx|\})$ does not depend upon t. Without loss of generality we suppose that $\mathbb{E}\{x(t)\} = 0$ so $\mathbb{E}\{Wx(t)\} = 0$.

The coefficients of the joint phase-modulus correlation matrix $\mathbb{E}\{\rho W x\,(\rho W x)^*\}$ are

$$\Big(\mathbb{E}\{\rho W x(t,j)\,\rho W x(t - 2^j\tau, j-a)^*\}\Big)_{t,\tau,j,a}.$$



(a) $\mathbb{E}\{Wx\,Wx^T\}$ (b) $\mathbb{E}\{Wx\,|Wx|^T\}$ (c) $\mathbb{E}\{|Wx|\,|Wx|^T\}$

FIGURE 2.4 – Modulus of the joint phase-modulus correlation $\mathbb{E}\{\rho Wx\,(\rho Wx)^T\}$ for S&P signal. The diagonal of this matrix ($\tau = 0, a = 0$) is the wavelet spectrum $\sigma_W^2(j)$ which is not constant. To remove such normalization effect, we plot the matrix where each coefficient (t, τ, j, a) is divided by $\sigma_W(j)\sigma_W(j-a)$. For the 3 matrices, each subblock is a Toeplitz correlation matrix along time $(t, t - 2^j\tau)$, for scales $(j, j-a)$ fixed, because of time stationarity. All correlation values are constant when j varies, which is a mark of wide-sense self-similarity.

They do not depend upon t because x is stationary. We estimate them from a single realization of x with a time average

$$\left\langle \rho Wx(t,j)\,\rho Wx(t-2^j\tau, j-a)^* \right\rangle_t .$$

The correlation matrix $\mathbb{E}\{\rho Wx\,(\rho Wx)^T\}$ is composed of four submatrices

$$\begin{pmatrix} \mathbb{E}\{Wx\,Wx^T\} & \mathbb{E}\{Wx\,|Wx|^T\} \\ \mathbb{E}\{|Wx|\,Wx^T\} & \mathbb{E}\{|Wx|\,|Wx|^T\} \end{pmatrix}$$

We show that $E\{Wx\,Wx^T\}$ and $\mathbb{E}\{Wx\,|Wx|^T\}$ are sparse matrices. On the other hand, we shall see that $\mathbb{E}\{|Wx|\,|Wx|^T\}$ may not be sparse. The coefficients of the wavelet auto-correlation matrix $\mathbb{E}\{Wx\,Wx^T\}$ are

$$\mathbb{E}\{x \star \psi_j(t)\, x \star \psi_{j-a}(t-2^j\tau)^*\}. \tag{2.11}$$

Section 2.2.3 explains that we can approximate it by its diagonal values which define the wavelet spectrum. For a signal of size T, since $1 \le j \le \log_2 N$ we only estimate $\log_2 N$ wavelet spectrum coefficients $\mathbb{E}\{|x \star \psi_j(t)|^2\}$.

The off-diagonal matrix $\mathbb{E}\{Wx|Wx|^T\}$ is a correlation between complex wavelet coefficients and their modulus, that we shall call *phase-modulus correlation*.

$$\mathbb{E}\{x \star \psi_j(t)\,|x \star \psi_{j-a}(t-2^j\tau)|\} \tag{2.12}$$

Since $\mathbb{E}\{x \star \psi_j(t)\} = 0$ these correlations are also covariance coefficients. Figure 2.4b shows that wavelet phase-modulus correlations are non-negligible for a scale shift $a \ge 0$ and $\tau = 0$. This is expected because the power spectrum of $x \star \psi_j$ and $|x \star \psi_{j-a}|$ have an overlapping support. When $\tau \neq 0$ they become negligible because of random phase fluctuations. Since $1 \le j - a < j \le \log_2 T$, we only estimate $2^{-1}\log_2^2 T$ wavelet phase-modulus cross-spectrum coefficients $\mathbb{E}\{x \star \psi_j(t)\,|x \star \psi_{j-a}(t)|\}$.

The coefficients of the wavelet *modulus auto-correlation* $\mathbb{E}\{|Wx|\,|Wx|^T\}$ are

$$\mathbb{E}\{|x \star \psi_j(t)|\,|x \star \psi_{j-a}(t - 2^j\tau)|\}. \tag{2.13}$$

Figure 2.4c shows that the wavelet modulus correlation is nearly a full matrix. Their covariance is obtained by subtracting the modulus means. These covariances are also a priori non-zero for all scale shift a, because the power spectra of $|x \star \psi_j|$ and $|x \star \psi_{j-a}|$ have overlapping supports. They can be non-negligible for large time shift τ, because all phases have been eliminated. The number of time shifts is nearly equal to the signal size T. For $1 \leq j - a \leq a \leq \log_2 T$, there are about $2^{-1}\,T\,\log_2^2 T$ potentially non-negligible wavelet modulus correlations, which is too large to estimate them directly from a single realization of x. Section 2.4 shows that one can reduce the number of correlation coefficients to $\log_2^3 N$, by applying a second wavelet transform on $|x \star \psi_j(t)|$ before calculating its auto-correlation.

2.3.2.3 Non-Gaussian properties

Phase and modulus wavelet correlations can be analyzed as particular cases of phase harmonic correlations introduced in [Mallat, 2020]. It captures non-Gaussian properties proved in [Zhang, 2021]. The following proposition transpose these results in our context. It proves that the existence of non-negligible wavelet modulus covariances across scales is a mark of non-Gaussianity. We write $\mathrm{Cov}\{A, B\} = \mathbb{E}\{AB^T\} - \mathbb{E}\{A\}\,\mathbb{E}\{B\}^T$.

Proposition 1. *If x is Gaussian and $\widehat{\psi}_j\,\widehat{\psi}_{j-a} = 0$ then for all τ*

$$\mathrm{Cov}\{\rho W x(t, j)\,,\, \rho W x(t - 2^j\tau, j - a)\} = 0.$$

We indeed saw that if $\widehat{\psi}_j\widehat{\psi}_{j-a} = 0$ then $x \star \psi_j(t)$ and $x \star \psi_{j-a}(t - 2^j\tau)$ are uncorrelated (2.5). If x is Gaussian then they are also Gaussian and hence independent. Applying a modulus preserves this independence and thus produces covariance coefficients which remain zero, proving the second equality. The condition $\widehat{\psi}_j\,\widehat{\psi}_{j-a} = 0$ is verified up to a very small error for $|a| > 1$. For $|a| = 1$, the supports of $\widehat{\psi}_j$ and $\widehat{\psi}_{j-a}$ have a small overlap so the product is small but not zero. It follows from the proposition that non-zero covariance coefficients across distant scales evidence that x is not Gaussian.

Time-asymmetry is another form of non-Gaussianity, often produced by causality phenomena. Let R be the time reversal operator $Rx(t) = x(-t)$. A process x is said to be time-reversible if the probability distributions of Rx and x are equal. Gaussian stationary processes are time-reversible. The following proposition shows that time-reversibility can be detected from phase correlation coefficients.

Proposition 2. *If x is time-reversible then the joint wavelet modulus correlation has a Hermitian symmetry along τ :*

$$\mathbb{E}\{\rho W x(t, j)\,\rho W x(t - 2^j\tau, j - a)^*\} = \mathbb{E}\{\rho W x(t, j)^*\,\rho W x(t + 2^j\tau, j - a)\}. \tag{2.14}$$

Time-reversibility means that x and Rx have the same distribution. Since $(Rx) \star \psi_j(t) = x \star \psi_j(-t)^*$ it implies the equality (2.14). This Hermitian symmetry is always satisfied by the wavelet auto-correlation coefficients (2.11) even if x is not time-reversible, but not necessarily by phase-modulus correlations and modulus auto-correlation coefficients (2.12,2.13). If they do not have the Hermitian symmetry (2.14) then x is not time-reversible, and hence non-Gaussian.

2.3.3 Wide-sense self-similarity

Self-similarity is defined in (2.1) and (2.8) on the process distributions, which are high-dimensional objects, on increments or wavelet coefficients. Such properties cannot be tested statistically on a single realization. Section 2.2 gives necessary conditions (2.4,2.9) over the marginals of increments and wavelet coefficients, which are simple to verify but provides relatively weak characterization of self-similarity. The same difficulty appears to test that a process has a stationary distribution. It cannot be tested statistically on a single realization. Conditions on marginals impose that the probability distributions of $x(t)$ for each t does not depend upon t. It can be tested numerically but it is a weak condition. More powerful characterizations of stationarity impose that $\mathbb{E}\{x(t)\}$ and the auto-correlation of x are invariant to time shift. The process x is then said to be wide-sense stationary. We follow the same approach for self-similarity by imposing a scale shift invariance on a normalized joint phase-modulus correlation matrix.

Self-similarity of increments distributions (2.1) or wavelet coefficients (2.8) are defined up to random multiplicative factors A_ℓ. We eliminate these multiplicative factors with a normalized phase-modulus correlation matrix where each correlation coefficient of $\mathbb{E}\{\rho W x\, (\rho W x)^T\}$ is normalized by a product of standard deviations given by $\sigma_W^2 = \mathbb{E}\{|x \star \psi_j(t)|^2\}$:

$$C_{\rho W}(\tau; j, a) = \frac{\mathbb{E}\{\rho W x(t, j)\, \rho W x(t - 2^j\tau, j - a)^*\}}{\sigma_W(j)\, \sigma_W(j - a)}.$$

A wavelet transform introduces explicitly a scaling parameter 2^j. However, the correlations of wavelet coefficients Wx vanish across scales, and thus can not be used directly to identify self-similarity across scales. We define a notion of wide-sense self-similarity as a translation and scale invariance of the mean and correlation matrix of $(Wx, |Wx|)$.

Theorem 1 (Wide-sense self-similarity)**.** *If x is self-similar up to the scale 2^J in the sense of (2.8) then there exist c_1, c_2, ζ_1, ζ_2 such that for all $j \leq J$*

$$\mathbb{E}\{|x \star \psi_j(t)|\} = c_1\, 2^{j\zeta_1}, \tag{2.15}$$

$$\mathbb{E}\{|x \star \psi_j(t)|^2\} = c_2\, 2^{j\zeta_2}. \tag{2.16}$$

and for all τ, $j \leq J$, a,

$$C_{\rho W}(\tau; j, a) \;\; = \;\; C_{\rho W}(\tau; 0, a). \tag{2.17}$$

The theorem is proved in A.3. A process x which satisfies the properties (2.15,2.16,2.17) is said to be wide-sense self-similar. The appendix proves that self-similarity implies wide-sense

similarity. Similarly to moment self-similarity (2.9), wide-sense self-similarity imposes the existence of scaling exponents ζ_q, but only for $q = 1$ and $q = 2$. The scaling exponent ζ_2 specifies the decay of the wavelet spectrum $\sigma_W(j) = \mathbb{E}\{|x \star \psi_j(t)|^2\}$. The ratio between first and second order moments is a sparsity measure on wavelet coefficients

$$s_W(j) = \frac{\mathbb{E}\{|x \star \psi_j(t)|\}}{\sigma_W(j)}. \tag{2.18}$$

If x is wide-sense self-similar then $s_W^2(j) = c_s\, 2^{j\zeta_s} \leq 1$, where $c_s = c_1^2\, c_2^{-1}$ and $\zeta_s = 2\zeta_1 - \zeta_2 \geq 0$. The lower $s_W^2(j)$ the sparser $x \star \psi_j(t)$. The exponent ζ_s is a sparsity rate which governs the increase of sparsity when the scale decreases. If x is Gaussian then $\zeta_s = 0$. If $\zeta_s > 0$ then the sparsity of $x \star \psi_j$ increases as j decreases. The constant c_s is a sparsity multiplicative factor. If x is Gaussian then $x \star \psi_j$ is also Gaussian and one can verify that $c_s = \pi/4$, which is the ratio between first and second order moments of complex Gaussian random variables.

Wide-sense self-similarity also imposes that the normalized phase-modulus correlation matrix $C_{\rho W}$ depends only on time shift τ and scale shift a. This is a powerful second order condition, which is sufficient to reveal existence of important self-similar properties in time-series. It applies to the non-negligible coefficients of each of the three submatrices of $C_{\rho W}$. Over diagonal wavelet auto-correlation coefficients, it is already specified by (2.16). Over non-negligible phase-modulus cross-spectrum coefficients, it imposes that

$$C_{W|W|}(0; j, a) = \frac{\mathbb{E}\{Wx(t, j)\, |Wx(t, j-a)|\}}{\sigma_W(j)\, \sigma_W(j-a)}$$

does not depend upon j. These coefficients are estimated by replacing each expected value by an average on t. For self-similar processes, since these moment do not depend on j, we can further improve this estimation by averaging them over scales. It defines an *scale invariant phase-modulus cross spectrum*

$$\overline{C}_{W|W|}(a) = \left\langle C_{W|W|}(0; j, a) \right\rangle_j. \tag{2.19}$$

For wavelet modulus auto-correlations (2.13), these conditions are translated into conditions over a scattering cross-spectrum which is introduced in the next section.

Processes such as fractional Brownian motion [Mandelbrot, 1968] or multifractal random walk [Bacry, 2001a] are self-similar and therefore wide-sense self-similar. Self-similarity cannot be tested statistically, whereas wide-sense self-similarity is a correlation property which can be estimated. Figure 2.4 shows that the S&P financial signal is wide-sense self-similar. Indeed $\log \mathbb{E}\{|x \star \psi_j(t)|\}$ and $\log \mathbb{E}\{|x \star \psi_j(t)|^2\}$ have a linear decay along j, and the normalized correlation $C_{\rho W}$ is constant along its diagonals when j varies.

2.4 Scattering cross-spectrum

Section 2.3.2 explains that there are too many wavelet modulus auto-correlation coefficients to estimate them from a single realization of x. We introduce a low-dimensional approximation of

this auto-correlation matrix by applying a second wavelet transform, which defines a scattering transform [Mallat, 2012]. The resulting scattering covariance is nearly diagonal and its spectra can be estimated from a single realization.

2.4.1 Diagonal scattering cross-correlation

Section 2.2.3 explains that if x has self-similarity properties then its auto-correlation matrix is well approximated by a diagonal matrix after applying a wavelet transform. Similarly, instead of computing directly the auto-correlation of $|Wx| = \left(|x \star \psi_j(t)|\right)_{t,j}$ we will compute the auto-correlation of its wavelet transform.

Applying a second wavelet transform W on $|Wx|$ defines a scattering transform $Sx = W\,|Wx|$, with

$$Sx(t; j, k) := |x \star \psi_j| \star \psi_k(t).$$

It is non-negligible only if $k > j$. Indeed, the Fourier transform of $|x \star \psi_j|$ is mostly concentrated in $[-2^{-j}\pi, 2^{-j}\pi]$. If $k \leq j$ then it does not intersect the frequency interval $[2^{-k}\pi, 2^{-k+1}\pi]$ where the energy of $\widehat{\psi}_k$ is mostly concentrated, in which case $Sx(t; j, k) \approx 0$.

Since $Sx = W\,|Wx|$, its auto-correlation is

$$\mathbb{E}\{Sx\,Sx^T\} = W\,\mathbb{E}\{|Wx|\,|Wx|^T\}\,W^T, \tag{2.20}$$

which specifies $\mathbb{E}\{|Wx|\,|Wx|\}$ because W is invertible. The coefficients of $\mathbb{E}\{Sx\,Sx^T\}$ are

$$\mathbb{E}\{|x \star \psi_j| \star \psi_k(t)\,|x \star \psi_{j-a}| \star \psi_{k'}(t-\tau)^*\}$$

for all time t, time shift τ, first wavelet scale j and scale shift a, and second wavelet scales k, k'. We impose that $k > j$ and $k' > j - a$ otherwise the scattering coefficients are negligible. Applying (2.10) to $y(t) = |x \star \psi_j| \star \psi_k(t)$ and $z(t) = |x \star \psi_{j-a}| \star \psi_{k'}(t-\tau)$ shows that this correlation is negligible if $k \neq k'$ because the spectra of y and z barely overlap. Indeed $\widehat{\psi}_k$ and $\widehat{\psi}_{k'}$ are concentrated over non-overlapping frequency intervals. Correlations for $k = k'$ have a fast polynomial decay away from $\tau = 0$ [Wornell, 1993] when the cross-spectrum of the modulus is regular, and we thus shall only consider these scattering correlations for $\tau = 0$. Scattering correlations are thus calculated only for $k = k' = j - b$ with $b < 0$ and $\tau = 0$.

We incorporate the normalization of the wavelet modulus auto-correlation by the wavelet spectrum $\sigma_W^2(j) = \mathbb{E}\{|x \star \psi_j(t)|^2\}$ to the scattering correlations. The normalized diagonal scattering coefficients for $k = k' = j - b$ define a *scattering cross-spectrum* whose coefficients are :

$$C_S(j, a, b) = \frac{\mathbb{E}\{|x \star \psi_j| \star \psi_{j-b}(t)\,|x \star \psi_{j-a}| \star \psi_{j-b}(t)^*\}}{\sigma_W(j)\,\sigma_W(j-a)}.$$

Since ψ_{j-b} has a Fourier transform mostly supported in $[2^{-j+b}\pi, 2^{-j+b+1}\pi]$, $C_S(j, a, b)$ can be interpreted as the cross-spectrum of $|x \star \psi_j|$ and $|x \star \psi_{j-a}|$ integrated over this frequency interval. These cross-spectra specify intermittency phenomena which appear when the wavelet modulus correlations (2.13) remain large on a long-range of time. If the modulus are uncorrelated in

time across scales then $C_S(j,a,b)$ is nearly constant along b for j,a fixed. On the contrary, if $C_S(j,a,b)$ has a fast decay in b then it implies that the modulus have long range correlations in time.

If x is of size T then there are at most $\log_2 T$ scales indices j, a and b. it shows that the scattering transform can provide an approximation of $T\,\log_2^2 T$ wavelet modulus auto-correlation coefficients with $\log_2^3 T$ scattering cross-spectrum coefficients.

2.4.2 Properties

The following proposition derives from Proposition 2 that the imaginary part of C_S captures time-asymmetry properties of x. The proof is in A.4.

Proposition 3. *If x is Gaussian and $\widehat{\psi}_j\,\widehat{\psi}_{j-a} = 0$ then*

$$C_S(j,a,b) = 0.$$

If x is time-reversible then for all j, a and b the imaginary part satisfies

$$\operatorname{Im} C_S(j,a,b) = 0.$$

The following theorem proves that the self-similarity condition (2.17) on $C_{\rho W}$ can be evaluated on non-zero scattering coefficients.

Theorem 2. *The scale invariance (2.17) of $C_{\rho W}$ implies that*

$$\forall j \le J \quad , \quad C_S(j,a,b) = C_S(0,a,b). \tag{2.21}$$

The theorem is proved in A.4. This condition on scattering cross-spectrum coefficients is necessary and almost sufficient to guarantee the scale invariance of the normalized wavelet modulus auto-correlation $\mathbb{E}\{|Wx|\,|Wx|^T\}$. To do so we would also need to impose an invariance condition on the off-diagonal coefficients of C_S but these coefficients have mostly a negligible amplitude. In the following, we shall thus systematically replace $\mathbb{E}\{|Wx|\,|Wx|^T\}$ by the scattering cross-spectrum correlation C_S and assess the scale invariance from (2.21). They are estimated by replacing expected values by a time averaging. For self-similar processes, C_s does not depend on j. These invariant coefficients are thus estimated by averaging them over scales. It defines a *scale invariant scattering cross-spectrum*

$$\overline{C}_S(a,b) = \langle C_S(j,a,b)\rangle_j\,. \tag{2.22}$$

For a signal of size T, there are at most $\log_2^2 T$ such coefficients.

2.5 Numerical dashboard for multi-scale processes

We show that the Scattering Spectra, defined as the multi-scale moments (2.6,2.18,2.19,2.22), specify intermittency, time-asymmetries and self-similar properties of multi-scale processes. Next

section studies standard mathematical models of multi-scale processes, and the following section considers numerical financial and turbulent time-series.

2.5.1 Models of self-similar processes

We consider Brownian motions, Poisson processes, multifractal random walks and Hawkes processes. To analyze their self-similarity properties we display and analyze their Scattering Spectra, composed of

— $\sigma_W^2(j)$: wavelet spectrum (2.6) shown in Figure 2.5a,

— $s_W^2(j)$: wavelet sparsity factor (2.18) in Figure 2.5b,

— $\overline{C}_{W|W|}(a)$: scale invariant phase-modulus cross-spectrum (2.19) in Figure 2.6.

— $\overline{C}_S(a,b)$: scale invariant scattering cross-spectrum (2.22) in Figure 2.7.

Table 2.1 gives the power-law decay parameters of σ_W^2 and s_W^2.

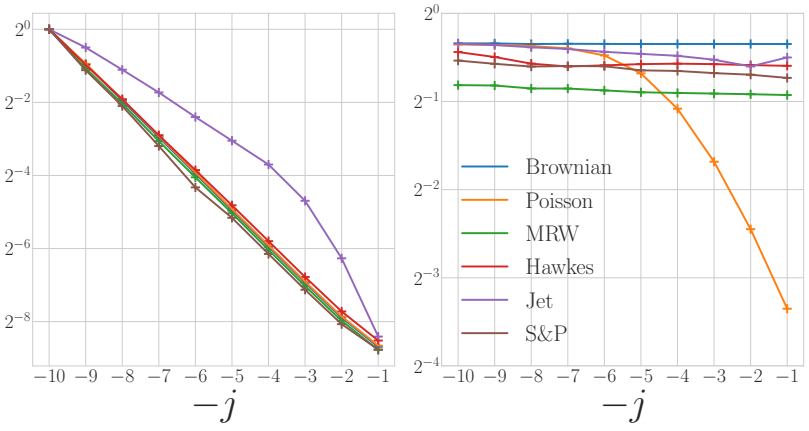
Expected values are estimated with empirical averages over T samples in time, as in (2.7). If x has a finite integrable scale s, which means that $x(t)$ and $x(t')$ are independent if $|t-t'| > s$ (see section 2.2.1), then we set the maximum wavelet scale 2^J to be equal to s. When the signal size T goes to ∞, time average estimators are consistent estimators of expected values. Indeed, if the wavelet has a support of size α then all Scattering Spectra coefficients are expected values of operators whose support sizes are at most $2\alpha 2^J$. One can thus verify that each empirical estimator averages at least $T(2\alpha s)^{-1}$ blocks of independent coefficients, which have the same mean because x is stationary. These empirical estimators thus converge to the expected value when T increases, with a variance which decays at least like $2\alpha s T^{-1}$.

In this section, we consider multi-scale processes whose integrable scale is not necessarily finite. The maximum wavelet scale 2^J is chosen to be smaller than the signal size T in order for large scale coefficients to be well estimated. Time average estimators of Scattering Spectra coefficients can then provide consistent estimators at all the scales smaller than 2^J. The variance decay of our estimators is not guaranteed mathematically, but it is verified numerically.

	Brown.	Poisson	MRW	Hawkes	Jet	S&P
ζ_2	1.0	1.0	1.0	1.0	0.66	1.0
c_s	0.79	×	0.66	0.65	0.67	0.61
ζ_s	0.00	×	0.016	0.013	0.025	0.016

TABLE 2.1 – Self-similarity parameters ζ_2, c_s, ζ_s, for different multi-scale processes. For the jet, ζ_2, c_s, ζ_s are given on the self-similar range.

Fractional Brownian motion. It is a Gaussian self-similar process, studied by Mandelbrot [Mandelbrot, 1968]. It has stationary increments and a generalized power spectrum $P_x(\omega) = c_2 |\omega|^{-\zeta_2 - 1}$. Computations are performed with $\zeta_2 = 1$, which corresponds to a standard Brownian motion. Wavelet sparsity coefficients have a power law decay with $c_s = \pi/4$ and $\zeta_s = 0$ in Table 2.1, because it is a Gaussian process. Propositions 1 and 3 prove that $\overline{C}_{W|W|}(a) \approx 0$ and $\overline{C}_S(a,b) \approx 0$ for $a > 0$, which is verified in Figure 2.6 and Figure 2.7. For $a = 0$, we also observe



(a) Wavelet power spectrum σ_W^2 (b) Sparsity factor s_W^2

FIGURE 2.5 – (a) Power spectrum $\sigma_W^2(j)$ in (2.6) as a function of $-j$, which is a log frequency index. (b) Sparsity factor $s_W^2(j)$ in (2.18) as a function of $-j$. Each process is shown with a different color. Non-linear curves reveal absence of self-similarity for Poisson process and the turbulent jet at fine scales.

in Figure 2.7 that $\overline{C}_S(0,b)$ is constant, which shows that $|x \star \psi_j(t)|$ are uncorrelated at time increments which are sufficiently large relatively to the scale 2^j. As proved by propositions 2 and 3, the phases of $C_{W|W|}$ and C_S are zero in Figures 2.6 and 2.7, because a Gaussian process is time-reversible. Since Brownian motions are self-similar, the phase-modulus cross-spectrum $C_{W|W|}$ and the scattering cross-spectrum C_S remain constant across scales 2^j.

Poisson process. It is a jump process having independent stationary increments. The number of jumps in an interval is proportional to the intensity λ. We further suppose that each jump is positive or negative with a probability $1/2$. Its power spectrum has a power law decay with $\zeta_2 = 1$, but it is non-Gaussian and not self similar, which clearly appears in its Scattering Spectra. Figure 2.5b shows that $\log s_W^2(j)$ in (2.23) has a slope which varies as a function of $-j$. Indeed, if $2^j \ll \lambda^{-1}$ then [Bruna, 2015] proves that $s_W^2(j) \sim 2^j$ because

$$\mathbb{E}\{|x \star \psi_j(t)|\}^2 \sim \lambda^2\, 2^{2j} \quad \text{and} \quad \mathbb{E}\{|x \star \psi_j(t)|^2\} \sim \lambda^2\, 2^j,$$

whereas if $2^j \gg \lambda^{-1}$ then $s_W^2(j) \sim 1$ because $x \star \psi_j(t)$ converges in distribution to a Gaussian white noise of variance $\lambda 2^j$ as 2^j goes to ∞ [Bruna, 2015], and hence

$$\mathbb{E}\{|x \star \psi_j(t)|\}^2 \sim \lambda\, 2^j \quad \text{and} \quad \mathbb{E}\{|x \star \psi_j(t)|^2\} \sim \lambda\, 2^j\ .$$

The non-self-similarity of Poisson process is also revealed by the large variations of $C_S(j,a,b)$ along j, which appears as large error bars in Figure 2.7.

Multifractal Random Walk (MRW). It is a non-Gaussian self-similar process, whose increments have scaling exponents ζ_q in (2.4) that are quadratic in q. Increments are computed as

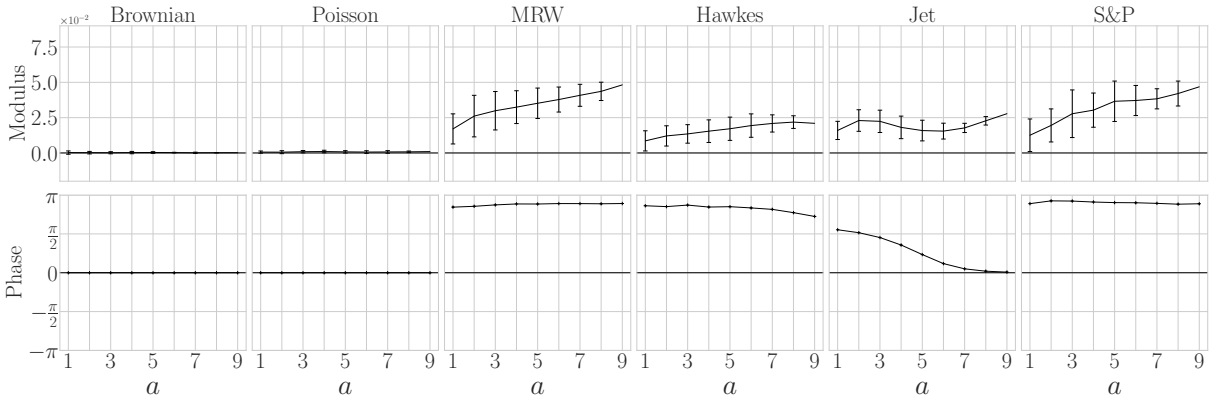


FIGURE 2.6 – Modulus and phase of the scale invariant phase-modulus cross-spectrum $\overline{C}_{W|W|}(a)$ (2.19). A skewed MRW and a quadratic Hawkes are shown. Error bars represent the mean-square variations of $C_{W|W|}(0; j, a)$ around $\overline{C}_{W|W|}(a)$. Non-zero phase coefficients reveal time-asymmetry.

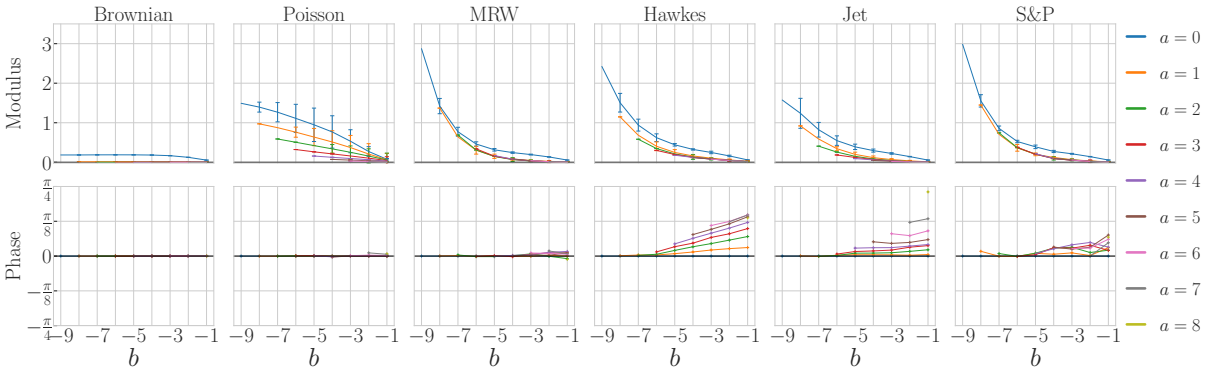


FIGURE 2.7 – Modulus and phase of the scale invariant scattering cross-spectrum $\overline{C}_S(a, b)$ (2.22) as a function of b. A skewed MRW and a quadratic Hawkes are shown. The parameter b is a log-frequency. Each color curve corresponds to a different scale shift a. Error bars represent the mean-square variations of $C_S(j, a, b)$ around $\overline{C}_S(a, b)$. Non-zero phases reveal time-asymmetry of wavelet modulus.

a product of the increment $\delta_j B$ of a Brownian motion with a log-normal process

$$\delta_j x(t) = \delta_j B(t)\, e^{\Omega_j(t)}.$$

The Gaussian process $\Omega_j(t)$ has an auto-correlation with a slow logarithmic decay :

$$\mathrm{Cov}\{\Omega_j(2^j t), \Omega_j(2^j t')\} = \lambda^2 \ln \theta(|t - t'|),$$

where θ decreases linearly and is specified in [Bacry, 2001a]. Since Ω_j is highly correlated in time, it creates wavelet modulus that are also highly correlated in time with bursts of activity. The parameter λ governs the intensity of this intermittency. If $\lambda = 0$ then the multifractal random walk is a Brownian motion. For MRW one can prove [Bacry, 2001a] that $\zeta_s = \lambda^2$, so Table 2.1 recovers the value $\lambda^2 = \zeta_s = 0.016$.

The scale invariant scattering cross-spectrum $\overline{C}_S(a,b)$ is the power spectrum of the modulus auto-correlation at two scales shifted by a, where b is a log-frequency index. Long range correlation of wavelet modulus appears in Figure 2.7, which shows that $\overline{C}_S(a,b)$ has a power-law decay for each a.

A skewed multifractal random walk incorporates a time-asymmetry by imposing that increments in the past are correlated with the future volatility [Pochart, 2002], in order to reflect the so-called leverage effect [Bekaert, 2000 ; Bouchaud, 2001]. This volatility is defined in Finance as the mean square average of increments over a fixed period of time. It amounts to replacing the Gaussian process Ω_j by $\Omega_j - k \star \delta_j B$ where $k(t)$ is a strictly causal power-law convolution kernel. The faster K decreases the shorter the scale of asymmetry. Figure 2.6 shows $\overline{C}_{W|W|}(a)$ for skewed MRW. As expected from Proposition 2, this time-asymmetry is revealed by the non-zero phase of $\overline{C}_{W|W|}$, which implies that its imaginary part is also non-zero.

Hawkes process. It is a non-homogeneous, causal self-excited point process [Bacry, 2014 ; Bacry, 2015], where each jump is positive or negative with probability 1/2. The jump intensity λ_t depends on the average of the past-jumps with power-law decaying kernel h

$$\lambda_t = \lambda_\infty + h \star |dN|_t,$$

where dN_s is the signed jump measure and $h \star |dN|_t := \int_{-\infty}^t h(t-s)|\mathrm{d}N_s|$. A linear feedback term $l \star dN_t$ can be added to account for correlation between past signed increments and future volatility.

A quadratic Hawkes model [Blanc, 2017] introduces time-asymmetric modulus dependencies with a causal quadratic feedback of kernel $k(t)$

$$\lambda_t = \lambda_\infty + h \star |dN|_t + l \star dN_t + |k \star dN_t|^2.$$

If $\int |h(t)|\,dt < 1$ then a quadratic Hawkes is stationary, and if $\int (|h(t)| + |k(t)|^2)\,dt < 1$ the mean intensity $\overline{\lambda}$ is finite [Aubrun, 2021]. In numerical simulation, as in [Blanc, 2017] we set $h(t) \sim |t|^{-1.2}$, $g(t) \sim e^{-0.01|t|}$ and $k(t) \sim e^{-0.03|t|}$ with $\int (|h(t)| + |k(t)|^2)\,dt = 0.9$. As expected from Proposition 3, Figure 2.7 shows that for this quadratic Hawkes, $\overline{C}_S(a,b)$ has a non-zero phase which reveals its modulus time-asymmetry [Zumbach, 2009].

2.5.2 Analysis of multi-scale time-series

Brownian motions, multifractal random walks and Hawkes processes are used as models of multi-scale time-series, particularly in Finance and Turbulence [Mandelbrot, 1968 ; Mandelbrot, 1997 ; Bacry, 2001b ; Mordant, 2002 ; Bacry, 2014]. The next two paragraphs analyze the Scattering Spectra of financial and turbulent time-series. Expected values are computed with time average estimators, which introduce an estimation error.

2.5.2.1 Finance

Finding stochastic models of financial time-series is important to compute the price of financial instruments which mitigate financial risks, such as options. A Brownian motion is a simple first order model, on which the Black and Scholes option pricing formula is based. However, many studies have shown strong deviations to Gaussianity, including the existence of bursts of activity and crises. Multifractal random walks [Bacry, 2001a], Hawkes processes [Blanc, 2017] and rough volatility models [Gatheral, 2018 ; El Euch, 2019b ; El Euch, 2019a] are among the most popular models used to capture non-Gaussian properties of financial markets [Mandelbrot, 1963 ; Bacry, 2001b ; Chicheportiche, 2014a ; Bacry, 2015 ; Blanc, 2017 ; Gatheral, 2018]. We consider the American stock index S&P log-prices sampled over 5 minutes from January 2000 to October 2018. A standard preprocessing is performed and is described in A.5. The resulting signal has $N = 7.5\,10^5$ samples.

Figure 2.5 and Table 2.1 show that S&P has a wavelet spectrum decay exponent $\zeta_2 = 1$, which is the same as a Brownian motion. However, its sparsity factor $c_s = 0.61 \neq \pi/4 \approx 0.79$ and exponent $\zeta_s = 0.016 \neq 0$, which shows that this time-series is clearly not Gaussian. These values are matched by the MRW model, and by a Hawkes process with calibrated parameters.

The S&P scale invariant phase-modulus cross-spectrum $\overline{C}_{W|W|}$ in Figure 2.6 is similar to a skewed multifractal random walk (SMRW), with a strong time-asymmetry shown by a non-zero phase, related to the leverage effect [Bekaert, 2000 ; Bouchaud, 2001]. The amplitude of $|\overline{C}_S(a,b)|$ in Figure 2.7 is similar for the S&P and the MRW. However, the phase of $\overline{C}_S(a,b)$ is non-zero for the S&P, which proves that wavelet modulus are also asymmetric in time. This is not well captured by the SMRW model. Such an effect is referred to as the Zumbach effect in Finance, which can be explained from causal agent reactions to past trends [Zumbach, 2009 ; Chicheportiche, 2014a ; Blanc, 2017]. Such a modulus time-asymmetry appears in the quadratic Hawkes model, but $|\overline{C}_S(a,b)|$ has different variations along the a direction for S&P and for the Hawkes model, presumably because of our choice of an exponentially decaying kernel $k(t)$ (which was calibrated on intraday data only). This analysis of the Scattering Spectra shows that none of these mathematical models fully capture all statistical properties of the S&P time-series.

Financial signals are often believed to be self-similar. We can estimate whether the S&P satisfies the wide-sense self-similarity properties of Theorem 1. The wavelet spectrum and sparsity factors in Figure 2.5 do indeed have a power law decay. Wide-sense self-similarity also imposes that $C_{W|W|}(0;j,a)$ and $C_S(j,a,b)$ do not depend upon the scale j. Figures 2.6 and 2.7 display their mean-square variations along j as error bars. They are of the same order as the estimation variance of each coefficients. We thus conclude that S&P time-series has no significant deviation from wide-sense self-similarity.

2.5.2.2 Turbulence

Kolmogorov introduced in 1941 a self-similar Gaussian model of Turbulence [Kolmogorov, 1941a ; Kolmogorov, 1 941 ; Kolmogorov, 1941b], with an asymptotic analysis of Navier-Stokes equations at high Reynolds numbers. This analysis predicts that the projection of the velocity

field on a line is a stationary process whose power spectrum has a power-law decay with $\zeta_2 = 2/3$. However, turbulent time-series are highly non-Gaussian, and Kolmogorov's initial theory was then refined to take into account intermittency and the presence of vortices [Kolmogorov, 1962 ; Frisch, 1991].

We study the Scattering Spectra of experimental velocity measurements along a single direction, measured over a turbulent gaseous helium jet at low temperature, with a high Reynolds number equal to 929 [Chanal, 2000]. This time-series has $N = 3.5\,10^7$ samples and is thus much larger than the S&P time-series, providing more accurate estimators of the Scattering Spectra. The non-zero phase of $\overline{C}_{W|W|}(a)$ and $\overline{C}_S(a,b)$ in Figures 2.6 and 2.7 shows a time-asymmetry, which results from the directionality of the jet propulsion. The quadratic Hawkes provides the best model of $\overline{C}_S(a,b)$ but it fails to accurately represent $\overline{C}_{W|W|}(a)$.

It thus appears that none of the existing mathematical model provides accurate models of this turbulent time-series.

Turbulence time-series may be self-similar on a frequency range limited by the Reynolds number at low frequencies and by the fluid viscosity at high frequencies. The wavelet power spectrum and sparsity factors in Figure 2.5 are indeed self-similar up to the finest scale $j = 2$, which is the diffusion scale created by the fluid viscosity. The self-similarity error bars in Figure 2.6 and Figure 2.7 are of the same order as for the S&P. However, their amplitude is statistically significant in this case because this time-series is 50 times larger than the S&P time-series and the estimation variance is thus much smaller. It shows that this turbulent time-series has significant deviations from wide-sense self-similarity.

2.6 Maximum entropy Scattering Spectra models

Brownian motions, multifractal random walks and Hawkes models are defined by a fixed number of parameters which does not depend upon their size T. They can be calibrated on data, but we saw they are typically too restrictive to capture all important properties of multi-scale time-series. On the other hand, neural network models [Oord, 2016 ; Eckerli, 2021] have a considerable flexibility but they can not be trained from a single time-series because the number of parameters is much larger than T.

This section introduces maximum entropy models computed from the scattering spectra energy vector $\Phi(x)$ of dimension smaller than $\log_2^3 T$ and hence much smaller than T. This energy vector is specified in the next section. It provides consistent estimators of the Scattering Spectra for random processes having a finite integral scale. We study the approximation of multi-scale time-series from such maximum entropy models. Section 2.6.3 shows that Scattering Spectra models are sufficiently flexible to approximate a wide range of mathematical processes, as well as real financial and turbulent data.

2.6.1 Scattering Spectra energy vector

We define maximum entropy models of the form $p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \theta, \Phi(x)\rangle}$ for a certain $\theta \in \mathbb{R}^M$, where the energy vector $\Phi(x)$ computes the Scattering Spectra estimation for any time-series x

of dimension T.

The Scattering Spectra vector $\Phi(x)$ is normalized by wavelet spectrum coefficients, which are constants estimated from a realization $\widetilde{x}$ of the random process x that we want to model

$$\widetilde{\sigma}_W^2(j) = \left\langle |\widetilde{x} \star \psi_j(t)|^2 \right\rangle_t .$$

If $x(t)$ has a finite integrable scale (see section 2.2.1), the estimator $\widetilde{\sigma}_W$ of σ_W is consistent as T goes to ∞.

Low-frequencies are captured by the low-pass filter $\varphi_J = \psi_{J+1}$ defined in (A.2). The Scattering Spectra energy is defined by four families of coefficients

$$\Phi(x) = \big(\Phi_1(x), \Phi_2(x), \Phi_3(x), \Phi_4(x)\big).$$

The first family provides J order 1 moment estimators squared, corresponding to wavelet sparsity coefficients (2.18)

$$\Phi_1(x)[j] = \frac{\langle |x \star \psi_j(t)| \rangle_t^2}{\widetilde{\sigma}_W^2(j)}. \tag{2.23}$$

The $J+1$ normalized second order wavelet spectrum associated to x are computed by

$$\Phi_2(x)[j] = \frac{\langle |x \star \psi_j(t)|^2 \rangle_t}{\widetilde{\sigma}_W^2(j)}. \tag{2.24}$$

There are $J(J+1)/2$ wavelet phase-modulus cross-spectrum coefficients

$$\Phi_3(x)[j,a] = \frac{\langle x \star \psi_j(t)\, |x \star \psi_{j-a}(t)| \rangle_t}{\widetilde{\sigma}_W(j)\, \widetilde{\sigma}_W(j-a)}. \tag{2.25}$$

Finally, it includes less than $J(J+1)(J+2)/6$ scattering cross-spectrum

$$\Phi_4(x)[j,a,b] = \frac{\left\langle |x \star \psi_j| \star \psi_{j-b}(t)\, |x \star \psi_{j-a}| \star \psi^*_{j-b}(t) \right\rangle_t}{\widetilde{\sigma}_W(j)\, \widetilde{\sigma}_W(j-a)}.$$

The dimension of $\Phi(x)$ is smaller than $J^3 \leq \log_2^3 T$ as soon as $T \geq 8$, this is much smaller than the dimension T of x.

2.6.2 Model validation with test moments

This section evaluates numerically the precision of maximum entropy models defined from the Scattering Spectra energy $\Phi(x)$ [1], to approximate the probability distributions of real mathematical processes and real data. The maximum entropy model $p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \theta, \Phi(x) \rangle}$ is sampled with a standard microcanonical algorithmic approach reviewed in A.6. It is computed with a gradient descent which avoids estimating the macrocanoncial parameters θ. The choice of the maximum wavelet scale 2^J amounts to defining an integrable scale equal to 2^J and hence nearly

1. the Scattering Spectra are a complex representation, for the sake of obtaining a real representation we concatenate its real and imaginary part

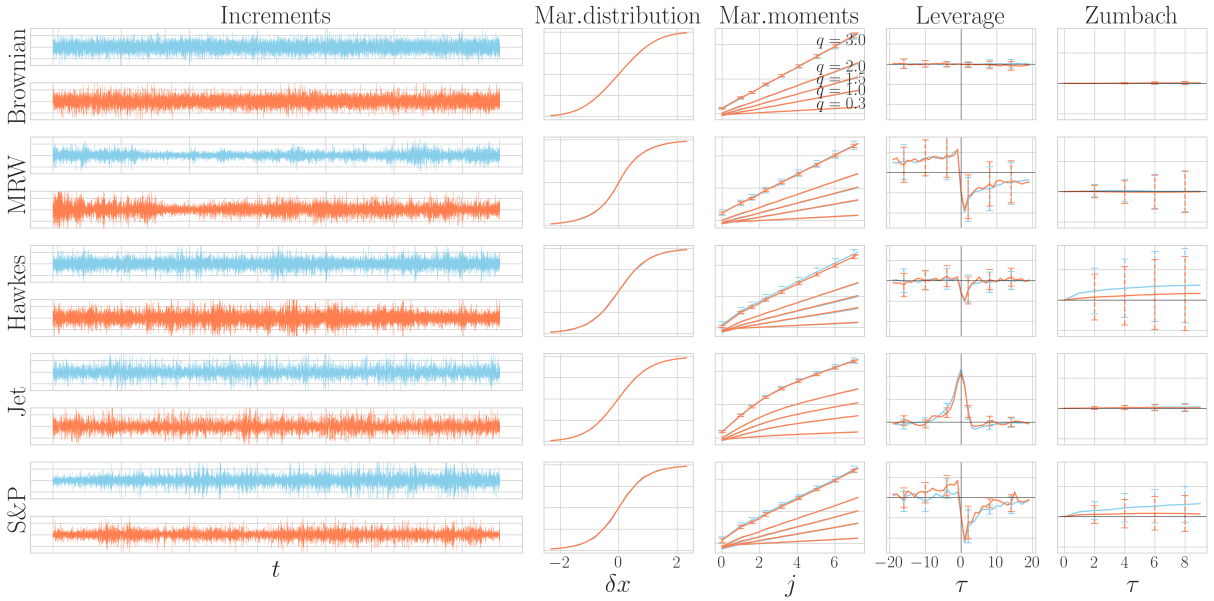


FIGURE 2.8 – Synthesis results. (Left) Increments $\delta_0 x(t)$ of original (blue) and generated (orange) time-series. (Right) Using the same color code : cumulative marginal distribution function $F(\delta x)$, log of marginal moments $\langle |\delta_j x(t)|^q \rangle_t$ as a function of j, leverage $\mathcal{L}_j(\tau)$ and Zumbach integral, for the Brownian, MRW (skewed), Hawkes (quadratic), turbulent jet and S&P. Leverage and Zumbach are shown for a specific j. Error bars show the standard deviation for test moments of order greater than 3. Though it is based on order 1 and order 2 moments, our model is able to capture higher order moments used to reveal scaling properties as well as time-asymmetries.

independent coefficients at distances larger than 2^J. Indeed, the Scattering Spectra model then does not impose any constraint on coefficients whose distance are much larger than 2^J so the entropy maximisation yields a random process whose coefficients are nearly independent at such distances.

Assessing the precision of a model from a dataset of samples drawn from an unknown distribution is an ill-defined problem. A maximum entropy model constrains the values of moments $\mathbb{E}\{\Phi(x)\}$. One may find errors by comparing moments which are not explicitly constrained. Such *test moments* are estimated on the original time-series and on time-series generated by the model itself. Following [Leonarduzzi, 2019], we describe test moments used in the Finance literature to identify important differences relatively to Brownian motions. Figure 2.8 compares moment values obtained from a Scattering Spectra model with the original time-series. The following test moments are computed on increments $\delta_j x(t) = x(t) - x(t - 2^j)$ which are stationary. While our model considers dyadic scales $j \in \mathbb{Z}$, we consider all scales $2^j \in \mathbb{R}^+$ for test moments.

Cumulative distribution $F(\delta x)$ of increments $\delta_0 x(t) = x(t) - x(t-1)$ at the finest scale.

Marginal high order moments described in (2.3) are estimated by time average $\langle |\delta_j x(t)|^q \rangle_t$ at all scales 2^j, for $0.3 \leq q \leq 3$. The multi-scale properties of these moments is reviewed in Section 2.2.2.

Leverage moments measure asymmetric dependencies between past and future increments in Finance [Bekaert, 2000 ; Bouchaud, 2001]. A leverage correlation for a time shift τ is an order 3

moment, at any scale 2^j :

$$\mathcal{L}_j(\tau) = \left\langle \delta_j x(t-\tau) |\, \delta_j x(t)|^2 \right\rangle_t .$$

If x has a time-reversible distribution then $\mathcal{L}_j(\tau) = -\mathcal{L}_j(-\tau)$. Results are shown in Figure 2.8 at an intermediate scale $2^j = 159$ that corresponds to the day in the case of S&P. The same scale is taken for all processes except for the Jet for which we take $2^j = 1$.

Zumbach moments evaluate the time-asymmetry of the volatility in Finance [Zumbach, 2009 ; Chicheportiche, 2014a]. The volatility is the energy of increments over a time period of size 2^j

$$\Sigma_j^2(t) = \left\langle |\delta_0 x(u)|^2 \right\rangle_{t-2^j \le u < t} .$$

A Zumbach moment for a time shift τ is an order 4 moment at a scale 2^j

$$\mathcal{Z}_j(\tau) = \left\langle |\delta_j x(t-\tau)|^2 \, \Sigma_j^2(t) \right\rangle_t .$$

If x is time-reversible then $\mathcal{Z}_j(-\tau) = \mathcal{Z}_j(\tau)$. To evaluate the time-asymmetry, Figure 2.8 shows $\int_0^t (\mathcal{Z}_j(s) - \mathcal{Z}_j(-s))\, \mathrm{d}s$ as a function of t for a scale $2^j = 159$, that corresponds to the day for S&P. This asymmetry coefficient on an order 4 moment is typically estimated with a large variance as we shall see.

2.6.3 Generation from Scattering Spectra models

Figure 2.8 gives results of Scattering Spectra models computed from a single realization of a Brownian motion, a skewed MRW, a quadratic Hawkes process, a turbulent jet and the S&P financial signal. It displays realizations generated by these models and compares test moments. For financial and turbulent data, the syntheses recover signals of size $N = 7.10^5$ with models computed over $J = 11$ scales. The resulting Scattering Spectra model has 375 parameters. Figure 2.8 shows that all test moments of order 2 or below are perfectly reproduced by Scattering Spectra models, for Brownian motion, MRW and Hawkes as well as for the turbulent jet and S&P financial data. Marginal moments of order $1/3 \le q \le 3$ are captured by our model on all processes, which is in accordance with the well reproduced cdf.

For test moments of order 3 or 4, including the leverage and Zumbach effects that capture time-asymmetries, we represent the variance of estimators with an error bar, which is quite large for the Zumbach effect. Leverage is well captured for MRW, Hawkes, Jet, and remains within the estimation error bar for S&P. Zumbach integral estimations have a much larger variance. The main information is in the sign of this integral, when significant. This test moment is again reproduced on both Hawkes and S&P within the estimation error. Its high variance clearly shows the importance of using low order moments, even for time-asymmetries. Scattering Spectra models reveal such non-Gaussian properties with a modulus and moments of order 1 and 2. In the case of the S&P, we believe that any remaining discrepancy for the Zumbach effect comes from our somewhat naive treatment of closing periods during the night.

2.7 Conclusion

We introduced the Scattering Spectra which give an interpretable low-dimensional representation of processes having stationary increments. It captures their power spectrum, multi-scale sparsity, and the dependencies of wavelet coefficients phase and modulus across scales. Wide-sense self-similar signals have Scattering Spectra which are invariant to scale shifts, and thus define a representation of even lower dimension.

For a time-series of size T, this Scattering Spectra is at most of dimension $\log_2^3 N$. We showed numerically that it reveals potential non-Gaussianity and self-similar properties. This was demonstrated on mathematical multi-scale models such as fractional Brownian motions, multifractal random walks and Hawkes processes, but also on real time-series in Finance and Turbulence. Maximum entropy Scattering Spectra models capture essential multi-scale dependency properties and can be efficiently sampled with a microcanonical approach.

Scattering Spectra models are related to generative convolutional neural networks based on covariance matrices [Gatys, 2015]. Similarly to a one-hidden layer convolutional neural network, it computes a cascade of two convolutions and a pointwise non-linearity. The network filters are wavelets which are not learned. It provides a much lower dimensional representation of random processes than usual deep convolutional neural networks, and it is furthermore interpretable. However, it only applies to signals which are stationary or have stationary increments.

Next chapter considers a first extension towards models of multivariate processes such as physical fields.

Chapitre 3

Scattering Spectra models of Physical fields

Foreword

Multivariate processes depending on multiple space variables are widely encountered in Physics. This chapter constitutes an extension of the previous chapter to this type of multivariate processes. Physicists need to characterize fields with a variety of structures, but building probabilistic models beyond the simple Gaussian model is often challenging, especially when the number of data samples is limited. We introduce Scattering Spectra models that characterize scale and angle dependencies on a field and make use of symmetry and regularity properties of physical fields and show that they can provide accurate and compact statistical descriptions for a wide range of fields. Providing both summary statistics and generative models, this representation can be used for data exploration, classification, parameter inference, and component separation in analyzing the ever-growing datasets in physics and beyond.

This chapter is adapted from the following submitted paper. Sihao Cheng, Rudy Morel, Erwan Allys, Brice Ménard, Stéphane Mallat. Scattering Spectra Models for Physics, 2023.

Contents

3.1 Introduction

An outstanding problem in statistics is to estimate the probability distribution $p(x)$ of high dimensional data x from few or even one observed sample. In physics, establishing probabilistic models of stochastic fields is also ubiquitous, from the study of condensed matter to the Universe itself. Indeed, even if physical systems can generally be described by a set of differential equations, it is usually not possible to fully characterize their solutions. Complex physical fields, described here as non-Gaussian random processes x, may indeed include intermittent phenomena as well as coherent geometric structures such as vortices or filaments. Having realistic probabilistic models of such fields however allows for considerable applications, for instance to accurately characterize and compare non-linear processes, or to separate different sources and solve inverse problems. Unfortunately, no generic probabilistic model is available to describe complex physical fields such as turbulence or cosmological observations. This chapter aims at providing such models for stationary fields, which can be estimated from one observed sample only.

At thermal equilibrium, physical systems are usually characterized by the Gibbs probability distribution, also called Boltzmann distribution, that depends on the energy of the systems [Landau, 2013]. For non-equilibrium systems, at a fixed time one may still specify the probability distribution of the field with a Gibbs energy, which is an effective Hamiltonian providing a compact representation of its statistics. Gibbs energy models can be defined as maximum entropy models conditioned by appropriate moments [Jaynes, 1957]. The main difficulty is to define and estimate the moments which specify these Gibbs energies.

For stationary fields, whose probability distributions are invariant to translation, moments are usually computed with a Fourier transform, which diagonalizes the covariance matrix of the field. The resulting covariance eigenvalues are the Fourier power spectrum. However, capturing non-Gaussian properties requires to go beyond second-order moments of the field. Third and fourth-order Fourier moments are called bispectrum and trispectrum. For a cubic d-dimensional stationary field of length L, the number of coefficients in the raw power spectrum, bispectrum and trispectrum are $O(L^d)$, $O(L^{2d})$ and $O(L^{3d})$ respectively. High-order moment estimators have high variance and are not robust, especially for non-Gaussian fields, because of potentially rare outliers which are amplified. It is thus very difficult to accurately estimate these high-order Fourier spectra from a few samples. Accurate estimations require to considerably reducing the number of moments and eliminating the amplification effect of high-order moments.

Local conservation laws for mass, energy, momentum, charge, etc. result in continuity equations or transport equations. The resulting probability distributions of the underlying processes thus are typically regular to deformations that approximate the local transport. These properties have motivated many researchers to use of a wavelet transform as opposed to a Fourier transform, which provides localized descriptors. Most statistical studies have concentrated on second-order and marginal wavelet moments [e.g., Bougeret, 1995 ; Vielva, 2004 ; Podesta, 2009] which fail to capture important non-Gaussian properties of a field. Other studies [Ha, 2021] use wavelet operator for interpretation with application to cosmological parameter inference, but rely on a trained neural network model.

In recent years, new representations have been constructed by applying point-wise non-linear operators on the wavelet transforms of non-Gaussian fields to recover their high-order statistics. The scattering transform, for instance, is a representation that is built by cascading wavelet transforms and non-linear modulus [Mallat, 2012 ; Bruna, 2013]. This representation has been used in astrophysics and cosmology [Cheng, 2021a], to study the interstellar medium [Allys, 2019 ; Saydjari, 2021], weak-lensing fields [Cheng, 2020 ; Cheng, 2021b], galaxy surveys [Valogiannis, 2022], or radio observations [Greig, 2022]. Other representations, which are built from covariances of phase harmonics of wavelet transforms [Mallat, 2020 ; Zhang, 2021], have also been used to model different astrophysical processes [Allys, 2020 ; Jeffrey, 2022 ; Régaldo-Saint Blancard, 2023]. Such models, which can be built from a single image, have in turn enabled the development of new component separation methods [Regaldo-Saint Blancard, 2021 ; Delouis, 2022], which can be directly applied to observational data without any particular prior model of the components of a mixture [Auclair, 2023].

These models however suffer from a number of limitations : they are not very good at reproducing vortices or long thin filaments, and they require an important number of coefficients to capture dependencies between distant scales, as well as angular dependencies. Building on those previous works, reduced scattering covariance representations have been introduced, but only for time-series, by leveraging scale invariance as we did in chapter 2. In this chapter, we present the Scattering Spectra, a low-dimensional representation that is able to efficiently describe a wide range of non-Gaussian processes encountered in physics. In particular, we show how it is possible to take into account the intrinsic regularity of physical fields to dramatically reduce the dimension of such representations. The first part of the chapter presents maximum entropy models and Scattering Spectra statistics, as well as their dimensional reduction. The second part of the chapter presents a quantitative validation of these models on various two-dimensional multiscale physical fields and discuss their limitations.

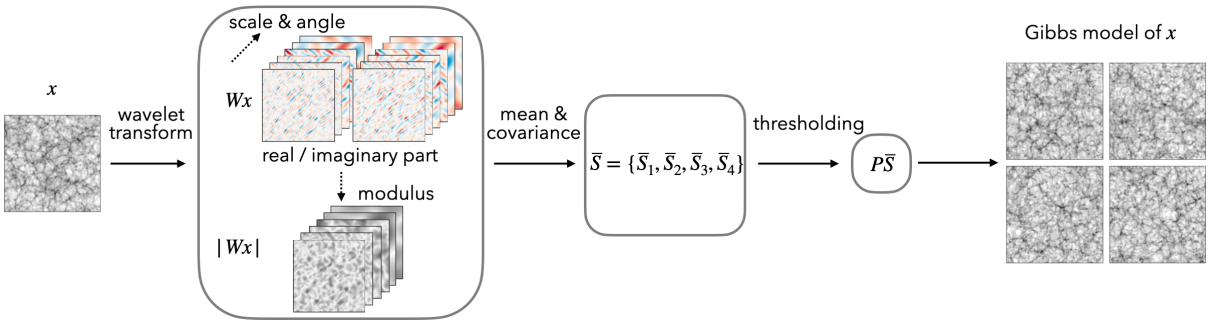


FIGURE 3.1 – Steps to build a feasible model for a random field x from only one or a few realizations. We first build a low-dimension representation $\Phi(x)$ of the random field, which specifies a maximum entropy model. The representation $\Phi(x)$ is obtained by conducting the wavelet transform Wx and its modulus $|Wx|$, and then computing the means and covariance of all wavelet channels $(Wx\,, |Wx|)$. Such a covariance matrix is further binned and sampled using wavelets to reduce its dimensionality, which is called the Scattering Spectra $\bar{S}(x)$. Finally, These Scattering Spectra are renormalized and reduced in dimension by thresholding its Fourier coefficients along rotation and scale parameters $\Phi = P\bar{S}$, making use of the regularity properties of the field. For many physical fields, this representation can be as small as only around $\sim 10^2$ coefficients.

3.2 Methods

3.2.1 Gibbs energy of stationary fields

We review the properties of Gibbs energies resulting from maximum entropy models conditioned by moment values [Geman, 1984 ; Zhu, 1997 ; Zhu, 1998]. We write $x(u)$ a field where the site index u belongs to a cubic d-dimensional lattice of size L. It results that $x \in \mathbb{R}^{L^d}$.

Assume that $x \in \mathbb{R}^{L^d}$ has a probability density $p(x)$ and consider Gibbs energy models linearly parameterized by a vector $\theta = \{\theta_m\}_{m \leq M}$ over a potential vector $\Phi(x) = \{\Phi_m(x)\}_{m \leq M}$ of dimension M

$$U_\theta(x) = \langle \theta, \Phi(x) \rangle = \sum_{m=1}^{M} \theta_m \, \Phi_m(x).$$

They define exponential probability models

$$p_\theta(x) = Z_\theta^{-1} \, e^{-\langle \theta, \Phi(x) \rangle}.$$

The model class is thus defined by the potential vector $\Phi(x)$, which needs to be chosen appropriately.

If it exists, the maximum entropy distribution conditioned by $\mathbb{E}\{\Phi(x)\}$ is a p_{θ_0} which belongs to this model class. It has a maximum entropy $H(p_{\theta_0}) = -\int p_{\theta_0}(x) \log p_{\theta_0}(x)\, dx$ under the expected value condition

$$\int \Phi(x)\, p_{\theta_0}(x)\, dx = \mathbb{E}\{\Phi(x)\}. \tag{3.1}$$

In statistical physics, p_{θ_0} is a macrocanonical model defined by a vector $\mathbb{E}\{\Phi(x)\}$ of observables.

One can verify that θ_0 also minimizes the Kullback-Liebler divergence within the class

$$D(p\|p_{\theta_0}) = \int p(x) \log \frac{p(x)}{p_{\theta_0}(x)}\, dx = H(p_{\theta_0}) - H(p). \tag{3.2}$$

The main topic of the chapter is to specify $\Phi(x)$ in order to define accurate maximum entropy models for large classes of physical fields, which can be estimated from a small number n of samples $\widetilde{x}_i$. In this section, we suppose that $n = 1$. Reducing the model error given by (3.2) amounts to defining Φ which reduces the excess entropy of the model. This can be done by enriching $\Phi(x)$ and building very high-dimensional models. However, we must also take into account the empirical estimation error of $\mathbb{E}\{\Phi(x)\}$ by $\Phi(\widetilde{x}_1)$, measured by $\mathbb{E}\{\|\Phi(x) - \mathbb{E}\{\Phi(x)\}\|^2\}$.

In this chapter, as in the rest of the thesis, macrocanonical models are approximated by microcanonical models, which have a maximum entropy over a microcanonical set of width $\epsilon > 0$

$$\Omega_\epsilon = \{x \in \mathbb{R}^{L^d} \mid \|\Phi(x) - \Phi(\widetilde{x}_1)\|^2 \leq \epsilon\}. \tag{3.3}$$

Appendix A.6 reviews a sampling algorithm for such model. It also explains how to extend the definition of Ω_ϵ for $n > 1$ samples $\widetilde{x}_i$ by replacing $\Phi(\widetilde{x}_1)$ by $\langle \Phi(\widetilde{x}_i) \rangle_i$. If $\Phi(x)$ concentrates around $\mathbb{E}\{\Phi(x)\}$ then the microcanonical model converges to the macrocanonical model when the system length L goes to ∞ and ϵ goes to 0. The concentration of $\Phi(x)$ generally imposes that its dimension M is small relatively to the dimension L^d of x. The choice of $\Phi(x)$ must thus incorporate a trade-off between the model error (3.2) and the distance between micro and macrocanonical distributions.

3.2.2 Fourier polyspectra potentials

Gaussian random fields are maximum entropy models conditioned on first and second-order moments. The potential vector $\Phi(x)$ is then an empirical estimator of first and second-order moments of x. For stationary fields, there is only one first-order moment $\mathbb{E}\{x(u)\}$ which can be estimated with an empirical average [1] over u : $\langle x(u) \rangle_u$. Similarly, the covariance matrix $\mathbb{E}\{x(u)x(u')\}$ only depends on $u - u'$, so only the diagonal coefficients in Fourier space are informative, which are called the power spectrum,

$$\mathbb{E}\{\hat{x}(\omega)\, \hat{x}(\omega')^*\} \quad \text{with} \quad \omega = \omega'. \tag{3.4}$$

The off-diagonal elements vanish because of phase cancellation under all possible translations, which means the second-order moments treat Fourier coefficients independently, and cannot describe relations or dependence between them. The diagonal elements, which can also be written as $|\hat{x}(\omega)|^2$, can be estimated from a single sample x by averaging $|\hat{x}(\omega)|^2$ over frequency bins that are large enough to reduce the estimator variance. A uniform binning and sampling along frequencies results in power spectrum estimators with $O(L^d)$ elements, so the Gaussian model is compact and feasible.

1. This single moment can be directly constrained, and we do not discuss it in the following.

However, the Gaussian random field model has limited power to describe complex structures. The majority of fields encountered in scientific research are not Gaussian. Non-Gaussianity usually means dependence between Fourier coefficients at different frequencies. The traditional way goes to higher orders moments of $\hat{x}$, the polyspectra [Brillinger, 1965], where phase cancellation implies that for stationary fields, only the following moments are informative,

$$\mathbb{E}\{\hat{x}(\omega_1)\,...\hat{x}(\omega_n)\} \quad \text{with} \quad \omega_1 + ... + \omega_n = 0, \tag{3.5}$$

while other moments are zero. These polyspectra at order $n > 2$ capture dependence between $n-1$ independent frequencies. As the leading term, the Fourier bispectrum specifies the non-zero third-order moments and has $O(L^{2d})$ coefficients. However, bispectrum is usually not sufficient to characterize non-Gaussian fields. For example, it vanishes if the field distribution is symmetric $p(x) = p(-x)$. One must then estimate fourth-order Fourier moments, the trispectrum, which has $O(L^{3d})$ coefficients.

There are two main problems for the polyspectra coefficients to become proper potential functions $\Phi(x)$ in the maximum entropy models. First, the number of coefficients increases sharply with the order. Second, high-order moments are not robust and difficult to estimate from a few realizations [Huber, 1981]. For random fields with a heavy tail distribution, which is ubiquitous in complex systems [Bak, 1987; Bouchaud, 1990; Coles, 1991; Kello, 2010; , 2017], higher order moments may not even exist. Those two problems are common for high-order moments and have been demonstrated in real-world applications [Dudok de Wit, 2004; Lombardo, 2014]. In the following two sections, we introduce modifications to this approach to solve those problems.

3.2.3 Wavelet polyspectra

Many physical fields exhibit multiscale structures induced by non-linear dynamics, which implies regularity of $p(x)$ in frequency. The wavelet transform groups Fourier frequencies by wide logarithmic bands, providing a natural way to compress the Fourier polyspectra. The compression not only reduces the model size but also improves estimator convergence. We use the wavelet transform to compute a compressed power spectrum estimate, as well as a reduced set of $O(\log^2 L)$ third and $O(\log^3 L)$ fourth order wavelet moments, allowing for efficient estimation of the polyspectra.

3.2.3.1 Wavelet transform

A wavelet is a localized wave-form $\psi(u)$ for $u \in \mathbb{R}^d$ which has a zero average $\int_{\mathbb{R}^d} \psi(u)\,du = 0$. We shall define complex-valued wavelets $\psi(u) = g(u)\,e^{i\langle \xi, u\rangle}$ where $g(u)$ is a real window whose Fourier transform $\hat{g}(\omega)$ is centered at $\omega = 0$ so that $\hat{\psi}(k) = \hat{g}(\omega - \xi)$ is localized in the neighborhood of the frequency ξ. Fig. B.1 shows ψ and $\hat{\psi}$ for a $d = 2$ dimensional Morlet wavelet described in appendix B.1. The wavelet transform is defined by rotating $\psi(u)$ with a rotation r

in $\mathbb{R}^d$ and by dilating it with dyadic scales $2^j > 1$. It defines

$$\psi_\lambda(u) = 2^{-jd}\,\psi(2^{-j}r^{-1}u) \;\text{ with }\; \lambda = 2^{-j}\,r\xi\ . \tag{3.6}$$

Its Fourier transform is $\hat{\psi}_\lambda(\omega) = \hat{g}(2^j r^{-1}(\omega - \xi))$, which is centered at the frequency λ and concentrated in a ball whose radius is proportional to 2^{-j}.

To decompose a field $x(u)$ defined over a grid of width L, the wavelet is sampled on this grid. Wavelet coefficients are calculated as convolutions with periodic boundary conditions

$$Wx(u,\lambda) = x \star \psi_\lambda(u) = \sum_{u'} x(u')\,\psi_\lambda(u-u'). \tag{3.7}$$

It measures the variations of x in a spatial neighborhood of u of length proportional to 2^j, and it depends upon the values of $\hat{x}$ in a frequency neighborhood of $\omega = \lambda$ of length proportional to 2^{-j}. The scale 2^j is limited to $1 \leq j \leq J$, and for practical application to fields with a finite size L, the choice of J is limited by $J < \log L$. Left part of Fig. 3.1 illustrates the wavelet transform of an image.

The rotation r is chosen within a rotation group of cardinal R, where R does not depend on L. Wavelet coefficients need to be calculated for $R/2$ rotations because $Wx(u,-\lambda) = Wx(u,\lambda)^*$ for real fields. In $d = 2$ dimensions, the R rotations have an angle $2\pi\ell/R$, and we set $R = 8$ in all our numerical applications, which boils down to 4 different wavelet orientations. The total number of wavelet frequencies λ is $RJ = O(\log L)^{\,2}$ as opposed to L^d Fourier frequencies.

A wavelet transform is also stable and invertible if ψ satisfies a Littlewood-Paley condition, which requires an additional convolution with a low-pass *scaling* function ψ_0 centered at the frequency $\lambda = 0$. The specifications are detailed in appendix B.1.

3.2.3.2 Wavelet power spectrum

Given scaling regularity, one can compress the $O(L^d)$ power spectrum coefficients into $RJ = O(\log L)$ coefficients using a logarithmic binning defined by wavelets. This is obtained by averaging the power spectrum with weight functions as the Fourier transform of wavelets, which are band-pass windows, $\left\langle \mathbb{E}\{|\hat{x}(\omega)|^2\}\,|\hat{\psi}_\lambda(\omega)|^2 \right\rangle_\omega$. The limited number of wavelet power spectrum coefficients has reduced estimation variance. In fact, they are also the diagonal elements of the wavelet covariance matrix, $Wx(u,\lambda)Wx(u,\lambda)^* = |Wx(u,\lambda)|^2$, therefore an empirical estimation can also be written as an average over u :

$$M_2(x)[\lambda] = \left\langle |Wx(u,\lambda)|^2 \right\rangle_u . \tag{3.8}$$

Similar to the power spectrum, phase cancellation due to translation invariance means that the off-diagonal blocks i.e. the cross-correlations between different wavelet frequency bands are

2. Here we assume the choice of R is independent of field dimension d. Another possible choice is to require a constant ratio between the radial and tangential sizes of the d-dimension oriented wavelets. Then, R is proportional to the ratio between the surface area of a d–1-sphere and the volume of a d–1-ball, proportionally to $\Gamma(n/2 + 1/2)/\Gamma(n/2)$. It results in an approximate scaling of $RJ = O(d \log L)$ when d is small and $O(\sqrt{d}\log L)$ when d is large.

nearly zero because the support of two wavelets $\hat{\psi}_\lambda$ and $\hat{\psi}_{\lambda'}$ are almost disjoint, as illustrated in Fig. 3.2(a).

3.2.3.3 Selected 3rd and 4th order wavelet moments

One may expect to compress the polyspectra in a similar manner with a wavelet transform, taking advantage of the regularities of the field probability distribution. However, it is non-trivial to logarithmically bin the polyspectra because more than one independent frequency is involved and the phase cancellation condition needs to be considered.

To solve this problem, let us revisit the phase cancellation of two frequency bands, which causes their correlation to be zero,

$$\mathbb{E}\{Wx(u,\lambda)\, Wx(u',\lambda')^*\} \sim 0\,,$$

for $\lambda \neq \lambda'$. To create a non-zero correlation, we must realign the support of $Wx(u,\lambda)$ and $Wx(u',\lambda')$ in Fourier space through non-linear transforms. As shown in Fig. 3.2(b), we may apply a square modulus to one band (shown in blue) in the spatial domain, which recenters its frequency support at origin. Indeed, $|x \star \psi_\lambda|^2 = (x \star \psi_\lambda)(x \star \psi_\lambda)^*$ has a Fourier support twice as wide as that of $x \star \psi_\lambda$, and will overlap with another wavelet band with lower frequency than λ. The transformed fields $|x \star \psi_\lambda|^2$ can be interpreted as maps of locally measured power spectra. Correlating this map with another wavelet band $x \star \psi'_\lambda$ gives some third-order moments

$$\mathbb{E}\{|Wx|^2(u,\lambda)\, Wx(u',\lambda')^*\}$$

that are a priori non-zero. Furthermore, for wide classes of multiscale processes having regular power spectrum, it suffices to only keep the coefficients at $u = u'$ because of random phase fluctuation (see appendix B.1). For stationary random fields, they can be estimated with an empirical average over u,

$$M_3(x)[\lambda,\lambda'] = \left\langle |Wx|^2(u,\lambda)\, Wx(u,\lambda')^* \right\rangle_u . \tag{3.9}$$

Now we obtain a set of statistics characterizing the dependence of Fourier coefficients in two wavelet bands in a collective way, which are selected third-order moments. They can be interpreted as a logarithmic frequency binning of certain bispectrum coefficients. There are about $R^2 J^2 = O(\log^2 L)$ such coefficients, which is a substantial compression compared to the $O(L^{2d})$ full bispectrum coefficients.

Similarly, we consider the cross correlation between two wavelet bands both transformed by the square modulus operation and obtain a wavelet binning of fourth-order moments,

$$\mathbb{E}\{|Wx(u,\lambda)|^2\, |Wx(u',\lambda')|^2\} - \mathbb{E}\{|Wx(u,\lambda)|^2\}\mathbb{E}\{|Wx(u',\lambda')|^2\}.$$

For stationary fields, this covariance only depends on $u - u'$. A further reduction of such a large covariance function is possible because its Fourier transform over $u - u'$ has two properties.

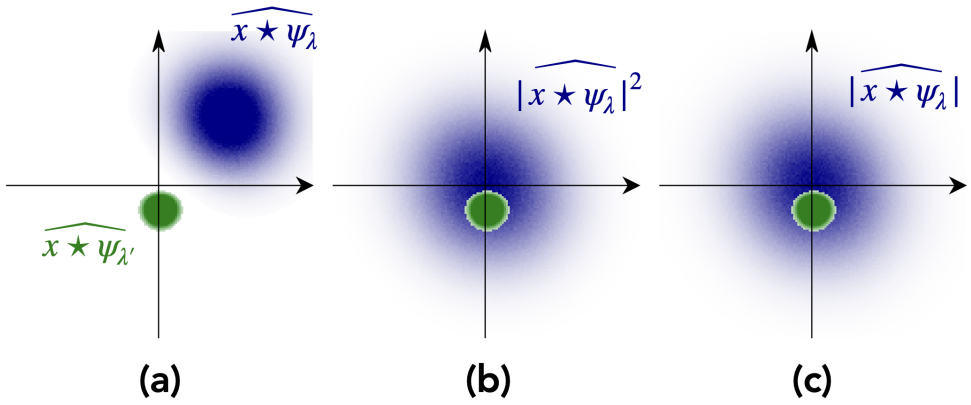


FIGURE 3.2 – (a) : For $\lambda \neq \lambda'$ the Fourier supports of $x \star \psi_\lambda$ (blue) and $x \star \psi_{\lambda'}$ (green) typically do not overlap. (b) : The Fourier support of $|x \star \psi_\lambda|^2$ is twice larger and centered at 0 and hence overlaps with $x \star \psi_{\lambda'}$ if $|\lambda'| \leq |\lambda|$. (c) : The Fourier support of $|x \star \psi_\lambda|$ is also centered at 0 and hence overlaps with $x \star \psi_{\lambda'}$ if $|\lambda'| < |\lambda|$.

First, it typically does not have higher frequency components than the initial wavelet transforms involved (see Fig. 3.2) as the phase fluctuations have been eliminated by the square modulus, and second, for fields with multiscale structures, it is regular and can be approximated with another logarithmic frequency binning. Thus, we can compress the large covariance function with a second wavelet transform, and estimate it by an empirical average over u :

$$M_4(x)[\lambda, \lambda', \gamma] = \left\langle W|Wx|^2[u, \lambda, \gamma] \;\; W|Wx|^2[u, \lambda', \gamma]^* \right\rangle_u , \tag{3.10}$$

where $(W|Wx|^2)[u, \lambda, \gamma] = |x \star \psi_\lambda|^2 \star \psi_\gamma(u)$, and the central frequencies of the second wavelets verifies $|\lambda| \geq |\lambda'| > |\gamma|$. There are about $R^3 J^3 = O(\log^3 L)$ such coefficients, which is also a substantial compression compared to the $O(L^{3d})$ full trispectrum coefficients.

3.2.4 Scattering Spectra

In general, the estimation of high-order moments has a high variance because high-order polynomials amplify the effect of outliers. A scattering approach [Mallat, 2012 ; Bruna, 2013 ; Cheng, 2021a] reduces the variance of these estimators by replacing $|Wx|^2$ by $|Wx|$. The resulting spectra only depend on the mean and covariance matrix of $(Wx, |Wx|)$, which are low-order transforms of the original field x.

Local statistics of wavelet modulus have been studied to analyze properties of image textures [Portilla, 2000]. Their mathematical properties have been analyzed to capture non-Gaussian characteristics of random fields [Mallat, 2020 ; Zhang, 2021] in relation to scattering moments [Mallat, 2012 ; Bruna, 2013]. Scattering Spectra have been defined on a univariate time-process in chapter 2, from the joint covariance of a wavelet transform and its modulus : $(Wx\,,\,|Wx|)$. We extend it to fields of arbitrary dimension d and length L, in relation to Fourier high-order moments, and define models of dimension $O(\log^3 L)$.

3.2.4.1 First and second wavelet moments, sparsity

For non-Gaussian fields x, wavelet coefficients $Wx(u,\lambda)$ define fields which are often sparse [Olshausen, 1996 ; Mallat, 1999]. This is a non-Gaussian property that can be captured by first-order wavelet moments $\mathbb{E}\{|Wx[u,\lambda]|\}$. If x is a Gaussian random field then $Wx(u,\lambda)$ remains Gaussian but complex-valued so, and we have $\frac{\mathbb{E}\{|Wx|\}^2}{\mathbb{E}\{|Wx|^2\}} = \frac{\pi}{4}$. This ratio decreases when the sparsity of $Wx[u,\lambda]$ increases. The expected value of $|Wx|$ is estimated by

$$S_1(x)[\lambda] = \langle |Wx[u,\lambda]| \rangle_u \tag{3.11}$$

and the ratio is calculated with the second-order wavelet spectrum estimator

$$S_2(x)[\lambda] = M_2(x)[\lambda] = \Big\langle |Wx|^2(u,\lambda) \Big\rangle_u . \tag{3.12}$$

3.2.4.2 Cross-spectra between scattering channels

A scattering transform is computed by cascading modulus of wavelet coefficients and wavelet transforms [Mallat, 2012 ; Bruna, 2013]. Let us replace $|Wx|^2$ by $|Wx|$ in the selected third and fourth-order wavelet moments described in the previous section. The third order moments (3.9) become $\mathbb{E}\{|Wx(u,\lambda)|\, Wx(u,\lambda')^*\}$. Such moments are a priori non-zero if the Fourier transforms of $|Wx(u,\lambda)| = |x \star \psi_\lambda(u)|$ and $Wx(u,\lambda') = x \star \psi_{\lambda'}(u)$ overlap. This is the case if $|\lambda'| < |\lambda|$ as illustrated in Fig. 3.2. Eliminating the square thus preserves non-zero moments which can capture dependencies between different frequencies λ and λ'. The third order moment estimators given by (3.9) can thus be replaced by lower cross-correlations between $|Wx|$ and Wx at $|\lambda| \geq |\lambda'|$

$$S_3(x)[\lambda,\lambda'] = \big\langle |Wx|(u,\lambda)\ Wx(u,\lambda')^* \big\rangle_u . \tag{3.13}$$

Replacing $|Wx|^2$ by $|Wx|$ in the fourth order wavelet moments (3.10) amounts to estimating the covariance matrix of wavelet modulus fields $|Wx|$. As the $u-u'$ dependency of this covariance can also be characterized by a second wavelet transform, this amounts in turn to estimate the covariance of scattering transforms $W|Wx|[u,\lambda,\gamma] = |x \star \psi_\lambda| \star \psi_\gamma(u)$

$$S_4(x)[\lambda,\lambda',\gamma] = \big\langle W|Wx|[u,\lambda,\gamma]\ \ W|Wx|[u,\lambda',\gamma]^* \big\rangle_u , \tag{3.14}$$

for $|\lambda| \geq |\lambda'| \geq |\gamma|$. It provides a wavelet spectral estimation of the covariance of $|Wx|$.

Combining the moment estimators of Eqs. (3.11,3.12,3.13,3.14) defines a vector of Scattering Spectra

$$S(x) = \Big(S_1(x)\,,\ S_2(x)\,,\ S_3(x)\,,\ S_4(x) \Big). \tag{3.15}$$

It provides a mean and covariance estimation of the joint wavelet and wavelet modulus vectors $(Wx, |Wx|)$. It resembles the second, third, and fourth-order Fourier spectra but has much fewer coefficients and better information concentration. Considering the conditions satisfied by λ, λ', and γ, the exact dimension of $S(x)$ is $RJ + R^2J(J-1)/8 + R^3J(J^2-1)/48$, of order $O(\log^3 L)$.

3.2.4.3 Renormalization

Scattering Spectra coefficients must often be renormalized to improve the sampling of maximum entropy models. Indeed, multiscale random processes often have a power spectrum that has a power law decay $\mathbb{E}\{|\hat{x}(\omega)|^2\} \sim |\omega|^{-\eta}$ over a wide range of frequencies, long-range correlations corresponding to a strong decay from large to small scales. The wavelet spectrum also has a power-law decay $\mathbb{E}\{|Wx(u,\lambda)|^2\} \sim |\lambda|^{-\eta}$. This means that if we build a maximum entropy model with $\Phi(x) = S(x)$ then the coordinate of $\Phi(x)$ of low-frequencies λ have a much larger amplitude and variance than at high frequencies. The microcanonical model is then dominated by low frequencies and is unable to constrain high-frequency moments. The same issue appears when computing the θ_0 parameters of a macrocanonical model defined in (3.1), for which it has been shown that renormalizing to 1 the variance of wavelet coefficients at all scales avoid numerical instabilities [Marchand, 2022] [3].

We renormalize the Scattering Spectra by the variance of wavelet coefficients, $\sigma^2[\lambda] = \langle S_2(\widetilde{x}_i)[\lambda]\rangle_i$, which can be estimated from a few samples. The renormalized Scattering Spectra are

$$\bar{S}(x) = \Big(\bar{S}_1(x)\,,\,\bar{S}_2(x)\,,\,\bar{S}_3(x)\,,\,\bar{S}_4(x)\Big)$$

defined by

$$\bar{S}_1(x)[\lambda] = \frac{S_1(x)[\lambda]}{\sigma[\lambda]} \quad , \quad \bar{S}_2(x)[\lambda] = \frac{S_2(x)[\lambda]}{\sigma^2[\lambda]} \tag{3.16}$$

$$\bar{S}_3(x)[\lambda,\lambda'] = \frac{S_3(x)[\lambda,\lambda']}{\sigma[\lambda]\,\sigma[\lambda']}\ ,\ \bar{S}_4(x)[\lambda,\lambda',\gamma] = \frac{S_4(x)[\lambda,\lambda',\gamma]}{\sigma[\lambda]\,\sigma[\lambda']}.$$

The microcanonical models proposed in this chapter are built from these renormalized statistics and/or their reduced version described below.

3.2.5 Dimensionality reduction for physical fields

Though much smaller than the polyspectra representation, the Scattering Spectra $\bar{S}$ representation still has a large size. Assuming isotropy and scale invariance of the field x, a first-dimensional reduction can be performed that relies on the equivariance properties of Scattering Spectra with respect to rotation and scaling (see appendix B.2). However, such invariances cannot be assumed in general. In this section, we propose to construct a low-dimensional representation by only assuming regularity under rotation or scaling of the scales involved in the Scattering Spectra representation. A simplified version of such a dimensional reduction has been introduced in [Allys, 2019]. We refer the reader to appendix B.3 for technical details.

The goal of the reduction is to approximate the covariance coefficients $\bar{S}_3$ and $\bar{S}_4$, the most numerous, using only a few coefficients. This can be seen as a covariance matrix estimation problem. To do so, we first use a linear transform to sparsify the covariance matrix and then perform a threshold clipping on the coefficients to reduce the representation. We consider a linear

3. Without such a normalization, the calculation of θ_0 parameters at different frequencies is ill-conditioned, which turns into a "critical slowing down" of iterative optimization algorithms. The proposed normalization is closely related to Wilson renormalization.

transform $F\bar{S} = (\bar{S}_1, \bar{S}_2, F\bar{S}_3, F\bar{S}_4)$ with a pre-determined linear transform F which stands for a 2D or 3D Fourier transform along all orientations, as well as a 1D cosine transform along scales, for $\bar{S}_3$ and $\bar{S}_4$. For fields with statistical isotropy or self-similarity, all harmonics related to the action of global rotation and scaling on the field x should be consistent with zero, except for the zeroth harmonic. For general physical fields, we expect the statistics $\bar{S}(x)$ to have regular variations to the action of rotation or scaling of the different scales involved in its computation, which implies that its Fourier harmonics $F\bar{S}(x)$ have a fast decay away from the 0-th harmonic and $F\bar{S}(x)$ is a sparse representation.

Thresholding on a sparse representation is widely used in image processing for compression [Chang, 2000]. We use threshold clipping on the sparse representation $F\bar{S}$ to significantly reduce the size of the Scattering Spectra. Furthermore, when empirically estimating large but sparse covariance matrices such as $F\overline{S}$, thresholding provides Stein estimators [Stein, 1956] which have lower variance and are consistent[e.g., Donoho, 1994 ; Bickel, 2008 ; Cai, 2011 ; Fan, 2013]. As $\bar{S}_1$ or $\bar{S}_2$ are already small, we keep all of their coefficients.

There are different strategies available to set the threshold for clipping. We adopt a simple strategy which keeps those coefficients with $\mu(F\bar{S}) > 2\sigma(F\bar{S})$, where $\mu(F\bar{S})$ and $\sigma(F\bar{S})$ are the means and standard deviations of individual coefficients of $F\bar{S}$. These adaptive thresholding estimators achieve a higher rate of convergence and are easy to implement [Cai, 2011]. With multiple realizations from simulations, $\mu(F\bar{S})$ and $\sigma(F\bar{S})$ can be estimated directly. In the case where only a single sample field is available, $\sigma(F\bar{S})$ can be estimated from different patches of that sample field [e.g., Sherman, 2018]. We call $P\bar{S}$ the coefficients after thresholding projection :

$$P\bar{S} = (\bar{S}_1, \bar{S}_2, P\bar{S}_3, P\bar{S}_4) = \text{thresholding } F\bar{S}. \tag{3.17}$$

The compact yet informative set of Scattering Spectra $P\bar{S}$ is the representation $\Phi(x) = P\bar{S}(x)$ proposed in this chapter to construct maximum entropy models.

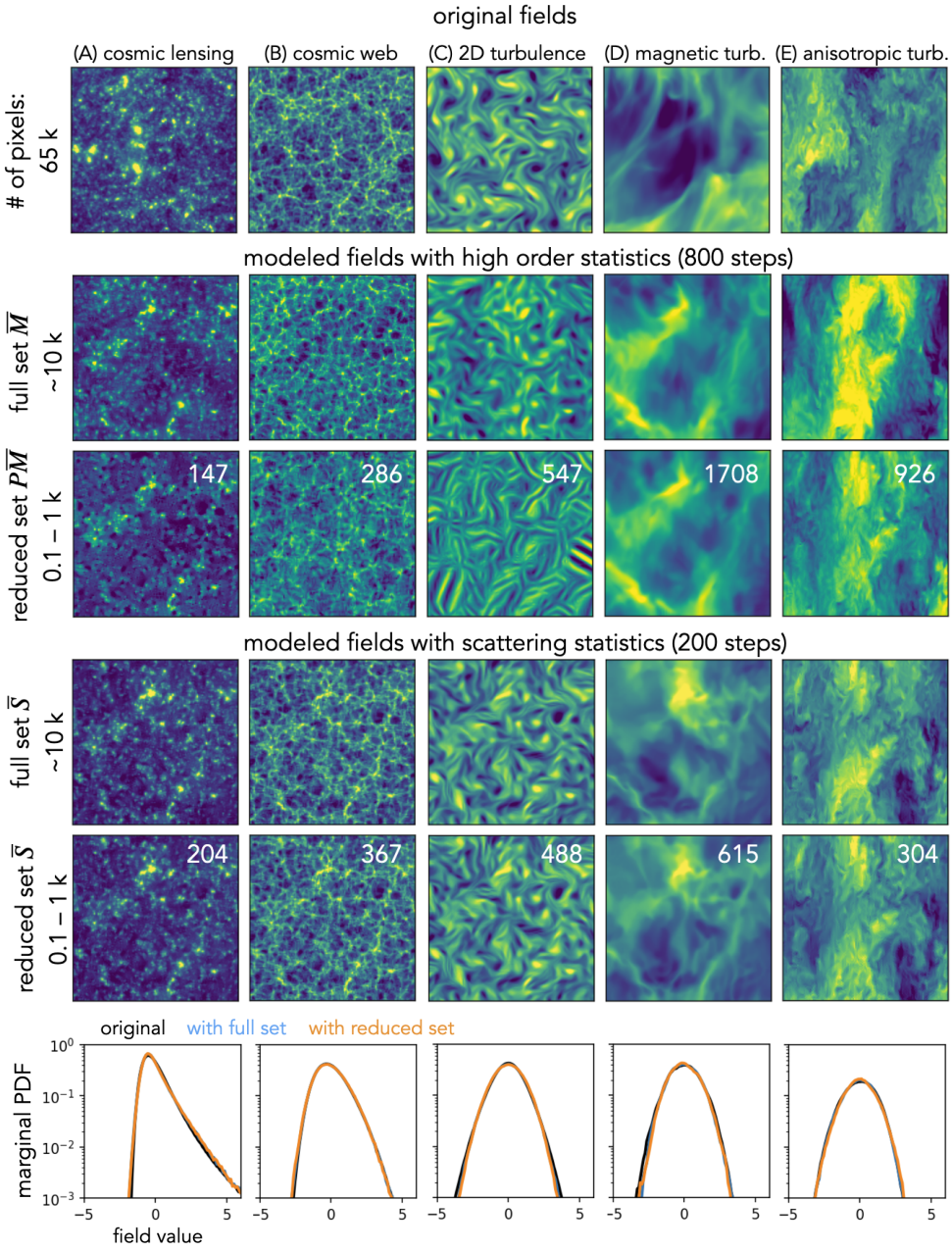


FIGURE 3.3 – Visual comparison of realistic physical fields and those sampled from maximum entropy models based on wavelet higher-order moments $\bar{M}$ and wavelet Scattering Spectra $\bar{S}$ statistics. The first row shows five example fields from physical simulations of cosmic lensing, cosmic web, 2D turbulence, magnetic turbulence, and squeezed turbulence. The second and third rows show syntheses based on the selected high-order wavelet statistics estimated from 100 realizations. They are obtained from a microcanonical sampling with 800 steps. The fourth and fifth rows show similar syntheses based on the Scattering Spectra statistics, with only 200 steps of the sampling run. This figure shows visually that the Scattering Spectra can model well the statistical properties of morphology in many physical fields, while the high-order statistics either fail to do so or converge at a much slower rate. To clearly show the morphology of structures at small scales, we show a zoom-in of 128 by 128 pixels regions. Finally, to quantitatively validate the goodness of the scattering model, we show the marginal PDF (histogram) comparison in the last row.

3.3 Numerical results

We have introduced maximum entropy models based on small subsets of $O(\log^3 L)$ Scattering Spectra moments $\bar{S}$ and projected moments $P\bar{S}$, claiming that it can provide accurate models of large classes of multiscale physical fields, and reproduce $O(L^{3d})$ power spectrum, bispectrum and trispectrum Fourier moments. This section provides a numerical justification of this claim with five types of 2D physical fields from realistic simulations. In order to reduce the variance of the validation statistics, we consider in this section a model estimated on several realizations of a field. However, our model also produces convincing realizations when estimated on a single realization (see Fig. B.2 for a visual assessment). All computations are reproducible with the software available on `https://github.com/SihaoCheng/scattering_transform`.

3.3.1 Dataset of physical fields

We use five two-dimensional physical fields to test the maximum entropy models. The five fields are chosen to cover a range of properties in terms of scale dependence, anisotropy, sparsity, and morphology :

(A) *Cosmic lensing :* simulated convergence maps of gravitational lensing effects induced by the cosmological matter density fluctuations [Matilla, 2016 ; Gupta, 2018].

(B) *Dark matter :* logarithm of 2D slices of the 3D large-scale distribution of dark matter in the Universe [Villaescusa-Navarro, 2020].

(C) *2D turbulence :* turbulence vorticity fields of incompressible 2D fluid stirred at the scale around 32 pixels, simulated from 2D Navier-Stokes equations [Schneider, 2006].

(D) *Magnetic turbulence :* column density of 3D isothermal magnetic-hydrodynamic (MHD) turbulent simulations [Allys, 2019]. The field is anisotropic due to a mean magnetic field in the horizontal direction.

(E) *Anisotropic turbulence :* two-dimensional slices of a set of 3D turbulence simulations [Li, 2008 ; Perlman, 2007]. To create anisotropy, we have squeezed the fields along the vertical direction.

These simulations are sampled on a grid of 256×256 pixels with periodic boundary conditions [4] and normalized to have zero mean and unity standard deviation, respectively. Samples of each field are displayed in the first row of Fig. 3.3. To clearly show the morphology of small-scale structures, we zoom in to a 128×128 region.

3.3.2 Model description and visual validation

We fit our maximum entropy model using wavelet polyspectra and Scattering Spectra, respectively, with the following constraint,

$$|| \left\langle \Phi(x_j) \right\rangle_j - \left\langle \Phi(\widetilde{x}_i) \right\rangle_i ||^2 \leq \epsilon \tag{3.18}$$

where the second average is computed on an ensemble of 100 realizations $\widetilde{x}_i$ for each physical simulation (for field D we use only 20 realizations due to the availability of simulations), and

4. When working without this condition, statistics can be computed by padding the images.

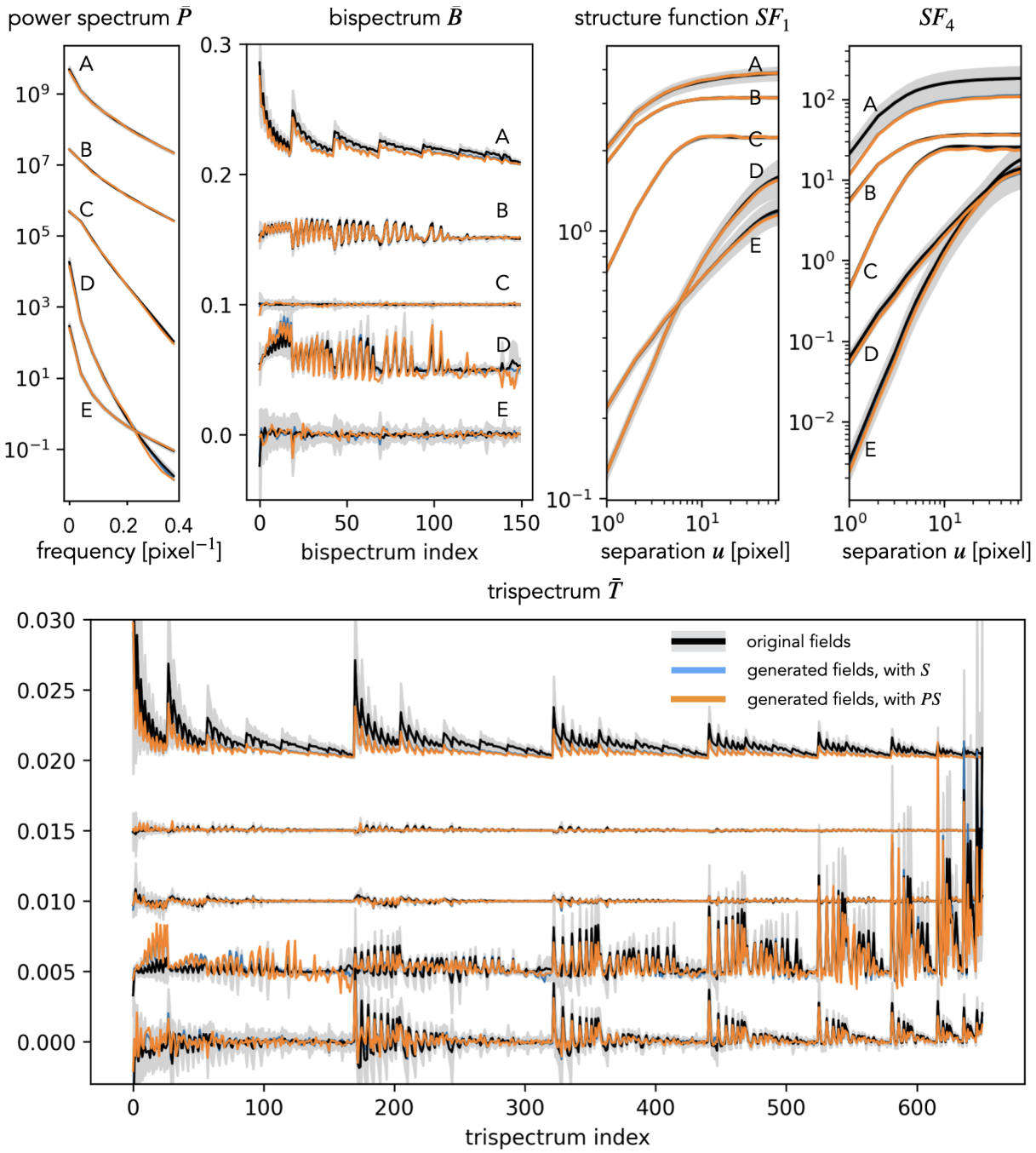


FIGURE 3.4 – Validation of the scattering maximum entropy models for the five physical fields A–E by various test statistics. The curves for field E represent the original statistics and those for A–D are shifted upwards by an offset. In general, our Scattering Spectra models well reproduce the validation statistics of the five physical fields.

the field generation is performed simultaneously for 10 fields x_j, making our microcanonical model closer to its macrocanonical limit. The microcanonical sampling algorithm is described in appendix A.6.

Examples of field generation results are given in Fig. 3.3. The second row shows samples generated based on the high-order normalized wavelet moments $\Phi(x) = \bar{M}(x) = (\bar{M}_2(x), \bar{M}_3(x), \bar{M}_4(x))$, where $\bar{M}_2 = \bar{S}_2$, $\bar{M}_3(x)[\lambda, \lambda'] = \frac{M_3(x)[\lambda,\lambda']}{\sigma^2[\lambda]\sigma[\lambda']}$ and $\bar{M}_4(x)[\lambda, \lambda'] = \frac{M_4(x)[\lambda,\lambda']}{\sigma^2[\lambda]\sigma^2[\lambda']}$ are defined similarly to $\bar{S}$ in (3.16). For the choice of wavelets, we use J=7 dyadic scales, and we set $R = 8$ which samples 4 orientations within π, resulting in $\dim \bar{M} = 11\,677$ coefficients for $\bar{M}$. The third row in Fig. 3.3 shows results from a reduced set $\Phi(x) = P\bar{M}(x)$, which is a 2σ Fourier thresholded representation of $\bar{M}$ defined in exactly the same way as $P\bar{S}$ in (3.17). The thresholding yields $\dim P\bar{M} = 147, 286, 547, 1708, 926$ for fields A–E, respectively. A visual check shows that these models fail to recover all morphological properties in our examples especially when a thresholding reduction is applied. This issue is a manifestation of the numerical instability of high-order moments.

In the fourth row, we present sample fields modeled with the Scattering Spectra $\overline{S}$ with $\dim P\bar{S} = 11\,705$ for J=7 and R=8. A visual check reveals its ability to restore coherent spatial structures including clumps, filaments, curvy structures, etc. The low-order nature and numerical stability of $\overline{S}$ also significantly fasten the sampling compared to the high-order moments $\bar{M}$ (200 vs. 800 steps to converge). The last row shows sample fields modeled by a much smaller set $P\overline{S}$, which has $\dim P\bar{S} = 204, 364, 489, 615, 304$ coefficients for fields A–E, respectively. This model is $\sim 10^2$ times smaller, while generating samples visually indistinguishable from the full set model with $\Phi(x) = \overline{S}(x)$. In addition, the ratio between the dimensionality of the field $\dim x = L^d$ (the number of pixels) and the model $\dim \Phi$ is more than 100.

3.3.3 Statistical validation

We now quantify the consistency between the Scattering Spectra models and the original fields using a set of validation statistics $V(x)$ defined below, including marginal PDF, structure functions SF_n, power spectrum P, and normalized bispectrum $\bar{B}$ and trispectrum $\bar{T}$. The validation statistics are shown in Figs. 3.3 and 3.4, where black curves represent the expected value μ_{original} of these statistics, estimated from 100 realizations $\bar{x}_i$ of the original simulated fields (except for field D for which we have only 20 realizations). Gray regions around the black curves represent the standard deviations σ_{original} of those statistics estimated on the original fields. Blue curves are statistics $\mu_{\bar{S},\text{model}}$ estimated on fields modeled with $\overline{S}$. Similarly, $\mu_{P\bar{S}\text{ model}}$ are estimated on fields modeled with the reduced set $P\overline{S}$. Both these averages are estimated from the 10 fields simultaneously sampled from the corresponding microcanonical models.

3.3.3.1 Validation statistics

The marginal probability distribution function (PDF) is measured as the histogram of sample fields and shown in Fig. 3.3. It averages out all spatial information and keeps only the overall asymmetry and sparsity properties of the field. The marginal information is not explicitly enco-

ded in the Scattering Spectra, but for all the five physical fields we examine here, it is recovered even with the reduced model $P\bar{S}$, where only $\sim 10^2$ Scattering Spectra coefficients are used.

Given that the high dimensionality of the full set of polyspectra coefficients, as well as the computational cost of estimating them properly, we adopt an isotropic shell binning for the power spectrum, bispectrum, and trispectrum. Although this reduces the number of coefficients as well as their variance, working with isotropic statistics prevents the characterization of anisotropic features, for instance in fields D and E, unlike with Scattering Spectra. Validation results with these isotropic polyspectra are given in Fig. 3.4.

The shell binning is defined as follow. We first divide the Fourier space into 10 annuli with the frequencies linearly spaced from 0 to 0.4 cycles/pixel. Then, we average the power and poly spectra coefficients coming from the same annulus combinations. For instance, the power spectrum yields :

$$P(i) = \langle \hat{x}(\omega)\hat{x}(-\omega)\rangle_{\omega \text{ in annuli } i}\,.$$

To decorrelate the information from the power spectrum and higher orders, we normalized the binned bi- and tri-spectra by $P[i]$:

$$\bar{B}(i_1, i_2, i_3) = \frac{\langle \hat{x}(\omega_1)\hat{x}(\omega_2)\hat{x}(\omega_3)\rangle_{\omega_n \text{ in annuli } i_n}}{\sqrt{P(i_1)P(i_2)P(i_3)}},$$

$$\bar{T}(i_1, i_2, i_3, i_4) = \frac{\langle \hat{x}(\omega_1)\hat{x}(\omega_2)\hat{x}(\omega_3)\hat{x}[\omega_4]\rangle_{\omega_n \text{ in annuli } i_n}}{\sqrt{P(i_1)P(i_2)P(i_3)P(i_4)}},$$

where the ω_n d-dimensional wave-vectors are respectively averaged in the i_n^{th} frequency annuli, and satisfy $\sum_n \omega_n = 0$. To clearly reveal the diversity of different type of physical fields, the trispectrum $\bar{T}$ coefficients shown in Fig. 3.4 are subtracted by the reference value of Gaussian white noise, evaluated numerically on 1000 independent realizations. Details about the numbers and the ordering of $\bar{B}$ and $\bar{T}$ are given in appendix B.4.

In Fig. 3.4 we also show the validation with structure functions, which are n-th order moments of the field increments as a function of the lag

$$SF_n[|\Delta u|] = \left\langle |x(u) - x(u - \Delta u)|^n \right\rangle_u .$$

Initially proposed by Kolmogorov for the study of turbulent flows [Kolmogorov, 1941b], they are widely used to analyze non-Gaussian properties of multiscale processes [Jaffard, 2004].

3.3.3.2 Comparison between original and modeled fields.

We quantify the discrepancy between the model and original field distributions by the outlier fraction of validation statistics outside the 2σ range,

$$|\mu_{\text{model}} - \mu_{\text{original}}|/\sigma_{\text{original}} > 2\,.$$

For each of the five types of fields, we observe the following fractions. The binned power spectrum has fractions of P : 0%, 0%, 20%, 0%, 0% for the models using all $\bar{S}$ statistics and 0%, 10%,

40%, 10%, 0% for the thresholding models with $P\bar{S}$. The power spectrum deviation of field C is likely caused by the longer convergence steps required by smooth fields, as our generative models start from white noise with strong small-scale fluctuations. Indeed increasing the steps to 800 reduces the outlier fraction of the $P\bar{S}$ model to 10%. For $\bar{B}$ and $\bar{T}$, the outlier fractions are all below 5% except for the models of field A, where the bispectrum coefficients have 13% of outliers. Those outliers all have the smallest scale involved, and disappear if the high-frequency cut is moved from 0.4 to 0.35 cycles/pixel. The low fractions demonstrate consistency between our maximum entropy models and ensembles of the original physical fields.

For field A, a similar deviation is also observed in high-order structure functions. For this field, it can be seen from Fig. 3.4 that even though many coefficients are not defined as outliers, they all tend to have a lower value than the original ones. This effect may originate from the log-normal tail of the cosmic density field [Coles, 1991], whose Gibbs potential includes terms in the form of $\log x$, in contrast to the form of $|x|$ in scattering covariance or x^n in high-order statistics. However, regardless of this difficulty, these outliers are all still within a 3σ range, demonstrating that the Scattering Spectra provide a good approximation though not exact model for fields with such heavy tails.

The marginal PDF, structure functions, power spectrum and polyspectra probe different aspects of the random field $p(x)$. The polyspectra especially probe a huge variety of feature configurations. For all the validation statistics, we observe general agreement between the model and original fields. Such an agreement is a non-trivial success of the Scattering Spectra model, as those statistics are not generically constrained by the Scattering Spectra for arbitrary random fields. They indeed significantly differ from the Scattering Spectra in the way they combine spatial information at different frequencies and in the non-linear operation adopted. The agreement implies, as we have argued, that symmetry and regularity can be used as strong inductive bias for physical fields and the Scattering Spectra, with those priors build-in, can efficiently and robustly model physical fields.

3.3.4 Visual interpretation of Scattering Spectra coefficients

The key advantage of the Scattering Spectra compared to usual convolutional neural networks is their structured nature : their computation corresponds to the combination of known scales and orientations in a fixed way. Beyond the limited number of symmetries, the structured nature of the Scattering Spectra allows us to both quantify and interpret the morphology of structures, which is one of the original goals to design these statistics.

The values of Scattering Spectra can be shown directly (see Fig. B.3) to analyze non-Gaussian properties of the field. Moreover, the meaning of its coefficients can also be visualized through our maximum entropy generative models. As one gradually changes the value of some summary statistics, the morphology of structures in the generated fields also changes. A similar exploration for a smaller set of scattering transform coefficients has been explored in [Cheng, 2021a], and we show such results with the much more expressive Scattering Spectra coefficients in Fig 3.5. Such exploration using synthesis is also similar to the feature visualization efforts for convolutional neural networks [Olah, 2017].

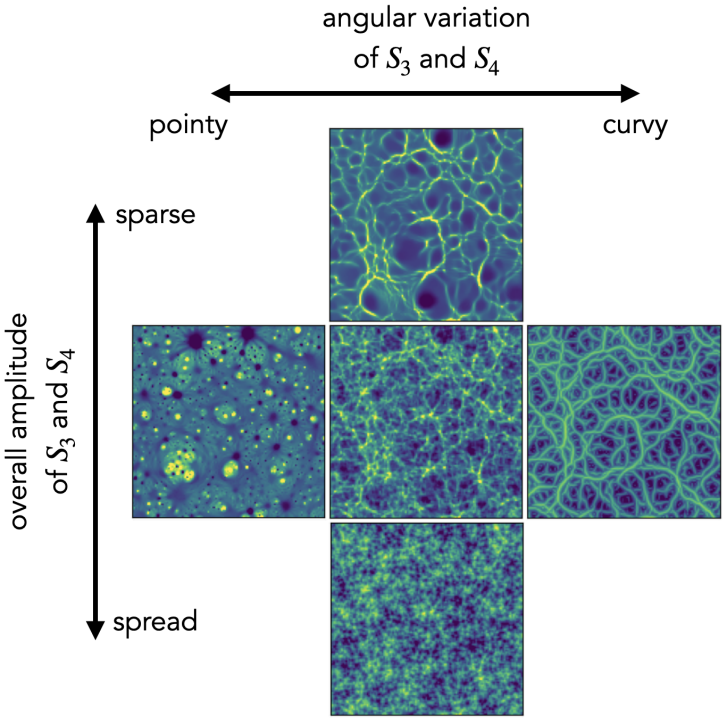


FIGURE 3.5 – Visual interpretation of the Scattering Spectra. The central field is one realization of field B in physical simulations. The other four panels are generated fields with two simple collective modifications of the Scattering Spectra coefficients.

The central panel is a realization of field B from physical simulations. The other four panels are generated fields with two collective modifications of the Scattering Spectra : the vertical direction shows the effect of multiplying all $\bar{S}_3$ and $\bar{S}_4$ coefficients by a factor of 1/3 or 3. It indicates that the amplitude of $\bar{S}_3$ and $\bar{S}_4$ controls the overall non-Gaussian properties of the field and in particular the sparsity of its structures. The horizontal direction corresponds to adjusting the orientation dependence. We set the coefficients with parallel wavelet configurations (i.e., $\bar{S}_3(x)[|\lambda|, |\lambda'|, l_1 = l_2]$ and $\bar{S}_4(x)[|\lambda|, |\lambda'|, |\gamma|, l_1 = l_2 = l_3]$) as references and keep them unchanged. Then, we make the difference from other coefficients to those references to be 2 times or –2 times the original difference. Visually, it controls whether structures are more point-like or more curvy-like in the field. In this experiment, the generated field is initialized with the original field instead of white noise, in order to clearly show the correspondence between the field structure and Scattering Spectra coefficients.

3.3.5 Application to identifying symmetry

As an expressive representation whose coefficients are equivariant under standard group transformation, the Scattering Spectra can also be used to detect and identify the various statistical invariances commonly present in physical fields. Besides the aforementioned rotation and scaling invariance, more can also be included, such as the flipping of coordinate or field values.

The simplest way to check asymmetry to a transformation like rotation or flip is to check if the Scattering Spectra S are changed after applying such a transform. A more sophistica-

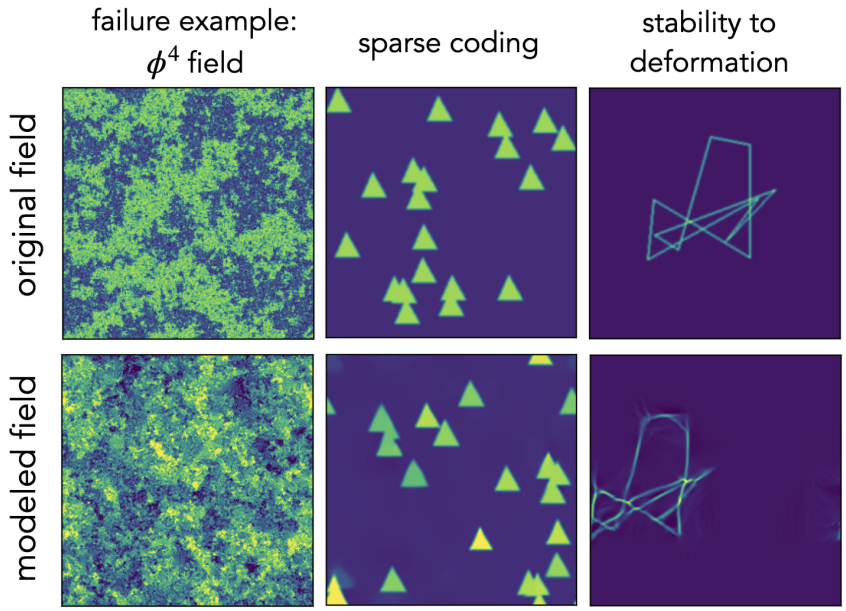


FIGURE 3.6 – Example of failures and applications beyond typical physical fields. The modeled fields of the central and right panels have been recentered for easier comparison with the original ones.

ted way that can also quantify partial symmetries is to linearly decompose $\bar{S}$ into symmetric and asymmetric parts and then compute the fraction of asymmetric coefficients surviving the thresholding reduction. We further normalize this fraction by that in the full set :

$$\text{asymmetry index} = \frac{\dim(P\bar{S}_{\text{asym}})}{\dim(P\bar{S})} \Big/ \frac{\dim(\bar{S}_{\text{asym}})}{\dim(\bar{S})}.$$

When it is zero, the random field $p(x)$ should be invariant to the transform up to the expressivity of our representation. For the five random fields analyzed in this study, we measure their asymmetry indices with respect to rotation and scaling. The corresponding anisotropy and scale dependence indices are (A) 0, 0.16 ; (B) 0, 0.53 ; (C) 0, 0.66 ; (D) 0.32, 0.45 ; (E) 0.28, 0.29. As expected, the cosmic lensing field (field A), which consists of haloes at all scales and strengths, is closest to isotropic and scale-free. The cosmic web (B) and 2D turbulence (C) fields are isotropic but have particular physical scales above which the field becomes Gaussian, so they are not scale-free. The last two turbulence fields have anisotropic physical input, but the latter largely probes the 'inertial regime' of turbulence, which is scale-free.

3.3.6 Limitations

While a broad range of physical fields satisfy the implicit priors of the scattering covariance, one does expect regimes for which the description will not be appropriate. The so-called φ^4 field in physics comes as a first problematic example. It is the maximum entropy field under the power spectrum and pointwise fourth-order moment x^4 constraints, but this characterization is unstable to specify a non-convex pdf which is a pointwise property as opposed to the delocalized

Fourier moments and it is highly unstable at critical points [Marchand, 2022]. The first column in Fig. 3.6 shows an original φ^4 field at its critical temperature and that generated from the full set of scattering covariance. In contrast to previous examples, this type of field is not successfully reproduced. More generally, our models fail to reproduce processes whose histograms are not regular, as in the second example of Fig. 3.6.

For physical fields with multi-scale structures, it is expected that the distribution function $p(x)$ does not change much under a slight deformation. When modeling such fields, it is important to have a representation that has the same property. Being built from wavelet decomposition and contracting operator, the Scattering Spectra also linearize small deformation in the field space, which plays an important role in lowering its variance (see [Bruna, 2013]). However, when modeling structured fields whose distribution functions are not regular under deformation, this means that the generative model will simply produce structures that are "close enough" up to small deformations. This typical type of failure is shown in the third example of Fig. 3.6.

3.4 Conclusion

We build maximum entropy models for non-Gaussian random fields based on the Scattering Spectra statistics. Our models provide a low-dimensional structured representation that captures key properties encountered in a wide range of stationary physical fields, namely : (i) stability to deformations as a result of local conservation laws in Physics for mass, energy, momentum, charge, etc ; (ii) invariance and regularity to rotation and scaling ; (iii) scale interactions typically not described by high-order statistics ; Those are the priors included in the Scattering Spectra.

Our models provide a practical tool for generating mock fields based on some example physical fields. In sharp contrast to neural network models, our representation has the key advantage of being interpretable and can be estimated on a few realizations. This is crucial in Physics where generating fields in experiments or simulations is costly or when non-stationarity limits the amount of clean recorded data. Our proposed approach enables a new range of data/simulation analyses [e.g. Regaldo-Saint Blancard, 2021 ; Delouis, 2022], involving extensions to the modeling of cross-regularities when multiple channels are available [e.g. Régaldo-Saint Blancard, 2023]. Next chapter considers a second extension towards multi-scale processes.

Chapitre 4

Models of multi-channel time-processes

Foreword

Many processes encountered in Medecine, Finance but also Physics, are recorded as a collection of time-series $x(t) = (x_1(t), \dots, x_C(t))$, also called multi-channel time-series. They exhibit a time variable and a channel variable which identifies the specific time-series. These are multivariate processes of a different type than in the previous chapter since the true notion of distance for this second variable, if it exists, may not be accessible. We thus cannot build upon the models introduced in the previous chapter. Instead, we leverage the Scattering Spectra models of univariate time-processes introduced in chapter 2. We introduce a model of multi-channel time-processes that can be estimated on limited data, in the specific case of the different stock prices of a financial index. For that, the process x is projected on selected directions across channels. The projected processes are called factors, they are univariate time-processes. Our model then constrains the time structure of such factors, through the Scattering Spectra. The key challenge is to find the few factors, whose time structure are the most informative on the joint structure of the process. By choosing a sparse basis along channels, we obtain a model that captures important non-linear dependencies across stocks, including order 3 moments and copula statistics introduced in the literature.

Contents

4.1 Introduction

Many multi-scale processes x of interest in Medicine, Physics and Finance are recorded as multi-channel time-series $x(t) = (x_1(t), \dots, x_C(t))$ with C channels, possibly large $C \gg 1$. In Medicine, one can think of electroencephalogram recordings of different areas of the brain. In Finance, one can think of the different stock prices of individual companies composing a price index. In general, the channel index c in $x_c(t)$ is of different nature than the time index t. The closest neighbors of a discrete time t are the times $t-1$ and $t+1$, but what are the "closest" channels of a time-series channel c? This requires a notion of proximity that is hardly accessible in the case of different stocks.

The goal in this chapter is to build models of multi-channel multi-scale time-processes that capture dependencies across channels and can be estimated from limited data. This chapter is thus a second extension of chapter 2 towards multivariate processes. We will focus on the specific case for which x is a financial index where the different channels n are the stocks composing it, e.g. the American index, and $x_c(t)$ is the log-price of stock c at date t. This problem is crucial for financial actors generally trading on multiple stocks at the same time and who are thus exposed to the joint distribution of the process x.

Dependencies across stocks can be detected through the linear correlation matrix, which is at the heart of portfolio allocation theory [Markowitz, 1952; Bouchaud, 2003]. Beyond requiring careful cleaning of such matrix whose estimation on limited data is a challenge [Potters, 2005; Tumminello, 2007], these moments do not capture important non-Gaussian properties as evidenced in [Chicheportiche, 2014b].

Factor models [Fama, 1993] decompose linearly the stocks on a small number of factors $f(t)$, with residuals $e(t)$, that evolve over time, $x(t) = \beta f(t) + e(t)$. The weights β parameterize the exposure of each stock to the common factors. For example, the weights and factors can be obtained through principal component analysis. Factor models then assume a certain stochastic structure for each factor and residuals, which are univariate time-processes, so as to capture important properties of the joint process x. While such models capture essential non-Gaussian properties of the process, they often make simplifying assumptions on the stochastic structure of the factors, for example they may not capture the joint time-asymmetry of the stocks x [Reigneron, 2011; Chicheportiche, 2015]. The main challenge is to find the few factors whose stochastic time structures, that should be accurately modeled, rule the joint stochastic structure of x.

Models of multi-channel multi-scale time-processes can be defined as a maximum entropy distribution conditioned by a vector of moments $\mathbb{E}\{\Phi(x)\}$ with $\Phi : \mathbb{R}^{C\times T} \mapsto \mathbb{R}^M$. If they exist, they have an exponential probability distribution

$$p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \theta, \Phi(x)\rangle}.$$

for $x = (x_1, \dots x_C) \in \mathbb{R}^{C\times T}$ and $\theta \in \mathbb{R}^M$, where M is the number of moments. Maximum entropy models depend only on the energy vector $\Phi(x)$, which needs to be chosen appropriately so as to specify dependencies across channels. Gaussian processes are maximum entropy models conditioned by first and second order moments. As in the last chapters, for estimation

and sampling purposes, we will consider microcanonical maximum entropy model which have a maximum entropy distribution on the set $\Omega_\epsilon = \{x \in \mathbb{R}^{C\times T} \mid \|\Phi(x) - \Phi(\widetilde{x})\| < \epsilon\}$ where $\widetilde{x}$ is the unique historical realization of stock prices. We refer to appendix A.6 for more details on microcanonical models.

The main contribution of this chapter is a maximum entropy factor model of stocks that captures important non-linear dependencies across stocks in the literature, including certain time-asymmetries evidenced from order 3 statistics across stocks. Our model of x makes use of the Scattering Spectra introduced in chapter 2 for univariate processes, by constraining the Scattering Spectra of certain selected factors, which are projections of the process x along given directions. Its dimension M scales in $\mathcal{O}(C \log_2^3 T)$ with the number C of channels and T of time steps. It can thus be estimated on a single realization $\widetilde{x} \in \mathbb{R}^{C\times T}$ of limited size.

Section 4.2 presents a simple model in which dependencies across stocks are imposed through linear correlation only. This, to better outline the limitations of linear correlations across stocks. Section 4.3 introduces our maximum entropy factor model. We investigate two types of factors, the first ones are based on principal component directions, well-studied in Finance. The second are based on sparse directions obtained via dictionary learning. Section 4.4 validates our proposed factor model by estimating standard non-linear statistics across stocks which are not directly imposed in our model.

4.2 Dependencies across channels through linear correlation

Maximum entropy models of multi-channel time-series can be written on the form

$$p_\theta(x) = Z_\theta^{-1}\, e^{-\langle \beta, \Phi_{\text{single}}(x)\rangle - \langle \gamma, \Phi_{\text{cross}}(x)\rangle}. \tag{4.1}$$

where β, γ are real vectors and $\Phi_{\text{single}}(x)$ depends solely on the univariate time-processes $x_1, \dots, x_C$ while $\Phi_{\text{cross}}(x)$ may depend on the joint distribution of $x = (x_1, \dots, x_C)$. A microcanonical maximum entropy model has a maximum entropy distribution on the set

$$\Omega_\epsilon = \{x \in \mathbb{R}^{C\times T} \mid \|\Phi(x) - \Phi(\widetilde{x})\| < \epsilon\} \tag{4.2}$$

where $\Phi = \big(\Phi_{\text{single}}(x), \Phi_{\text{cross}}(x)\big)$ and $\widetilde{x}$ is the unique historical realization of stock prices. They can be sampled through an approximate gradient descent algorithm (see appendix A.6).

Chapter 2 focused on building models of univariate processes. Section 4.2.1 builds on this work to define a $\Phi_{\text{single}}(x)$ which characterizes the univariate stochastic structure of single stocks. In order to build a model of the joint process, section 4.2.2 proposes a choice for $\Phi_{\text{cross}}(x)$ based on the linear correlations across stocks. This strategy has the advantage of dealing with a well-studied object that reveals information of sectors for example. In section 4.2.3 we investigate the ability of such model to capture dependencies across stocks.

4.2.1 Univariate distribution of stocks

We write $x_c(t)$ the log-price of stock c on day t. Its increment on day t is called *log-return* and is written

$$\delta x_c(t) = x_c(t) - x_c(t-1).$$

We assume that the log-return vector $\delta x(t) = (\delta x_1(t), \dots, \delta x_C(t))^T$ is a stationary process. Without loss of generality we also assume that different stocks have the same zero-trend $\mathbb{E}\{\delta x_c(t)\} = 0$ and are of constant average volatility $\mathbb{E}\{|\delta x_c(t)|^2\} = 1$. We write I the index of the stock

$$I(t) = \frac{1}{C}\sum_{c=1}^{C} x_c(t) \tag{4.3}$$

where we choose to average uniformly the stocks, regardless of their capitalization.

In chapter 2 we built a representation $\Phi : \mathbb{R}^T \mapsto \mathbb{R}^{M_1}$, the Scattering Spectra, with M_1 coefficients, which were shown to capture essential properties of a univariate multi-scale time-process x_1. A univariate model of x_c for $1 \leq c \leq C$ can be built through the statistics $\Phi(x_c)$. Single stocks share common properties, for example the absence of auto-correlation over time, that would enable arbitrage. In this chapter instead of modeling the time structure of each stock individually, which is costly, we instead constrain the average distribution of single stocks by computing the average Scattering Spectra over stocks

$$\Phi_{\text{single}}(x) = \langle \Phi(x_c) \rangle_c . \tag{4.4}$$

These statistics are invariant to permutations of the stocks. They characterize the average time structure of single stocks. This average improves the estimation of $\mathbb{E}\{\Phi(x_c)\}$ and reduces by a factor C the number of coefficients used to model the distributions of individual stocks.

4.2.2 Correlation and principal directions

We aim at defining cross-statistics $\Phi_{\text{cross}}(x)$ that characterize dependencies across stocks and that can be estimated on limited data $\widetilde{x} \in \mathbb{R}^{C \times T}$. Linear correlation $\Sigma = \mathbb{E}\{\delta x(t)^T \, \delta x(t)\}$ between stocks at same time t is a standard choice in the literature. Its estimation through empirical average $\Phi_{\text{cross}}(x) = \left\langle \delta x(t)^T \delta x(t) \right\rangle_t$ is difficult on a single realization $\widetilde{x}$ because the matrix Σ contains C^2 coefficients which is fairly significant compared to the number of data points in $\widetilde{x}$ of size $C \times T$. The number of stocks C is typically of a few hundreds and the number of days T is typically of the order of a few thousands.

A stationary Gaussian process of zero-mean is fully characterized by the linear correlation matrix Σ. It has a probability distribution of the form (4.1) with $\beta = 0$ and $\langle \gamma, \Phi_{\text{cross}}(x) \rangle = \frac{1}{2} \left\langle \delta x(t) \Sigma^{-1} \delta x(t)^T \right\rangle_t$, provided that Σ is invertible. The empirical correlation matrix is diagonalized in the PCA basis that we note P. Its columns, $v^1, \dots, v^N$, are the PCA vectors. One can show that a Gaussian model is a maximum entropy distribution under energy constraints along PCA directions

$$\Phi_{\text{cross}}(x) = \left(\left\langle |\langle v^1, \delta x(t) \rangle|^2 \right\rangle_t, \dots, \left\langle |\langle v^r, \delta x(t) \rangle|^2 \right\rangle_t \right) \tag{4.5}$$

with $r = C$, the number of stocks. Indeed, a maximum entropy model under such constraint satisfies that $P^T \delta x^T$ is a multivariate independent Gaussian process. In that specific case, the number of constraints is C instead of C^2. Of course, this requires determining the principal component directions accurately, which is a challenge.

In a general non-Gaussian case, instead of imposing the correlations one by one, we also choose to impose correlation through the energy constraint along PCA directions (4.5), where r can be strictly lower than C, depending on the regularity of the spectrum of the correlation matrix Σ.

4.2.3 Failure to capture joint non-Gaussianity

We construct a benchmark microcanonical model with $\Phi(x) = (\Phi_{\text{single}}(x), \Phi_{\text{cross}}(x))$ where $\Phi_{\text{single}}(x)$ is the average Scattering Spectra over stocks (4.4), and $\Phi_{\text{cross}}(x)$ characterizes dependencies across stocks through linear correlation by imposing the energy along the $r = 10$ first PCA directions (4.5). This model is non-Gaussian because the single stocks have a non-Gaussian distribution characterized by $\Phi_{\text{single}}(x)$. It is estimated on a realization of the $C = 253$ S&P stocks on $T = 3269$ days from January 2000 to December 2012. Fig. 4.1 shows that the 10 first directions account for most of the variance.

First, we verify that the PCA directions are correctly captured, as it should be in the Gaussian case. Fig. 4.1 shows that the first two PCA vectors are approximately captured by the benchmark model.

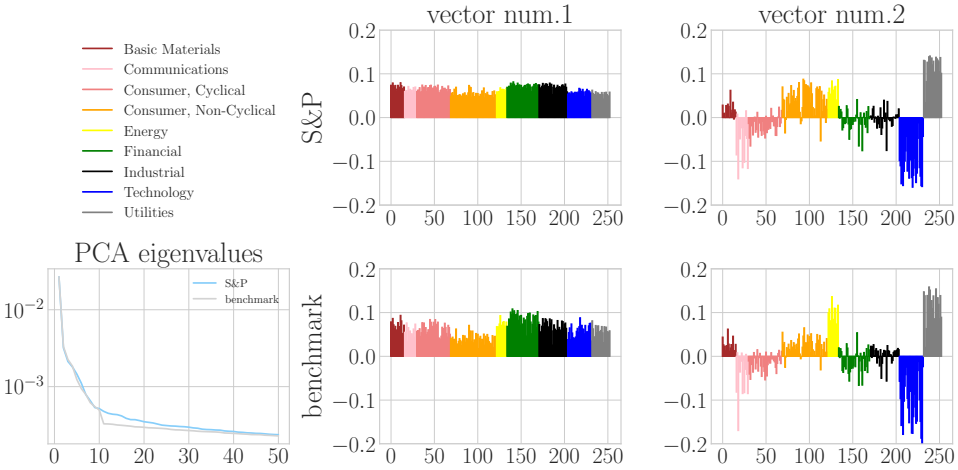


FIGURE 4.1 – Two first PCA directions estimated on the single S&P realization $\tilde{x}$ and in the benchmark model which characterizes dependencies across stocks through linear correlation only.

In order to investigate non-linear dependencies across stocks we consider the index leverage effect introduced in [Reigneron, 2011]. The index leverage is the standard leverage effect [Bekaert, 2000 ; Bouchaud, 2001] measured on the stock index I (4.3). It states that the correlation

$$\mathcal{L}_I(\tau) = \frac{\langle \delta I(t-\tau)\, |\delta I(t)|^2 \rangle_t}{\langle |\delta I(t)|^2 \rangle_t} \tag{4.6}$$

is asymmetrical in $\tau = 0$ as a consequence of the time-asymmetry of I : negative index log-return values in the past tend to be correlated to increase in volatility in the future.

Authors in [Reigneron, 2011] decompose this effect on stock indices into two contributions. We write $\sigma(t)$ and $\rho(t)$ respectively the instantaneous stock volatility and instantaneous correlation between all pairs

$$\sigma(t)^2 = \frac{1}{C}\sum_{c=1}^{C} |\delta x_c(t)|^2 \quad , \quad \rho(t) = \frac{1}{C(C-1)} \sum_{c \neq c'=1}^{C} \frac{\delta x_c(t)\delta x_{c'}(t)}{\sigma^2(t)}$$

Authors in [Reigneron, 2011] evidenced the presence of two partial leverage effects

$$\mathcal{L}_\sigma(\tau) = \frac{\langle \delta I(t-\tau)\, |\sigma(t)|^2 \rangle_t}{\langle |\delta I(t)|^2 \rangle_t} \quad , \quad \mathcal{L}_\rho(\tau) = \frac{\langle \delta I(t-\tau)\, \rho(t) \rangle_t}{\langle |\delta I(t)|^2 \rangle_t}.$$

The first one measures the following time-asymmetry : negative index log-returns tend to be followed by increase in the overall simultaneous volatility. The second states that negative index returns also tends to be followed by increase in the correlation of the stock pairs, interpreted as a panic effect.

Leverage statistics $\mathcal{L}_I(\tau)$, $\mathcal{L}_\sigma(\tau)$ and $\mathcal{L}_\rho(\tau)$ are combinations of order 3 moments across channels at separate times. $\langle \delta x_k(t-\tau)\delta x_c(t)\delta x_{c'}(t) \rangle_t$. They are shown on Fig. 4.2. The benchmark model poorly replicates these statistics. This shows the limitations of a model based solely on the linear correlations.

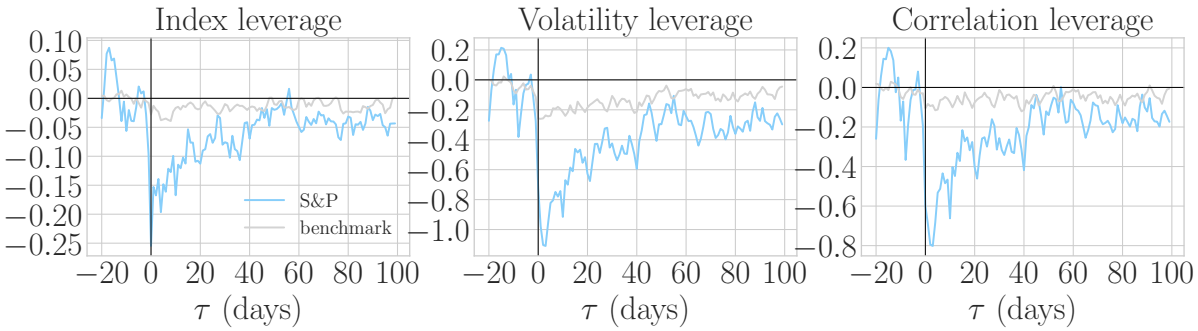


FIGURE 4.2 – Index leverage correlation $\mathcal{L}_I(\tau)$ with its volatility and correlation contributions $\mathcal{L}_\sigma(\tau)$, $\mathcal{L}_\rho(\tau)$, estimated on the $S\&P$ and in the benchmark model. The leverage effect, both its amplitude and asymmetry are poorly captured by the benchmark.

4.3 Factor model based on sparse directions

In order to capture non-linear dependencies across stocks, one could consider high-order moments. However their estimation on limited data is hard because of the variance induced by large events and their number is exponential in the number of stocks C.

Instead, factor models try to capture non-linear dependencies across stocks by considering the projections of the process along few selected directions. The projected process, that becomes

a univariate time-process is called a factor. A main challenge is to identify the few factors, whose univariate time structure should be constrained in a maximum entropy model so as to accurately approximate the distribution of the joint process x.

Section 4.3.1 introduces maximum entropy models based on selected factors. The factors can be defined from PCA directions, that are delocalized directions across stocks. In section 4.3.2 we introduce directions of another nature, sparse directions, obtained via dictionary learning.

4.3.1 Maximum entropy factor model

Given a vector $w = (w_1, \dots, w_C) \in \mathbb{R}^C$ along stocks, the projected process along this direction, called a factor, is a univariate process

$$\langle w, x \rangle(t) = \sum_{c=1}^{C} w_c x_c(t). \tag{4.7}$$

The factors are typically non-Gaussian processes. For example, a direction $w = (0, \dots, 1, \dots 0)$ with all zeros, except for one coordinate equal to 1, yields the single stocks which are non-Gaussian [Fama, 1965]. The direction $w = (1, \dots, 1)$ yields the index $I(t)$ that is also non-Gaussian. More generally, the abundance of co-jumps in stock markets [Bormetti, 2013], which are price jumps occurring simultaneously on multiple stocks, reveals that certain sparse directions w yield non-Gaussian factors that help describe the joint distribution of the process.

Given r directions $w^1, \dots, w^r \in \mathbb{R}^C$, a maximum entropy factor model of the joint process x constrains statistics on the r factors $\langle w^1, x \rangle, \dots, \langle w^r, x \rangle$ through

$$\Phi_{\text{cross}}(x) = \Big(\Phi(\langle w^1, x \rangle), \dots, \Phi(\langle w^r, x \rangle) \Big) \tag{4.8}$$

It is an extension of the statistics (4.5) that only considered the average volatility of the factors. The main challenge is to find the directions $w^1, \dots w^r$ such that the microcanonical model supported on (4.2), which constrains the time structure of the corresponding factors through (4.8), is a good approximation of the joint distribution p, while using as few factors as possible in order to reduce the variance of $\Phi_{\text{cross}}(x)$.

If x is a multivariate Gaussian process then any factor $\langle w, x \rangle$ is a Gaussian process. In that case, if r is the rank of the linear correlation matrix, one needs no more than r factors for model (4.1) to describe exactly the joint distribution of x.

The previous section considered PCA directions which provide a certain information on the dependencies across stocks. Fig. 4.3 shows the Scattering Spectra $\Phi(\langle v, x \rangle)$ of the projected process $\langle v, x \rangle$ averaged on the first 10 PCA directions v, averaged on the following 40 and the remaining 203. It shows that the further the PCA vector, the more Gaussian is the projected process, up to the Scattering Spectra. Indeed, a Gaussian process has constant sparsity factors Φ_1, zero cross-spectrum $|\Phi_3|$, and flat cross-spectrum $|\Phi_4|$. In particular, while the 10 first factors exhibit sign-asymmetry (measured by $|\Phi_3|$) and time-asymmetry (measured by $\operatorname{Arg} \Phi_3$), the next 50 PCA factors do not exhibit time-asymmetry any more ($\operatorname{Arg} \Phi_3 \approx 0$) and the 200 remaining

PCA factors do not exhibit either of these characteristics.

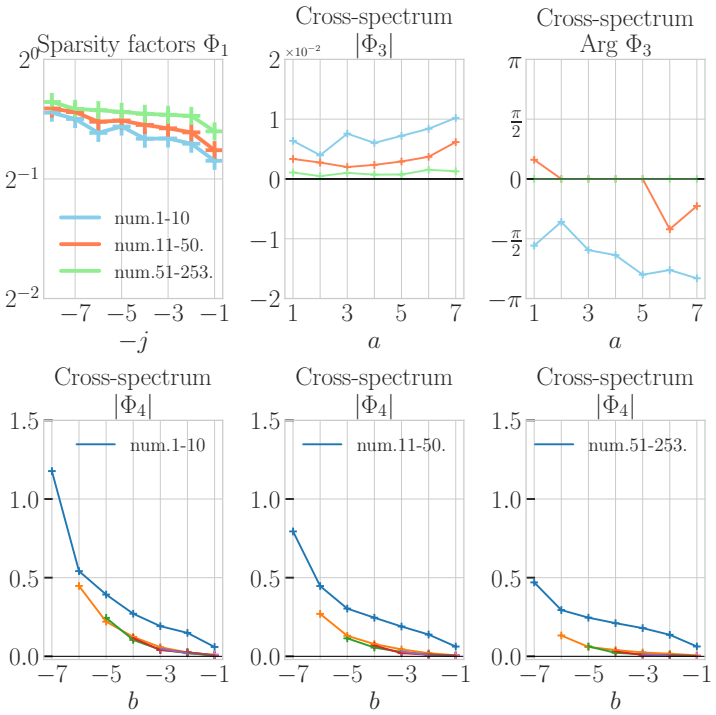


FIGURE 4.3 – Scattering Spectra $\Phi(\langle v, x\rangle)$ of PCA factors $\langle v, x\rangle$. These are averaged on the 10 first PCA vectors, the 40 next vectors and the 203 remaining vectors. The first PCA factors are highly non-Gaussian. The further the PCA vector, the more Gaussian is the process x projected on this direction. The full interpretation of this dashboard is provided under section 2.5 of chapter 2.

Thus, the direction that the most "factorize" the variance of δx are also more non-Gaussian. This suggests considering a microcanonical model whose dependencies across stocks are specified by (4.8) with $w^1, \dots, w^r$ the $r = 10$ first PCA directions that we call PCA factor model.

4.3.2 Sparse directions

As mentioned above, there exist vectors $w \in \mathbb{R}^C$ with a small number of non-zero coordinates, called sparse directions, such that the factor $\langle w, x\rangle$ has large events, called co-jumps. This proves that there exist specific sparse directions w that may reveal non-Gaussian structure across stocks, but how to find them ?

The PCA vectors are generally delocalized in two extents. First, the PCA vectors are not sparse and generally affect all coordinates significantly, see Fig. 4.5. Second, at each time t, all the PCA vectors are generally affected by $x(t)$ i.e. $P^T x(t)$ is not sparse, see Fig. 4.4.

Let us write $D \in \mathbb{R}^{C\times r}$ a matrix whose columns are the directions $w^1, \dots, w^r$. One can show that the r first PCA vectors satisfy the following optimization problem [Mallat, 1999]

$$\arg\min_{D} \ \mathbb{E}\{\|x - DD^T x\|_2^2\} \tag{4.9}$$

where we dropped here the dependence in t and considered x as a random column vector. The vector $D^T x$ contains the projection of vector x on the r directions that are the columns of D.

To promote sparsity [Donoho, 2006], a *dictionary* [Olshausen, 1996 ; Raina, 2007] optimizes the following loss with sparsity parameter $\lambda > 0$

$$\arg\min_D \ \mathbb{E}\{\frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1\}. \tag{4.10}$$

where $z = z(x, D, \lambda)$ is a certain function of x, D and λ. Optimization problem (4.9) is obtained by taking $\lambda = 0$ and $z = D^T x$. Vector z is called a sparse code and can be chosen so as to minimize the following loss

$$z(x, D, \lambda) = \arg\min_z \ \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1 \tag{4.11}$$

This problem is known as basis-pursuit [Chen, 2001].

In order to simplify the search for a dictionary D which is solution to the optimization problem (4.10) with the above choice of z we assume that the columns of D are orthogonal : $D^T D = I_r$ which imposes $r \leq C$.

Solving the optimization problem (4.10) can be done by an alternating direction method [Lin, 2011]. It alternates optimization on variable z for D fixed and on variable D for z fixed. For D fixed, one can prove that the optimal z solution to (4.11) is $z = \rho_\lambda(D^T x)$ where $\rho_\lambda(u) = \text{sign}(u)\max(0, |u| - \lambda)$ for $u \in \mathbb{R}$ is the soft-thresholding operator [Zarka, 2019] which is applied on each coordinate of vector $D^T x$. This operation puts to zero the smallest values in $D^T x$, this is a sparsity step. For z fixed, the optimal D solution to (4.10) is obtained through a SVD of the matrix $\mathbb{E}\{xz^T\} = U\Delta V^T$ trough $D = UV^T$ [Lin, 2011]. The optimization algorithm 1 consists

Algorithm 1 Learning sparse directions

Require: stocks price realization $\widetilde{x}$ cut into several days, sparsity threshold λ
  $D \leftarrow Id$
  $z \leftarrow 0$
  for step = 1 to 3000 **do**
    Step1. (sparse coding)
    $z \leftarrow \rho_\lambda(D^T \widetilde{x})$
    Step2. (reconstruction)
    $U, V \leftarrow \text{SVD}(\widetilde{x} z^T)$
    $D \leftarrow UV^T$
  end for
  return D (learned dictionary)

in alternating between a sparse-coding step (soft-thresholding) and a $\|\cdot\|_2$-reconstruction step (SVD).

For a given value of λ we obtain an approximate solution to 4.10. We then choose the λ so as to minimize the value of $\mathbb{E}\{\|D^T x\|_1\}$. We obtain $\lambda \approx 0.375$. We now refer to the columns of such optimal D obtained with $r = C$ as *sparse directions*.

We compare the PCA and sparse directions. Fig. 4.4 shows the projections $D^T x(t)$ for dif-

ferent times t along the first 50 vectors for the PCA and sparse bases on 500 selected days. As expected, at a given day t only a few sparse directions are activated, while more PCA directions are chosen. In particular, we see that the map

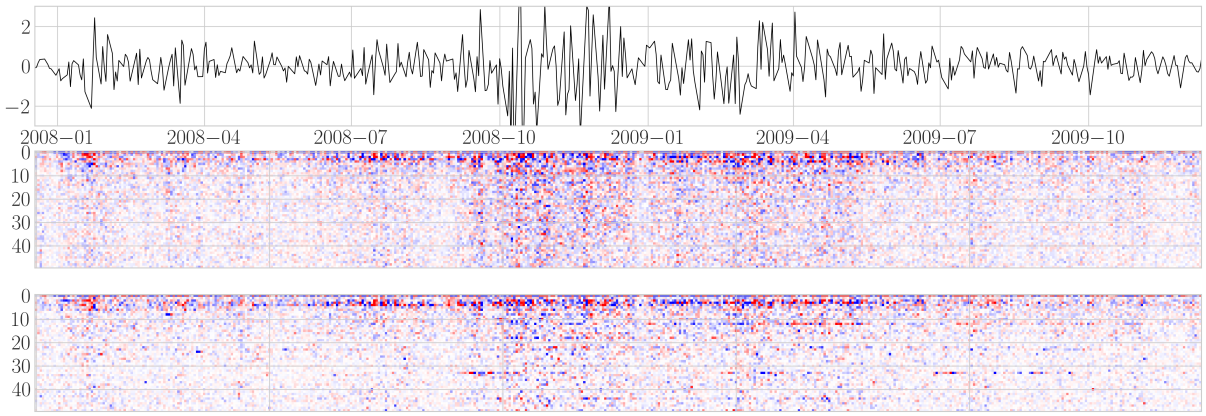


FIGURE 4.4 – Activation map $D^T x(t)$ of the first 50 vectors in D for several times t and for two choices of bases D : PCA or sparse basis.

Fig. 4.5 compares the PCA directions with sparse directions. While the first two sparse directions almost coincide with the PCA directions, the following directions becomes sparser and have supports localized on a few stocks, while the PCA vectors remain delocalized. The support of those sparse vectors seem to contain information on the dependencies across stocks. For example, the four largest coordinates on the 9^{th} sparse direction (see Fig.4.5) correspond to four stocks among which two are health maintenance organizations, one is a medical instrument company and one is a pharmacy service company.

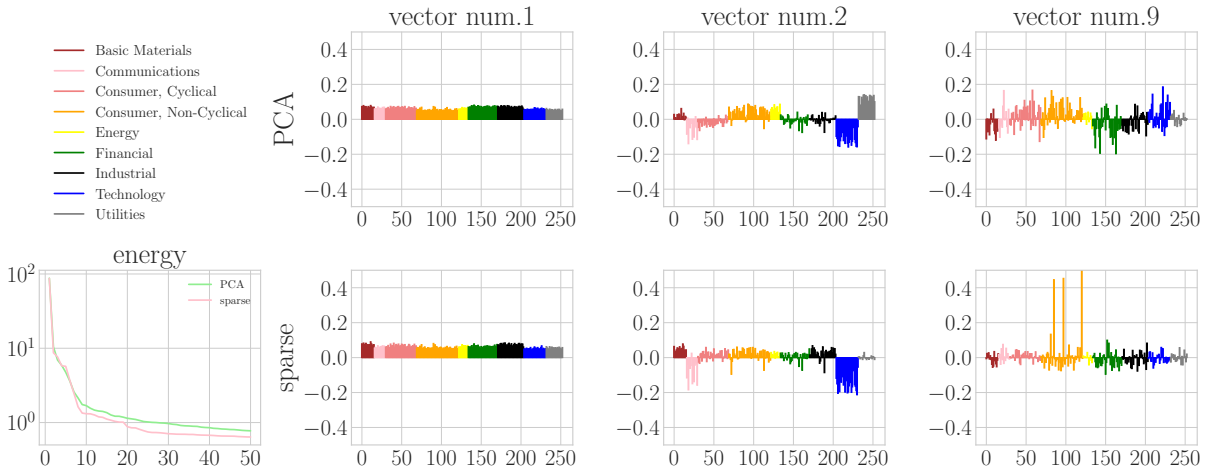


FIGURE 4.5 – Comparison between selected PCA vectors (top) and sparse basis (bottom). While the first vectors are close, sparse directions often have sparse coordinates. The largest coordinates corresponding to stocks whose classification are similar.

Next section assesses the accuracy of a maximum entropy factor model based on PCA directions or sparse directions.

4.4 Numerical validation

This section evaluates the accuracy of a maximum entropy factor model based on statistics $\Phi_{\text{single}}(x)$ (4.4) and $\Phi_{\text{cross}}(x)$ (4.8). We compare a factor model that chooses the first $r = 10$ PCA directions to a factor model that chooses the first $r = 10$ sparse directions. Both models contain 1914 order 1 or order 2 coefficients, which corresponds to an average of 8 coefficients per stock, and can thus be estimated on limited data. We evaluate these models using *test moments* estimated on the observed S&P stock time-series $\widetilde{x} \in \mathbb{R}^{C\times T}$ and estimated on time-series generated by the model. In addition to the index leverage reviewed in section 4.2.3, we describe test moments based on copulas used in the literature to evidence non-linear dependencies across stocks.

4.4.1 Non-linear statistics

The index leverage effect (4.6) presented in section 4.2.3 is a moment of order 3 across stocks. Fig., 4.6 shows estimates in a PCA factor model and a sparse factor model. Compared to the

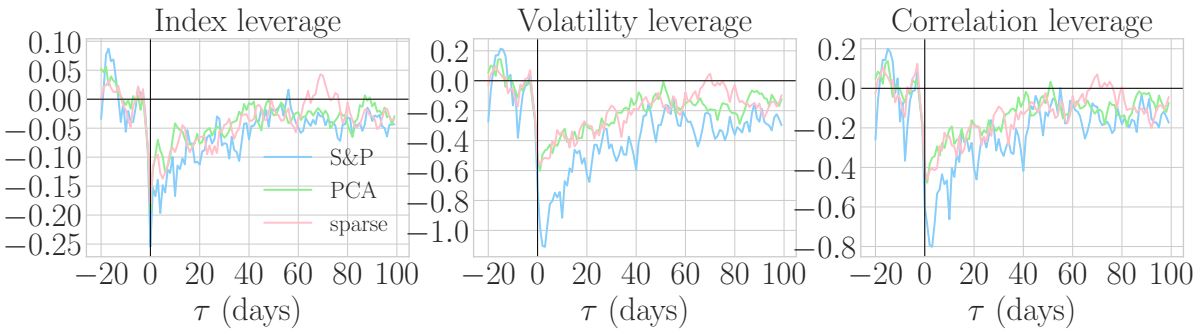


FIGURE 4.6 – Index leverage correlation $\mathcal{L}_I(\tau)$ with its volatility and correlation contributions $\mathcal{L}_\sigma(\tau)$,$\mathcal{L}_\rho(\tau)$, estimated on the *S&P* and in two maximum entropy factor models, one using PCA directions, the other using sparse directions.

benchmark model (see Fig. 4.2), a maximum entropy factor model based on either PCA or sparse directions captures the index leverage. There is also a volatility leverage asymmetry and correlation leverage asymmetry in our model, whose amplitude is partially reproduced by the two models. The fact that the two models reproduce this time-asymmetry equally is to be linked to the two previous observations. The first PCA and sparse direction coincide (see Fig. 4.5) and these are the first factors that bear the more time-asymmetry (see Fig. 4.3).

Copulas have been used to evidence non-linear dependencies across stocks [Chicheportiche, 2014b]. The copula of a pair of random variables (X, Y), typically a pair of stocks, is the joint cdf of $F_X^{-1}(X)$ and $F_Y^{-1}(Y)$ where F^{-1} is an inverse of the cdf

$$C_{(X,Y)}(p, q) = \mathbb{P}(X \leq F_X^{-1}(p), Y \leq F_Y^{-1}(q)).$$

It is the probability that X and Y are both below their p-quantile and q-quantile respectively. One can show that together with the laws of X and Y, the copula characterizes the joint law of

(X, Y) [Sklar, 1959].

Following [Chicheportiche, 2014b] we focus on a subset of the copulas, namely the diagonal copulas $p = q$ and anti-diagonal copulas $p = 1 - q$. The copulas depend on the correlation of the pair and we average the diagonal and anti-diagonal copulas on all the pairs with a given correlation coefficient ρ

$$C(p,p)(\rho) = \Big\langle C_{(x_c,x_{c'})}(p,p) \Big\rangle_{c\neq c',\mathrm{corr}(x_c,x_{c'})=\rho} \quad , \quad C(p,1-p)(\rho) = \Big\langle C_{(x_c,x_{c'})}(p,1-p) \Big\rangle_{c\neq c',\mathrm{corr}(x_c,x_{c'})=\rho}$$

where corr is the Pearson correlation coefficient. On a finite number of stocks, the copulas are averaged over pairs whose correlation is in a certain bin around ρ.

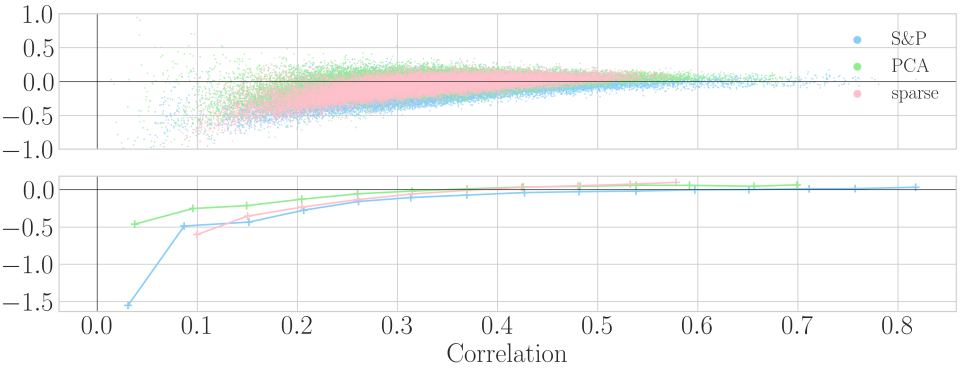


FIGURE 4.7 – (Top) $\ln|\rho/\cos(2\pi C(\frac{1}{2},\frac{1}{2}))|$ versus ρ for each stock pair. (Bottom) average of the point cloud over bins of correlations. The model based on sparse directions better captures the medial copulas $C(\frac{1}{2},\frac{1}{2})$.

We consider the medial point $C(\frac{1}{2},\frac{1}{2})(\rho)$ which is the probability that two stocks are simultaneously below their median. In [Chicheportiche, 2014b] authors have shown that the relation $\ln|\rho/\cos(2\pi C(\frac{1}{2},\frac{1}{2})(\rho))| = 0$ holds for a variety of models in the literature that are called elliptical models. Fig. 4.7 shows the quantities $\ln|\rho/\cos(2\pi C(\frac{1}{2},\frac{1}{2})(\rho))|$ versus ρ for all the pairs in the S&P realization and in our factor models. It shows that the factor model based on sparse directions provide a slightly better description of the medial copulas than the model based on PCA directions.

Authors in [Chicheportiche, 2014b] propose to evidence further non-Gaussian properties by considering the normalized copulas which subtract copulas C_G of a Gaussian pair

$$\Delta(p,p)(\rho) = \frac{C(p,p)(\rho) - C_G(p,p)(\rho)}{p(1-p)} \quad , \quad \Delta(p,1-p)(\rho) = \frac{C(p,1-p)(\rho) - C_G(p,1-p)(\rho)}{p(1-p)}.$$

It is proved in [Chicheportiche, 2014b] that these quantities have a constant limit when $p \to 0$ or $p \to 1$.

They are shown on Fig. 4.8. These curves are significantly non-zero, proving that such non-linear dependencies are present in our factor models. The model based on sparse directions partially captures the amplitude of the curve and the evolution of its concavity along different correlation bins, better that the model based on PCA directions, especially for correlations

$\rho > 0.35$.

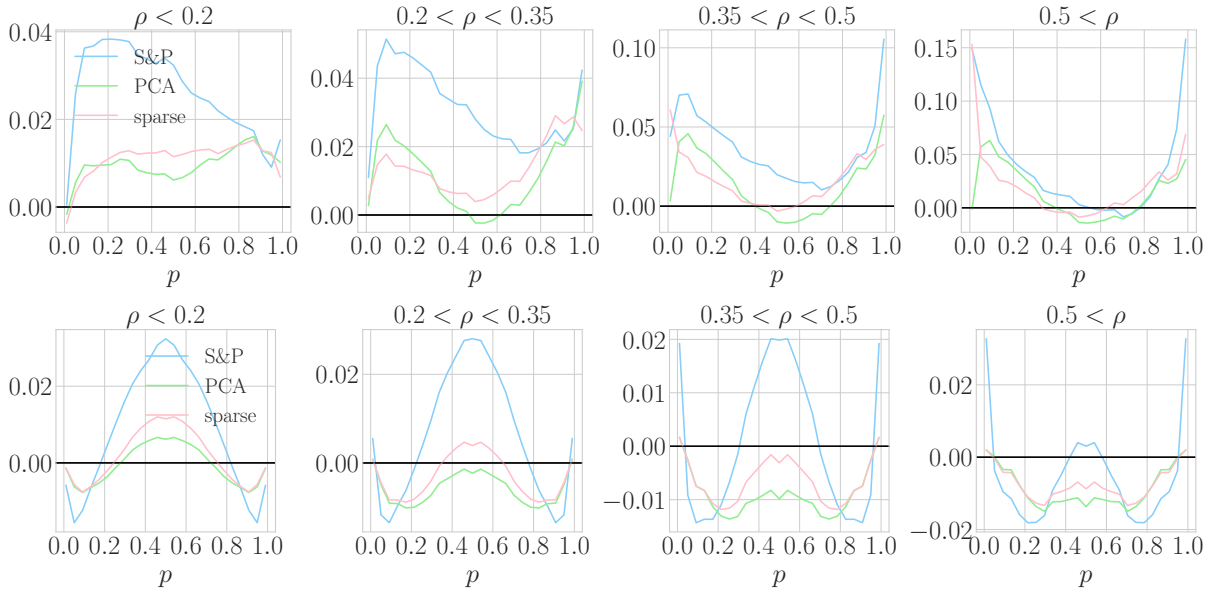


FIGURE 4.8 – Diagonal copulas $C(p,p)(\rho)$ (top) and anti-diagonal copulas $C(p,1-p)(\rho)$ of stock pairs averaged over four bins of correlations ρ. The concavity of the curve is better reproduced in a model based on sparse directions.

4.4.2 Random directions : a few directions to rule them all ?

In our maximum entropy factor model, the time structure along selected few factors is constrained via the Scattering Spectra Φ. We wish to assess to which extent the time structure of other factors is reproduced.

We consider the following score

$$\mathbb{E}\{\|\Phi(\langle w, x\rangle) - \Phi(\langle w, \widetilde{x}\rangle)\|^2\} \tag{4.12}$$

where w is a random direction independent on x. This score measures how good is the model along random directions, up to our Scattering Spectra Φ. The Scattering Spectra $\Phi(\langle w, x\rangle)$ and $\Phi(\langle w, \widetilde{x}\rangle)$ are normalized by the wavelet power spectrum of $\langle w, x\rangle$ and $\langle w, \widetilde{x}\rangle$ respectively (see chapter 2), so that the error $\|\Phi(\langle w, x\rangle) - \Phi(\langle w, \widetilde{x}\rangle)\|^2$ does not depend on the energy of x along the direction $w \in \mathbb{R}^C$.

We consider two types of random directions w. The first ones are uniform on the sphere $w \sim \mathcal{U}(\mathbb{S}^{C-1})$. We also wish to consider directions w that contain a few non-zero coordinates. Such directions are called "baskets" in Finance and are of particular importance for financial agents who often trade the same restricted number of stocks. For that we choose $w \sim \mathcal{U}(\mathbb{S}^{C-1} \cap \{w \mid \|w\|_0 \leq 5\})$ where $\|w\|_0$ counts the number of non-zero coordinates. Table 4.1 shows this score for the benchmark model presented in section 4.2.3, and the two factor models, based on PCA directions or on sparse directions. It is estimated on 10 000 random directions w. It shows that both models reproduce the stochastic structure of random factors much better than the

benchmark model. Both factor models performs very closely.

	random directions	random baskets
benchmark	1.4	1.35
PCA directions	0.22	0.16
sparse directions	0.23	0.17

TABLE 4.1 – Error on the Scattering Spectra of random factors $\langle w, x \rangle$ along uniformly random directions or random baskets of less than 5 stocks. Both factor models perform equally, much better than the benchmark.

We thus conclude that the factor model based on sparse directions outperforms the model based on PCA directions.

4.5 Conclusion

We introduced a maximum entropy factor model of multi-channel multi-scale time-processes that can be estimated on limited data, in the specific case of different stocks of an index. We made use of the Scattering Spectra introduced in chapter 2 in order to constrain the time structure of the projection of the process x along certain directions.

With a sparsity criterion we identified directions, different than the well-studied PCA directions, that better characterize the joint stochastic structure of the process. A factor model based on these sparse directions captures important non-linear dependencies across channels, such as joint time-asymmetry observed on an order 3 moment and copulas statistics.

As a matter of fact, our model relies on an implicit regularity assumption on the process which states that the joint stochastic structure is driven by a few number of factors. Prospective works include quantifying more accurately this regularity. The Scattering Spectra of a process x, introduced in chapter 2 through correlation, can be extended to Scattering cross-Spectra by considering the same correlations across two processes x, y. We believe that a good start in understanding the regularity across channels is to consider the matrix of scattering cross-spectra $\left(\Phi(x_c, x_{c'})\right)_{c,c'}$. This very large matrix characterizes scale dependencies across all pairs of stocks with $M_1 C^2$ coefficients, much more than the data points in a single realization $\widetilde{x}$. We thus expect it to have a low-dimensional structure as a consequence of the regularity across channels, that is yet to be discovered.

Chapitre 5

Unearthing InSights into Mars : Unsupervised Source Separation with Limited Data

Foreword

The previous three chapters involved constructing models of multi-scale processes. We now transition to the realm of inverse problems, which frequently arise in Physics. Source separation involves the ill-posed problem of retrieving a set of source signals that have been observed through a mixing operator. Solving this problem requires prior knowledge, which is commonly incorporated by imposing regularity conditions on the source signals, or implicitly learned through supervised or unsupervised methods from existing data. While data-driven methods have shown great promise in source separation, they often require large amounts of data, which rarely exists in planetary space missions. To address this challenge, we propose an unsupervised source separation scheme for domains with limited data access that involves solving an optimization problem in the wavelet Scattering Spectra space introduced in chapter 2. We present a real-data example in which we remove transient, thermally-induced microtilts—known as glitches—from data recorded by a seismometer during NASA's InSight mission on Mars. Thanks to the wavelet Scattering Spectra ability to capture non-Gaussian properties of stochastic processes, we are able to separate glitches using only a few glitch-free data snippets.
This chapter is adapted from the following publication. Ali Siahkoohi, Rudy Morel, Maarten de Hoop, Erwan Allys, Grégory Sainton, Taichi Kawamura. Unearthing InSights into Mars : Unsupervised Source Separation with Limited Data. *International Conference on Machine Learning*, 2023.

Contents

5.1 Introduction

Source separation is a problem of fundamental importance in the field of signal processing, with a wide range of applications in various domains such as telecommunications [Chevreuil, 2014 ; Gay, 2012 ; Khosravy, 2020], speech processing [Pedersen, 2008 ; Chua, 2016 ; Grais, 2014], biomedical signal processing [Adali, 2015 ; Barriga, 2003 ; Hasan, 2018] and geophysical data processing [Ibrahim, 2014 ; Kumar, 2015 ; Scholz, 2020]. Source separation arises when multiple source signals of interest are combined through a mixing operator. The goal is to determine the original sources with minimal prior knowledge of the mixing process or the source signals themselves. This makes source separation a challenging problem, as the number of sources is usually unknown, and the sources are often non-Gaussian, nonstationary, and multi-scale.

Classical signal-processing based source separation methods [Cardoso, 1989 ; Jutten, 1991 ; Bingham, 2000 ; Nandi, 1996 ; Cardoso, 1998 ; Starck, 2004 ; Jutten, 2004 ; Bobin, 2007] while being extensively studied and well understood, often make simplifying assumptions regarding the sources, e.g., sources being distributed according to Gaussian or Laplace distributions, which might negatively bias the outcome of source separation [Cardoso, 1998 ; Parra, 2003]. To partially address the shortcomings of classical approaches, deep learning methods have been proposed as an alternative approach for source separation, which exploit the information in existing datasets to learn prior information about the sources. In particular, supervised learning methods [Jang, 2003 ; Hershey, 2016 ; Ke, 2020 ; Kameoka, 2019 ; Wang, 2018] commonly rely on existence of labeled training data and perform source separation using an end-to-end training scheme. However, since they require access to ground truth source signals for training, supervised methods are limited to domains in which labeled training data is available.

On the other hand, unsupervised source separation methods [Févotte, 2009 ; Drude, 2019 ; Wisdom, 2020 ; Liu, 2022 ; Denton, 2022 ; Neri, 2021] do not rely on the existence of labeled training data and instead attempt to infer the sources based on the properties of the observed signals. These methods make minimal assumptions about the underlying sources, which make them a suitable choice for realistic source separation problems. Despite their success, unsupervised source separation methods often require tremendous amount of data during training [Wisdom, 2020], which is often infeasible in certain applications such as problem arising in planetary space missions, e.g., due to challenges associated with data acquisition. Moreover, generalization concerns preclude the use of data-driven methods trained on synthetic data in real-world applications due to the discrepancies between synthetic and real data.

To address these challenges, we propose an unsupervised source separation method applicable to domains with limited access to data. In order to achieve this, we leverage a multi-scale prior on the sources through the use of the Scattering Spectra introduced in chapter 2. They capture non-Gaussian multi-scale characteristics of the sources. We perform source separation by solving an optimization problem over the unknown sources that entails minimizing multiple carefully selected and normalized loss functions in the wavelet Scattering Spectra representation space. These loss function are designed to : (1) ensure data-fidelity, i.e., enforce the recovered sources to explain the observed (mixed) data ; (2) incorporate prior knowledge in the form of limited

(e.g., ≈ 50) training examples from one of the sources ; and (3) impose a notion of statistical independence between the recovered sources. Our proposed method does not require any labeled training data, and can effectively separate sources even in scenarios where access to data is limited.

As a motivating example, we apply our approach to data recorded by a seismometer on Mars during NASA's Interior Exploration using Seismic Investigations, Geodesy and Heat Transport (InSight) mission [Giardini, 2020 ; Golombek, 2020 ; Knapmeyer-Endrun, 2020]. The InSight lander's seismometer—known as the SEIS instrument—detected marsquakes [Horleston, 2022 ; Ceylan, 2022 ; Panning, 2023 ; InSight Marsquake Service, 2023] and transient atmospheric signals, such as wind and temperature changes, that provide information about the Martian atmosphere [Stott, 2022] and enable studying the interior structure and composition of the Red Planet [Beghein, 2022]. The signal recorded by the InSight seismometer is heavily influenced by atmospheric activity and surface temperature [Lognonné, 2020 ; Lorenz, 2021], resulting in a distinct daily pattern. Among different types of noise, transient thermally induced microtilts, commonly referred to as glitches [Scholz, 2020 ; Barkaoui, 2021], are a significant component of the noise and one of the most frequent recorded events. These glitches, hinder the downstream analysis of the data if left uncorrected [Scholz, 2020]. We show that our method is capable of removing glitches from the recorded data by only using a few snippets of glitch-free data.

In the following sections, after describing the related work, as a means to perform source separation in domains with limited data, we introduce our source separation approach that involves solving an optimization problem with loss functions defined in the wavelet Scattering Spectra space. We present two numerical experiments : (1) a synthetic setup in which we can quantify the accuracy of our method ; and (2) examples involving seismic data recorded during the NASA InSight mission.

5.2 Related work

Regaldo-Saint Blancard et al. [Regaldo-Saint Blancard, 2021] introduced the notion of components separation through a gradient descent in signal space with indirect constraints with applications to to the separation of an astrophysical emission (polarized dust emission in microwave) and instrumental noise. In an extensive study, Delouis, J.-M. et al. [Delouis J-M, 2022] attempts to separate the full sky observation of the dust emission with instrumental noise using similar techniques via wavelet Scattering Spectra representations. Authors take the non-stationarity of the signal into account by constraining statistics on several sky masks. Contrarily to a usual denoising approach, both of these works focus primarily on recovering the statistics of the signal of interest. In a related approach, Jeffrey et al. [Jeffrey, 2022] use a scattering transform generative model to perform source separation in a Bayesian framework. While very efficient, this approach requires training samples from each component, which are often not available. Finally, Xu et al. [Xu, 2022] similarly aim to remove glitches and they develop a supervised learning based on deglitched data obtained by existing glitch removal tools. As a result, the accuracy of their result is limited to the accuracy of the underlying data processing

tool, which our method avoid by being unsupervised. As we show in our examples, we are able to detect and remove glitches that were undetected by the main deglitching software [Scholz, 2020] developed closely by the InSight team.

5.3 Problem setup

Consider a linear mixing of unknown sources $s_i(t),\ i = 1, \dots, N$ via a mixing operator A,

$$x(t) = As(t) + \nu(t) = a_1^T s_1(t) + n(t), \tag{5.1}$$

with

$$\begin{aligned} s(t) &= (s_1(t), \dots, s_N(t))^T,\ A = \left[a_1^T\ \cdots\ a_N^T\right], \\ n(t) &= \nu(t) + \sum_{i=2}^{N} a_i^T s_i(t). \end{aligned} \tag{5.2}$$

In the above expressions, x represents the observed data, and ν is the measurement noise. Here we capture the noise and the mixture of all the sources except for s_1 through the mixing operator in n that does not longer depends on s_1. The matrices x and s have dimensions of $C \times T$ and $N \times T$, respectively, where T represents the number of time samples. The mixing operator A has dimensions of $C \times N$. The product of a_1^T and $s_1(t)$ yields a vector of size C and we note $a_1^T s_1$ the $C \times T$ resulting matrix, which corresponds to the contribution of source s_1 exclusively in x.

Objective. The aim is to obtain a point estimate s_1 given a single observation x with the assumption that a_1 is known and that we have access to a few realizations $\widetilde{n}_1, \dots, \widetilde{n}_K$ as a training dataset. For example, in the case of separating glitches from seismic data recorded during the NASA InSight mission, we will consider $\widetilde{n_k}$ to be snippets of glitch-free data and a_1 to encodes information regarding polarization. We will drop the time dependence of the quantities in equations (5.1) and (5.2) for convenience.

5.4 Principle of the method

The inverse problem of estimating s_1 from the given observed data x, as presented in equation (5.1), is ill-posed since the solution is not unique. To constrain the solution space of the problem, we incorporate prior knowledge in the form of realizations $\widetilde{n}_1, \dots \widetilde{n}_K$. We achieve this through a loss function that emphasizes the wavelet Scattering Spectra representation of $x - a_1^T s_1$ to be close to that of $\widetilde{n}_k,\ k = 1, \dots, K$:

$$\mathcal{L}_{\text{prior}}(s_1) := \frac{1}{K} \sum_{k=1}^{K} \left\| \Phi(x - a_1^T s_1) - \Phi(\widetilde{n}_k) \right\|_2^2. \tag{5.3}$$

In the above expression, Φ is the wavelet Scattering Spectra mapping as described in section 2.6.1 of chapter 2. With the prior loss defined, we impose data-consistency via :

$$\mathcal{L}_{\text{data}}(s_1) := \frac{1}{K}\sum_{k=1}^{K}\left\|\Phi(a_1^T s_1 + \widetilde{n}_k) - \Phi(x)\right\|_2^2. \tag{5.4}$$

The data consistency loss function $\mathcal{L}_{\text{data}}$ promotes estimations of s_1 such that for any training example from $\widetilde{n}_1, \ldots, \widetilde{n}_K$ the wavelet Scattering Spectra representation of $a_1^T s_1 + \widetilde{n}_k$ is close to that of the observed data.

To promote the independence of sources we make use of the Scattering cross-Spectra. The Scattering Spectra, introduced in chapter 2, computes a diagonal approximation of a correlation-based representation of the form $\langle \mathcal{R}x\, \mathcal{R}x^* \rangle$ where $\mathcal{R}x = (Wx, |Wx|)$ with W a wavelet operator and $\langle . \rangle$ performs an average over time. The Scattering cross-Spectra, written $\Phi(x, y)$, between two signals x, y, are defined as the same diagonal approximation, but now on the cross-correlation matrix $\langle \mathcal{R}x\, \mathcal{R}y^* \rangle$. It captures scale dependencies across two signals x and y. For the cross-Spectra, we do not take the low-pass wavelet in W. One can prove that if two processes x, y are independent then $\mathbb{E}\{\Phi(x, y)\} = 0$ so that $\Phi(x, y) \approx 0$ up to estimation error.

To promote the independence of sources, we penalize the Scattering cross-Spectra between $a_1^T s_1$ and $\widetilde{n}_k$

$$\mathcal{L}_{\text{cross}}(s_1) := \frac{1}{K}\sum_{k=1}^{K}\left\|\Phi(a_1^T s_1, \widetilde{n}_k)\right\|_2^2. \tag{5.5}$$

5.5 Loss normalization

The losses described previously do not contain any weighting term for the different coefficients of the Scattering Spectra representation. We introduce in this section a generic normalization scheme, based on the estimated variance of certain Scattering Spectra distributions. This normalization, which has been introduced in Delouis, J.-M. et al. [Delouis J-M, 2022], allows to interpret the different loss terms in a standard form, and to include them additively in the total loss term without overall loss weights. Let us consider first the loss term given by equation (5.3), which compares the distance between $x - a_1^T s_1$ and available training samples $\widetilde{n}_1 \ldots, \widetilde{n}_K$ in the wavelet Scattering Spectra representation space. Specifying explicitly the sum on the M wavelet Scattering Spectra coefficients $\Phi_m,\ m = 1, \ldots, M$, it yields

$$\mathcal{L}_{\text{prior}}(s_1) = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\left|\Phi_m(x - a_1^T s_1) - \Phi_m(\widetilde{n}_k)\right|^2.$$

Let us consider the second sum in this expression. In the limit where $\Phi_m(x - a_1^T s_1)$ is drawn from the same distribution as the $(\Phi_m(\widetilde{n}_k), 1 \le k \le K)$, the difference $\Phi_m(x - a_1^T s_1) - \Phi_m(\widetilde{n}_k)$, seen as a random variable, should have zero mean, and the same variance as the $(\Phi_m(\widetilde{n}_k), 1 \le k \le K)$ up to a factor 2. Denoting $\sigma^2(\Phi_m(n_k))$ as this variance, which can be estimated from

$(\Phi_m(\widetilde{n}_k), 1 \leq k \leq K)$, this gives a natural way of normalizing the loss :

$$\mathcal{L}_{\text{prior}}(s_1) = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{\left|\Phi_m(x - a_1^T s_1) - \Phi_m(\widetilde{n}_k)\right|^2}{\sigma^2(\Phi_m(\widetilde{n}_k))}$$

or in a compressed form

$$\mathcal{L}_{\text{prior}}(s_1) = \frac{1}{K} \sum_{k=1}^{K} \frac{\left\|\Phi(x - a_1^T s_1) - \Phi(\widetilde{n}_k)\right\|_2^2}{\sigma^2(\Phi(\widetilde{n}_k))}, \tag{5.6}$$

which takes into account the expected standard deviation of each coefficient of the Scattering Spectra representation. This normalization allows for two things. First, it removes the normalization inherent to the multi-scale structure of Φ. Indeed, coefficients involving low frequency wavelets tend to have a larger norm. Second, it allows to interpret the loss value, which is expected to be at best of order unity and to sum different loss terms of same magnitude.

We can introduce a similar normalization for the other loss terms. Loss term (5.4) should be normalized by the M-dimensional vector $\sigma^2(\Phi(a_1^T s_1 + \widetilde{n}_k))$ that we approximate by $\sigma^2(\Phi(x + \widetilde{n}_k))$, in order to have a normalization independent on s_1, yielding

$$\mathcal{L}_{\text{data}}(s_1) := \frac{1}{K} \sum_{k=1}^{K} \frac{\left\|\Phi(a_1^T s_1 + \widetilde{n}_k) - \Phi(x)\right\|_2^2}{\sigma^2(\Phi(x + \widetilde{n}_k))}. \tag{5.7}$$

Finally, loss term (5.5) should be normalized by $\sigma^2(\Phi(a_1^T s_1, \widetilde{n}_k))$ that we approximate by $\sigma^2(\Phi(x, \widetilde{n}_k))$

$$\mathcal{L}_{\text{cross}}(s_1) = \frac{1}{K} \sum_{k=1}^{K} \frac{\left\|\Phi(a_1^T s_1, \widetilde{n}_k)\right\|_2^2}{\sigma^2(\Phi(x, \widetilde{n}_k))}, \tag{5.8}$$

We can now sum the normalized loss terms defined in equations (5.6)–(5.8) to get the final optimization problem to perform source separation

$$\overline{s}_1 := \arg\min_{s_1} \Big[\mathcal{L}_{\text{data}}(s_1) + \mathcal{L}_{\text{prior}}(s_1) + \mathcal{L}_{\text{cross}}(s_1)\Big]. \tag{5.9}$$

Due to the delicate normalization of the three terms, we expect that further weighting of the three losses using weighting hyperparameters is not necessary. We propose to initialize the optimization problem in equation (5.9) with $s_1 := 0$. Such choice means that $n = x - a_1^T s_1$ is initialized to x, which contains crucial information on the sources, as will be explained in the next section.

We have observed that as soon as we know the statistics $\Phi(n)$, our algorithm retrieves the unknown statistics of the source $\Phi(a_1^T s_1)$. In other words the algorithm successfully separates the sources in the Scattering Spectra space, this constitutes a convergence result, that can be proved under simplifying assumptions (see theorem 3). Of course, in many cases as we will see in the next section, our algorithm retrieves point estimates of s_1 that is stronger.

Theorem 3. *Let $x = a_1^T s_1 + n$ with s_1 and n two independent processes. Let us assume we have two processes $\overline{s}_1$ and $\overline{n}$ with $x = a_1^T \overline{s}_1 + \overline{n}$.*

Under the following assumptions :

1. *n has a maximum entropy distribution under moment constraints $\mathbb{E}\{\Phi(n)\}$*
2. *$\overline{n}$ has a maximum entropy distribution under moment constraints $\mathbb{E}\{\Phi(\overline{n})\}$*
3. *$\mathbb{E}\{\Phi(\overline{n})\} = \mathbb{E}\{\Phi(n)\}$*
4. *$\overline{s}_1$ and $\overline{n}$ are independent*
5. *The Fourier transform $\widehat{p}_n$ of the distribution p_n of n is non-zero everywhere.*

one has $n \overset{d}{=} \overline{n}$ and $a_1^T s_1 \overset{d}{=} a_1^T \overline{s}_1$ where the equality is on the distribution of the processes.

Essentially, it means that when the source n is statistically characterized by its Scattering Spectra descriptors, the algorithm is able to retrieve statistically the other source. The theorem is proved and its assumptions are discussed in appendix C.1. This emphasizes the choice of a representation Φ that characterizes efficiently the stochastic structure of multi-scale processes, which was the subject of chapter 2.

5.6 Numerical experiments

The main goal of this chapter is to derive a unsupervised approach to source separation that is applicable in domain with limited access to training data, thanks to the wavelet Scattering Spectra representation. To provide a quantitative analysis to the performance of our approach, we first consider a stylized synthetic example that resembles challenges of real-world data. To illustrate how our method performs in the wild, we apply our method to data recorded on Mars during the InSight mission. We aim to separate transient thermally induced microtilts, i.e., glitches [Scholz, 2020 ; Barkaoui, 2021], from the recorded data by the InSight lander's seismometer. The code for partially reproducing the results can be found on GitHub. Our implementation is based on the original PyTorch code for wavelet Scattering Spectra.

5.6.1 Stylized example

We consider the problem of separating glitch-like signals from increments of a multifractal random work process [Bacry, 2001a]. This process is a typical non-Gaussian noise exhibiting long-range dependencies and showing bursts of activity, e.g., see Figure C.1 in the appendix for several realizations of this process. The second source signal is composed of several peaks with exponentially decaying amplitude, with possibly different decay parameters on the left than on the right. To obtain synthetic observed data, we sum increments of a multifractal random walk realization, which plays the role of n in equation (5.1), with a realization of the second source. The top three images in Figure 5.1 are the signal of interest, secondary added signal, and the observed data, respectively.

In order to retrieve the multifractal random walk realization, we solve the optimization problem in equation (5.9) using the L-BFGS optimization algorithm [Liu, 1989] using 500 iterations.

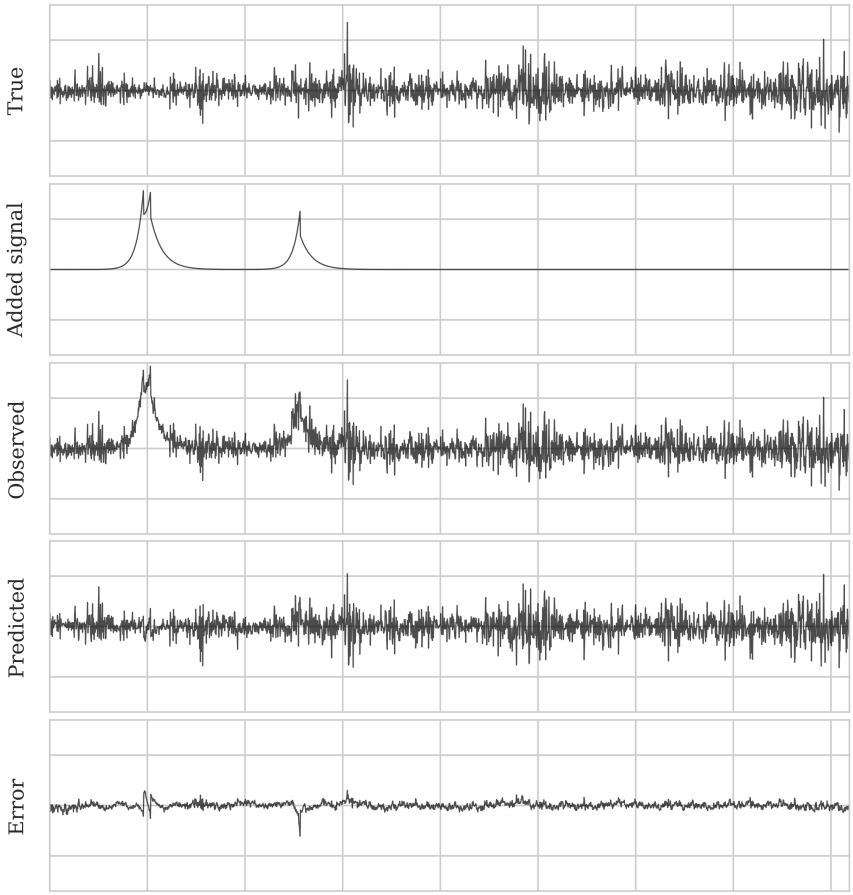


FIGURE 5.1 – Unsupervised source separation applied to the multifractal random walk data. The vertical axis is the same for all the plots.

We use a training dataset of 100 realizations of increments of a multifractal random walk, $\widetilde{n}_k$. We compute Scattering Spectra with $J = 8$ octaves. Given an input signal dimension of $T = 2048$, this choice of parameters yields a 174-dimensional wavelet Scattering Spectra space. The bottom two images in Figure 5.1 summarizes the results. We are able to recover the ground-truth multifractal random walk realization up to small, mostly incoherent, and seemingly random error. To see the effect of number of training realizations on the signal recovery, we repeated the above examples and used varying number of training samples. Figure 5.3 shows that, as expected, the signal-to-noise ratio of the recovered sources increases the more training samples we have.

We also investigate the behavior of our source separation algorithm in case there are no additional sources present in the signal, i.e., the observed data is a realization of the same stochastic process as the data snippets $\widetilde{n}_1, \ldots, \widetilde{n}_K$. Ideally, the source separation algorithm should not unnecessarily remove important signals. We present the results of this experiment in Figure 5.2, which indeed confirms that only a negligible amount of energy has been removed from the observed data in this case. We argue that the undesired separated signal from the observed data by our method is mainly due to errors in estimating the Scattering Spectra statistics using a finite amount of data snippets.

To show our method can also separate sources that are not localized in time, we consider

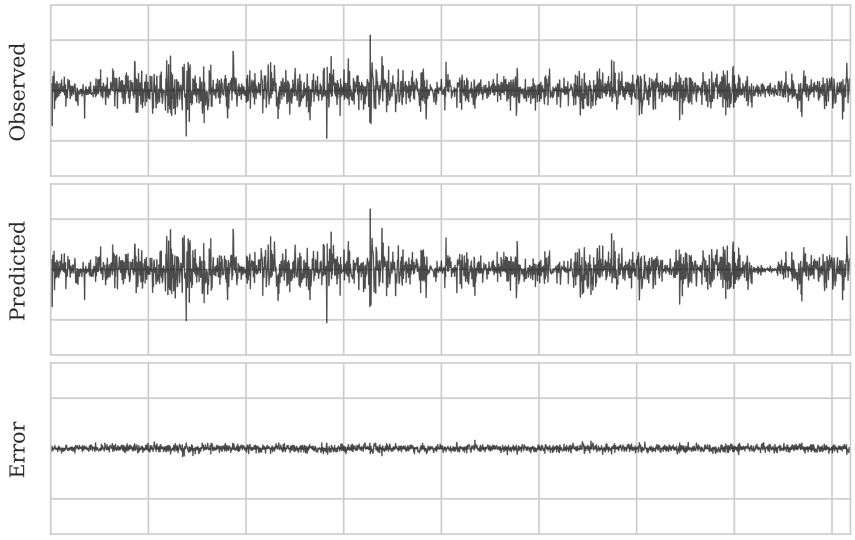


FIGURE 5.2 – The behavior when there are no sources to be removed, i.e., the observed data is a realization of the same stochastic process as the data snippets. The vertical axis is the same for all the plots.

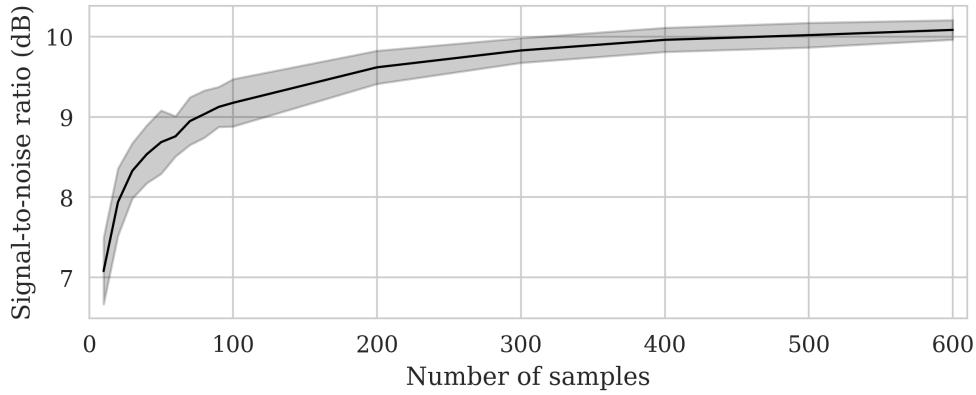


FIGURE 5.3 – Signal-to-noise ratio of the predicted multifractal random walk data versus number of unsupervised samples. Shaded area indicates the 90% interval of this quantity for ten random source separation instances.

contaminating the multifractal random walk data with a turbulent signal (see second image from the top in Figure 5.4. Without any prior knowledge regarding this turbulent signal and by only using 100 realizations of increments of a multifractal random walk as training samples, we are able to recover the signal of interest with arguably low error : juxtapose the ground truth and predicted multifractal random walk realization in Figure 5.4. The algorithm correctly removes the low frequencies content of the turbulent jet, and makes a small, uncorrelated, random error at high frequencies. In this case the two signals having different power spectra helps disentangling them at high frequencies. In the above synthetic examples, the signal low frequencies are well separated and the algorithm infers correctly the high frequencies. In the earlier example, the presence of time localized sources would facilitate the algorithm to "interpolate" the background noise knowing its Scattering Spectra representation. This case makes it more evident that the initialization $s_1 = 0$ informs the algorithm of the trajectory of the unknown source.

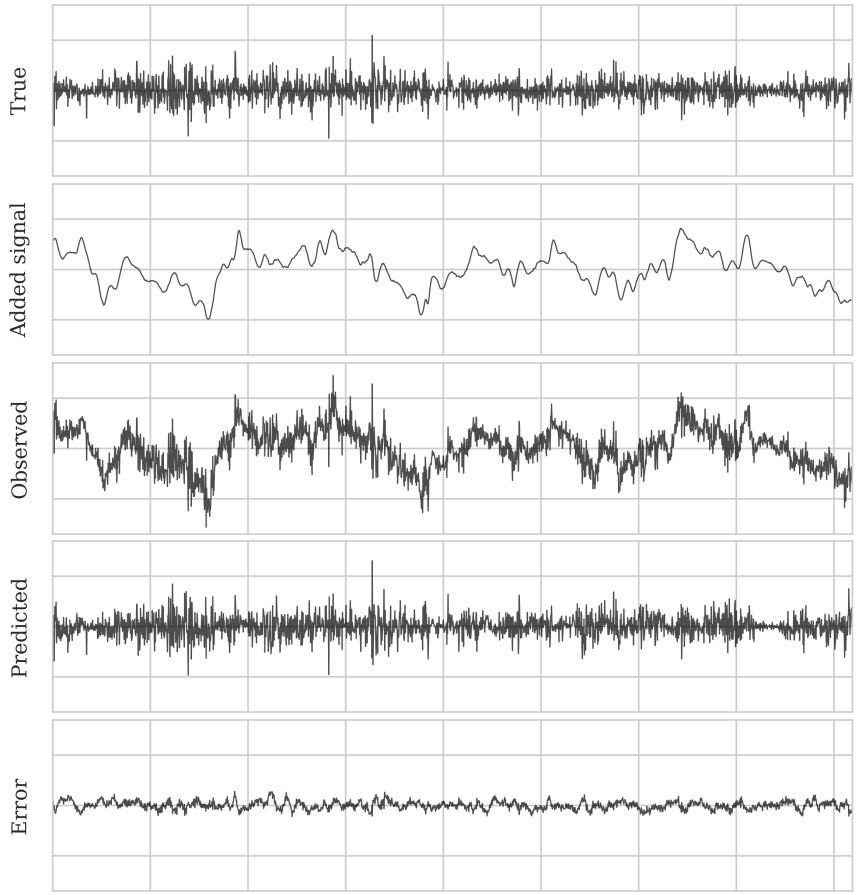


FIGURE 5.4 – Unsupervised source separation applied to the multifractal random walk data with a turbulent additive signal. The vertical axis is the same for all the plots.

5.6.2 Application to data from the InSight mission

InSight lander's seismometer, SEIS, is exposed to heavy wind and temperature fluctuations. As a result, it is subject to background noise. Glitches are a widely occurring family of noise caused by a variety of causes [Scholz, 2020]. These glitches often appear as one-sided pulses in seismic data and significantly affect the analysis of the data [Scholz, 2020]. In this section we will explore the application of our proposed method in separating glitches and background noise from the recorded seismic data on Mars.

5.6.2.1 Separating glitches

We propose to consider glitches as the source of interest s_1 in the context of equation (5.1). To perform source separation using our technique, we need snippets of data that do not contain glitches. We select these windows of data using an existing catalog and glitches [Scholz, 2020] and by further eye examination to ensure no glitch contaminates our dataset. In total, we collect 50 windows of length 102.4 s during sol 187 (6 June 2019) for the U component. We show four of these windows of data in Figure 5.5. We perform optimization for glitch removal using the same underlying scattering network architecture as the previous example using 50 training samples

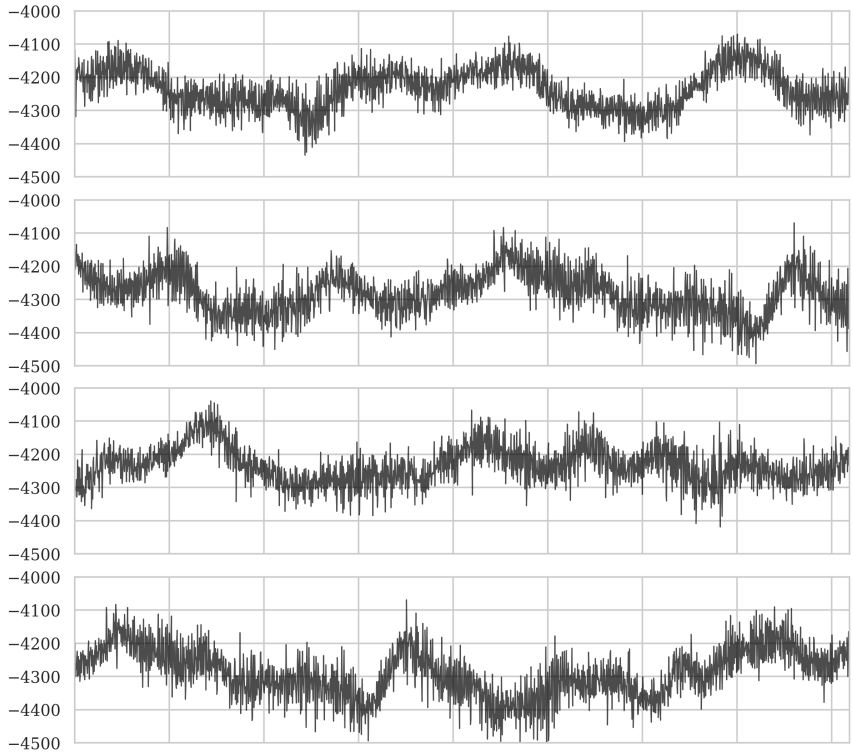


FIGURE 5.5 – Glitch-free snippets of the seismic data from Mars (U component).

and 1000 L-BFGS iterations. Figure 5.6 summarizes the results. The top-left image shows the raw data. Top-right image is the baseline [Scholz, 2020] (see Appendix C.2 for description) prediction for the glitch signal. Finally, the bottom row (from left to right) shows our predicted deglitched data and the glitch signal separated by our approach. As confirmed by experts at the InSight team, indeed our approach has removed a glitch that the baseline has ignored (most likely due the spike right at the beginning of the glitch signal). More deglitching examples can be seen in Figures C.2–C.5.

It is important to note that the separated glitch in our experiments may comprise some non-transient, non-seismic signals, potentially arising from atmosphere-surface interactions, as opposed to the the baseline glitch. Consequently, we anticipate the separation of these non-seismic signals in addition to the glitch when applying our approach. This results in "noisy" predicted glitches when compared to the baseline, which might be due to the the non-seismic signal. With this in mind, our approach extends the notion of glitch (as understood by the InSight team). This is one of the benefits of our unsupervised approach as the method—based on the statistics of the training data—identifies and removes events that do not seem to belong to the training data distribution.

Thanks to the interpretability of wavelet Scattering Spectra representations, stemming from our comprehension of scattering coefficients and covariances, we can perform a source separation quality control in domain where there is no access to ground truth source—as in our example. Figure 5.7 compares the power spectra of the reconstructed background noise (recorded data), a deglitched realization of the background noise and the mixed signal (observed data). It can

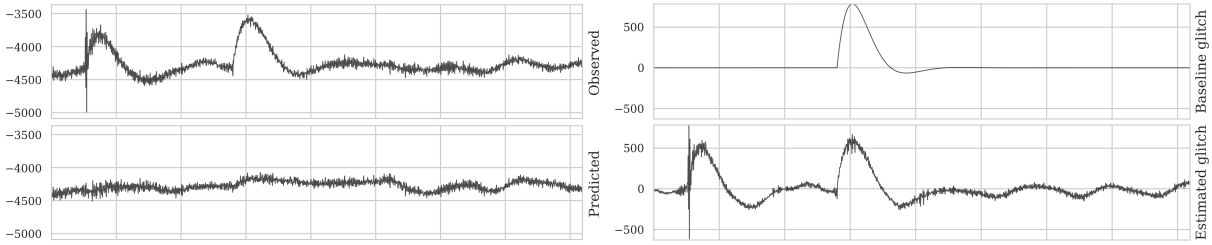


FIGURE 5.6 – Unsupervised source separation for glitch removal. Juxtapose the predicted glitches on the right. Our approach is able to remove a glitch whereas the baseline approach fails to detect it.

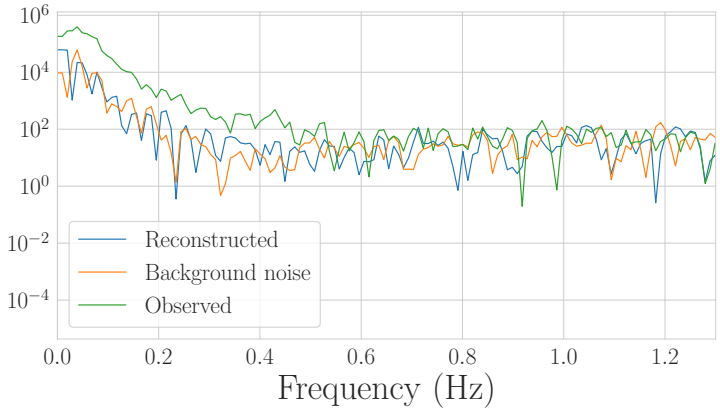


FIGURE 5.7 – Power spectrum of the observed signal x, the background noise n and the reconstructed background noise $x - a_1^T \bar{s}_1$. We see that the reconstructed component statistically agrees with a Mars seismic background noise n. The algorithm efficiently removed the low-pass component of the signal corresponding to a glitch.

be seen that the power spectrum of the background noise is correctly retrieved. In fact, the Scattering Spectra statistics, which extend the power spectrum, are correctly retrieved, which is due to the loss term in equation (5.3).

5.6.2.2 Marsquake background noise separation

Marsquakes are of significant importance as they provide useful information regarding the Mars subsurface, enabling the study of Mars' interior [Knapmeyer-Endrun, 2021 ; Stähler, 2021 ; Khan, 2021]. Recordings by the InSight lander's seismometer are susceptible to background noise and transient atmospheric signals, and here we apply our proposed unsupervised source separation approach to separate background noise from a marsquake [InSight Marsquake Service, 2023]. To achieve this, we select about 30 hours of raw data (except for a detrending step)—from the U component with a 20Hz sampling rate—to fully characterize various aspects of the background noise through the wavelet Scattering Spectra representation. Next, we window the data and use the windows as training samples from background noise (n_k in the context of equation (5.1)) with the goal of retrieving the marsquake recorded at February 3, 2022 [InSight Marsquake Service, 2023].

We use the same network architecture as previous examples to setup the wavelet Scattering

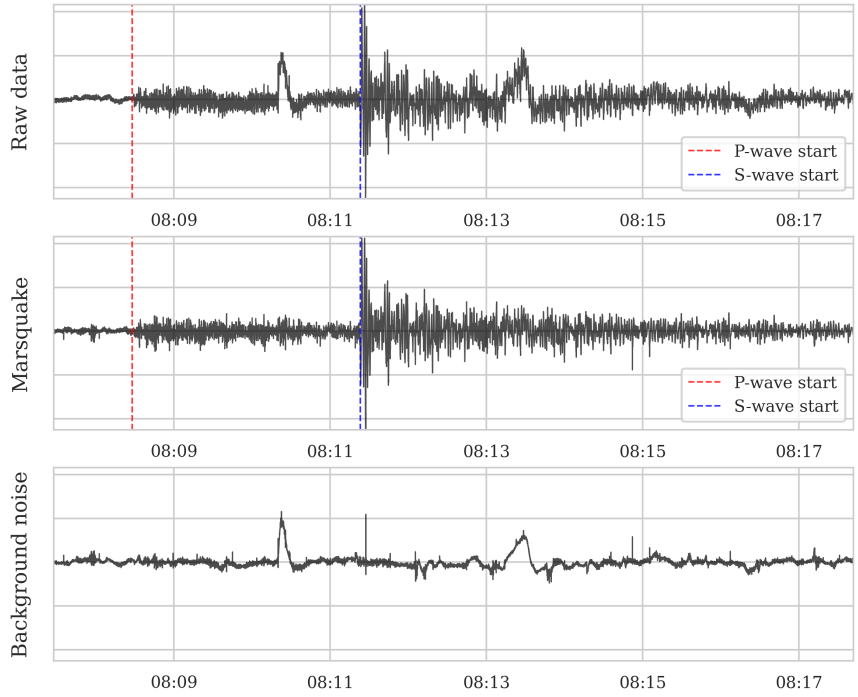


FIGURE 5.8 – Unsupervised separation of background noise, including thermally induced microtilts (glitches), from a marsquake recorded by the InSight lander's seismometer on February 3, 2022 [InSight Marsquake Service, 2023]. Approximately 30 hours of raw data from the U component, with no recorded marsquakes, were utilized for background noise separation without any explicit prior knowledge of marsquakes or glitches. The horizontal axis represents the UTC time zone.

Spectra representation. We use a window size of 204.8 s and solve the optimization problem in equation (5.9) with 200 L-BFGS iterations. The results are depicted in Figure 5.8. There are clearly two glitches that we have successfully separated, along with the background noise. This results is obtained merely by using 30 hours of raw data, allowing us to identify the marsquake as a separate source due to differences in wavelet Scattering Spectra representation.

5.7 Conclusion

For source separation to be effective, prior knowledge concerning unknown sources is necessary. Data-driven source separation methods extract this information from existing datasets during pretraining. In most cases, these methods require a large amount of data, which means that they are not suitable for planetary science missions. To address the challenge posed by limited data, we proposed an approach based on Scattering Spectra introduced in chapter 2. Using a Scattering Spectra space optimization problem, we were able to separate thermally induced microtilts (glitches) from data recorded by the InSight lander's seismometer with only a few glitch-free data samples. In addition, we applied the same strategy to separate marsquakes from background noise and glitches using only several hours of data with no recorded marsquake. Our approach did not require any knowledge regarding glitches or marsquakes, and proved to be

more robust in separating glitches from recorded seismic data on Mars than existing techniques. An important characteristic of our approach is that it serves as an exploratory method for unsupervised learning, particularly beneficial for investigating complex and real-world datasets.

Chapitre 6

Path Shadowing Monte-Carlo

Foreword

This chapter considers prediction on multi-scale time-series from limited data. For financial time-series, this presents a double challenge. Firstly, accurately learning the association between past and future is difficult due to the limited historical data available Additionally, price paths are very noisy i.e. the price process has a high entropy and any two snippets of data are very distant. Given a current date with an observed past time-series, it is thus difficult to look for similar occurrences in the past, that could offer insights into potential future scenarios. In order to face these challenges we introduce a prediction framework called *Path Shadowing Monte-Carlo*. It provides prediction of future paths given any generative model. At any given date, it averages future quantities over generated price paths whose past history "shadows" the actual (observed) history. We test our approach using paths generated from a maximum entropy model of financial prices based on the Scattering Spectra introduced in chapter 2 that we interpret here as multi-scale extensions of the standard skewness and kurtosis widely embrassed in Finance. This model promotes diversity of generated paths while reproducing the main statistical properties of financial prices, including stylized facts on volatility roughness. Our method yields state-of-the-art predictions for future realized volatility and allows one to determine conditional option smiles for the S&P500 that outperform both the Path Dependent Volatility model and the option market itself.

This chapter is adapted from the following submitted paper. Rudy Morel, Stéphane Mallat, Jean-Philippe Bouchaud. Path Shadowing Monte-Carlo, 2023.

Contents

6.1 Introduction

Modelling future price scenarios is crucial for risk control, for pricing and hedging contingent claims (like options), and, possibly, for detecting arbitrage opportunities. Recently, machine learning models such as transformers [Vaswani, 2017 ; Wen, 2022] propose to learn from data the distribution $p(x|x_{\text{past}})$ of log-prices x conditioned on past history x_{past}. When trained with a prediction loss, such models generally achieve excellent prediction results. However, their training requires very large amount of data which is usually not available for financial prices.

On the other hand, low-parameterized generative models, i.e. models p_θ of $p(x)$ with few parameters θ, have been extensively studied in the financial literature [Heston, 1993 ; Bacry, 2013 ; Gatheral, 2018 ; Wu, 2022 ; Guyon, 2022]. However, two main challenges come to the fore. First, these models may not reproduce some important statistics of real financial prices due to flawed assumptions, or due to the fact that they are calibrated on external data such as observed option smiles. Second, it may not be straightforward to condition these models on the realized past at a specific date, in other words, obtaining a model of $p(x|x_{\text{past}})$. Whereas conditioning is eased by considering Markovian models with a small number of factors [Guyon, 2022], such a strong assumption is often much too simplistic. In this chapter, we attempt to address both challenges.

Our main contribution is to introduce a new method, that we call *Path Shadowing Monte-Carlo* (PS-MC), which can be used within any generative model of $p(x)$ to yield a model of $p(x|x_{\text{past}})$. Our approach for modelling the distribution $p(x)$, summarized in section 6.2, is to define a minimal set of statistics describing financial prices that should be reproduced by the generating process. This set should be small enough to avoid over-fitting but should focus on "relevant" features, in a sense made precise below. This question was addressed in chapter 2, where it was shown that a good description of multi-scale processes can be achieved through the Scattering Spectra. Here we present such statistics, in a Finance context, as a multi-scale extension of the classical skewness and kurtosis, this is motivated in section 6.2.

A model based on these statistics captures all important stylized facts such as fat-tail distributions, intermittency, leverage effect and the "Zumbach effect", see chapter 2. Section 6.3 characterizes the average shape of option smiles generated by our model and shows that it correctly reproduces non-trivial power-law behaviors as a function of maturity, which were recently argued to be a specific feature of rough volatility models [Gatheral, 2018 ; Fukasawa, 2017].

"Path shadowing" is presented in section 6.4. It consists in softening the conditioning on a given past history x_{past}. In a nutshell, it amounts to scanning a large generated dataset, in search of paths whose history closely "shadows" the actual history, see Fig. 6.3. Path Shadowing Monte-Carlo methods then average the quantity of interest over the future of such matching paths. This method can effectively be seen as a kernel method, with a causal path embedding to reduce the dimensionality of recent past history.

Compared to other recent kernel methods, such as signature kernels [Salvi, 2021 ; Alden, 2022] that relies on a low-parametric model for $p(x|x_{\text{past}})$, e.g. a Heston model, Path Shadowing Monte-Carlo relies solely on a model of $p(x)$ and thus circumvent the conditioning of a gene-

rative model to a given past history x_{past}. Its performance depends directly on the accuracy of this generative model and its ability to produce a variety of paths with correct statistical dependencies. Section 6.4.3 shows that when performed with our maximum entropy Scattering Spectra model of financial prices, PS-MC yields state-of-the-art volatility prediction.

Section 6.5 uses Path Shadowing Monte-Carlo for obtaining *conditional* option smiles (i.e. option prices at a given date) through Hedged Monte-Carlo with shadowing paths. By construction, such smiles depend only on the log-price process distribution $p(x)$ and provide a counterpart to smiles obtained from option market data. A "trading game" then allows us to show that our option smiles correctly anticipate non-trivial future price movements, and outperforms state of the art models such as the Path Dependent Volatility model of ref. [Guyon, 2022]. Codes for both our generative model and Path Shadowing Monte-Carlo are available at `https://github.com/RudyMorel/path_shadowing`.

6.2 A multi-scale statistical model for financial prices

Statistical models of financial prices aim at reproducing statistics of the price process only. Price time-series exhibit numerous non-Gaussian features, which are difficult to capture within standard low-parametric models, whose number of parameters have been incrementally increased in the literature over the past decades, see e.g. [Heston, 1993 ; Bacry, 2013 ; Gatheral, 2018 ; Delemotte, 2023 ; Guyon, 2022]. An alternative route is to define a set of characteristic statistics of (log-)prices and impose that they should be accurately reproduced by the model. We denote as $\Phi(x)$ such statistics, for example the empirical mean and variance of log-returns $\Phi(x) = (\langle \delta x(t) \rangle_t , \langle |\delta x(t)|^2 \rangle_t)$. Section 6.2.1 presents maximum entropy models that allow defining models from a given vector of statistics $\Phi(x)$. In the simple case of mean and variance, the maximum entropy model coincides with the Gaussian random walk.

The set $\Phi(x)$ must be chosen carefully. It should contain enough relevant statistics of prices such that the model is realistic and accurate. However, in order to avoid over-fitting, such statistics must be well estimated on the only available historical realization of x. The construction of a set Φ that meets these requirements, called the *Scattering Spectra*, was introduced in chapter 2 in the general case of multi-scale processes, which fortunately includes financial data [Bacry, 2013 ; Borland, 2005].

We show in section 6.2.2 that such statistics correspond to natural low moment multi-scale extensions of the classical skewness and kurtosis of log-returns. We show that even the most recent low-dimensional parametric models fail to accurately account for these statistics. Such discrepancies turn out to be highly relevant when one wants to predict future realized volatility and option smiles, and highlights the limitations of traditional models, which our approach allows one to overcome.

In chapter 2, we have shown that a Scattering Spectra model properly captures the main properties of financial log-returns, in particular of the S&P500 (the US major stock index). In the following, we show that it also quantitatively reproduces the average behavior of option smiles of different maturities, in particular the maturity-dependent skewness that reflects volatility

roughness [Fukasawa, 2017] and the so-called skew-stickiness ratio [Bergomi, 2009 ; Vargas, 2015].

6.2.1 Maximum entropy models

We denote as $\widetilde{x} \in \mathbb{R}^N$ the observed historical realization of log-prices over N days [1]. Given a vector of M statistics $\Phi(\widetilde{x}) \in \mathbb{R}^M$ estimated on $\widetilde{x}$, a maximum entropy model p_θ with moment constraint $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \Phi(\widetilde{x})$, if it exists, has an exponential probability distribution

$$p_\theta(x) = Z_\theta^{-1} e^{-\langle \theta, \Phi(x) \rangle}. \tag{6.1}$$

for certain $\theta \in \mathbb{R}^M$.

Maximum entropy models depend only on the vector of statistics $\Phi(x)$. The model bias can be improved by enriching the set $\Phi(x)$. However, we must take into account the problem of estimating $\Phi(x)$ from the single realization of the process $\widetilde{x}$. The Scattering Spectra model imposes $\mathbb{E}_{p_\theta}\{\Phi(x)\} = \Phi(\widetilde{x})$, thus for p_θ to be a good approximation of the true distribution p, one needs $\Phi(\widetilde{x})$ to be close to the true $\mathbb{E}_p\{\Phi(x)\}$. This amounts to having low-variance statistics Φ. In the next section we present a good choice of Φ, the Scattering Spectra, introduced in chapter 2, that is we interpret in the context of Finance.

A microcanonical maximum entropy model has a maximum entropy distribution on the set

$$\Omega_\epsilon = \{x \in \mathbb{R}^N \mid \|\Phi(x) - \Phi(\widetilde{x})\|_2 < \epsilon\}.$$

Drawing samples from such model is performed through an approximate algorithm based on gradient descent, see Appendix A.6 for more details.

6.2.2 The Scattering Spectra

A standard way of characterizing the price process is through their trend, volatility, skewness and kurtosis. These are obtained from moments of order 1, 2, 3 and 4 on log-returns

$$\mathbb{E}\{\delta x(t)\} \ , \ \mathbb{E}\{\delta x(t)^2\} \ , \ \mathbb{E}\{\delta x(t)^3\} \ , \ \mathbb{E}\{\delta x(t)^4\} \tag{6.2}$$

However such moments do not characterize the time-structure of log-returns, but rather their one-point distribution. One could consider the same moments on multi-scale increments

$$\delta_\ell x(t) = x(t) - x(t - \ell) \tag{6.3}$$

for different lags ℓ, but we still obtain a poor description of x. For example, these moments do not pick up time-asymmetry, since changing $\delta x(t)$ into $\delta x(-t)$ leaves these moments unchanged. Another disadvantage of multi-scale increments (6.3) is that they exhibit as many scales $1 \leq \ell \leq N$ as the number of days N, which seems redundant, specially in view of the known scale-invariant properties of x.

1. In this chapter, we reserve the notation T for the maturity of an option, considered in next sections.

The construction of an appropriate statistics $\Phi(x)$ was studied in chapter 2 where we introduced the *Scattering Spectra*, applicable to multi-scale processes. They capture the main non-Gaussian properties of financial prices : fat tailed log-return distributions, sign-asymmetry, time-asymmetry, volatility clustering and volatility roughness. It consists of $M = \mathcal{O}(\log_2^3 N)$ statistics only, that are low-order moments (order 1 and 2 only) and can thus be accurately estimated on the historical realization $\widetilde{x}$ of size N.

We present here the main steps for building such Φ and we refer the reader to chapter 2 for more details about the construction.

Step 1. Wavelet increments.

Log-prices variation have interesting structure at all scales. However, it is not necessary to consider all scales ℓ in (6.3) to characterize them efficiently. Standard increments at scale ℓ (6.3) are obtained by convolution of x with the filter $g_\ell = \delta_0 - \delta_\ell$. Wavelet increments replace g_ℓ by wavelet filters ψ_j obtained by dilation of a regular mother wavelet ψ

$$Wx(t,j) = x \star \psi_j(t) \quad \text{where} \quad \psi(t,j) = 2^{-j}\psi(2^{-j}t). \tag{6.4}$$

The mother wavelet ψ has a zero average $\int \psi(t)\mathrm{d}t = 0$ and its Fourier transform $\widehat{\psi}(\omega) = \int \psi(t)\, e^{-i\omega t}\, \mathrm{d}t$, which is real, is mostly concentrated at frequencies $\omega \in [\pi, 2\pi]$. All numerical calculations in this chapter are performed with a complex Battle-Lemarié wavelet [Battle, 1987 ; Lemarié, 1988]. Fig. 2.1 from chapter 2 shows the real and imaginary parts of ψ as well as its Fourier transform. We refer the reader to section 2.2.3 of chapter 2 for more properties.

Analogous to (6.3), wavelet increments (6.4) can be seen as multi-scale increments at scales $\ell = 2^j$. However, scales are now defined as bins of frequencies $[2^{-j}\pi, 2^{-j+1}\pi]$ corresponding to the supports of wavelet filters ψ_j. The largest scale 2^J is chosen to be smaller than the size N of $\widetilde{x}$. This yields at most $\log_2 N$ scales instead of N lags ℓ.

Histograms of generalized increments Wx can be constrained by order 1 and order 2 moments $\mathbb{E}\{|Wx(t,j)|\}$, $\mathbb{E}\{|Wx(t,j)|^2\}$ which are estimated through empirical averages. The quantity

$$\Phi_1(x)[j] = \frac{\langle |Wx(t,j)| \rangle_t^2}{\langle |Wx(t,j)|^2 \rangle_t} \tag{6.5}$$

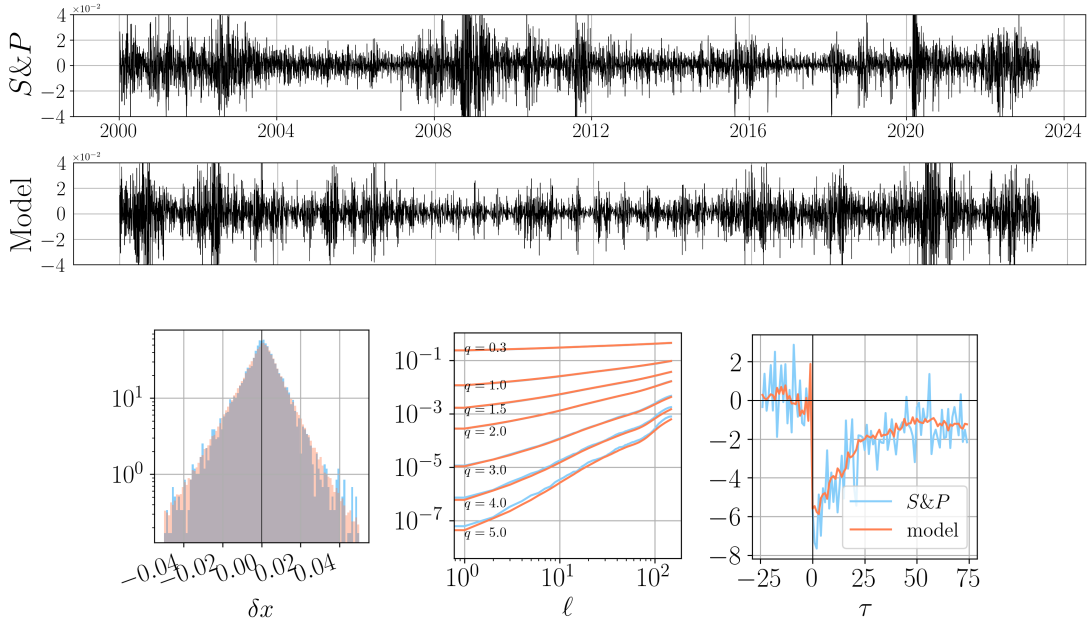
is a low-moment measure of kurtosis. Compared to its order 4 counterpart, it is less sensitive to large values. The more peaked at zero the distribution, the smaller the value of $\Phi_1(x)$ and the higher the kurtosis [Bouchaud, 2013]. The order 2 moment is

$$\Phi_2(x)[j] = \left\langle |Wx(t,j)|^2 \right\rangle_t \tag{6.6}$$

and quantifies the average volatility at scale 2^j on the period.

Step 2. Time-scale dependencies.

Multi-scale increments $Wx(t,j)$ are indexed by time t and scale 2^j. Such map exhibits dependencies across time and scales that are crucial to characterize the distribution of financial prices. For example, volatility clustering is attested by the fact that $Wx(t,j)$ has long-range time correlations. Beyond this well-known stylized fact, we have shown in chapter 2 that scale



(a) Histogram (b) Struct. functions (c) Leverage

FIGURE 6.1 – Standard statistics of log-returns in the Scattering Spectra model (orange) compared to S&P observed data (blue). Top graphs : time-series of the S&P and generated by the model. Bottom graphs : (a) Histogram of daily log-returns δx. (b) Structure functions $\langle |\delta_\ell x(t)|^q \rangle_t$. (c) Leverage correlation $\langle \delta x(t-\tau) |\delta x(t)|^2 \rangle_t$ on normalized increments. Remarkably, the model based on low-moment spectra is able to capture up to order 5 statistics.

dependencies are crucial to fully characterize the non-Gaussian nature of time-series. Natural descriptors for such scale dependencies are order 2, 3 and 4 moments

$$\mathbb{E}\{Wx\, Wx^*\}\ ,\ \mathbb{E}\{Wx\, |Wx|^2\}\ ,\ \mathbb{E}\{|Wx|^2\, |Wx|^2\}$$

where the products are taken across times t, t' and scales j, j'. In practice, estimating order 3 and order 4 moments is very difficult because of the variance induced by large events. In order to circumvent this problem, we replace $|Wx|^2$ by $|Wx|$ and define the following non-linear correlations of wavelet increments

$$\mathbb{E}\{Wx\, Wx^*\}\ ,\ \mathbb{E}\{Wx\, |Wx|\}\ ,\ \mathbb{E}\{|Wx|\, |Wx|\} \tag{6.7}$$

Owing to the compression properties of wavelets, the first matrix $\mathbb{E}\{Wx\, Wx^*\}$ is quasi-diagonal and its diagonal coefficients are already estimated by (6.6), see chapter 2.

Step 3. Low-moment multi-scale skewness and kurtosis.

Just like for standard skewness and kurtosis that are normalized moments, we normalize the second and third matrices $\mathbb{E}\{Wx\, |Wx|\}$ and $\mathbb{E}\{|Wx|\, |Wx|\}$ in (6.7) by $\mathbb{E}\{|Wx|^2\}$. One can show that the only non-negligible coefficients in the third matrix are obtained for $t = t'$ and $j \geq j'$,

they are estimated through

$$\Phi_3(x)[j,j'] = \frac{\langle Wx(t,j)\,|Wx(t,j')|\rangle_t}{\langle |Wx(t,j)|^2\rangle_t^{\frac{1}{2}}\,\langle |Wx(t,j')|^2\rangle_t^{\frac{1}{2}}}. \tag{6.8}$$

These are multi-scale extensions of the standard low-moment skewness $\mathbb{E}\{Y|Y|\}$ of a normalized random variable Y. Other than sign-asymmetry, these complex coefficients also measure time-asymmetry through their phase. Indeed, if log-returns are time-reversible $\delta x(-t) \overset{d}{=} \delta x(t)$ then $\operatorname{Im}\Phi_3(x) = 0$. One typical example is the leverage asymmetric correlation.

The fourth matrix $\mathbb{E}\{|Wx|\,|Wx|\}$ in (6.7) contains kurtosis information. If x is Gaussian, then for different scales $j \neq j'$ the Gaussian processes $Wx(t,j)$ and $Wx(t,j')$ are decorrelated, thus independent. It follows that $\mathbb{E}\{|Wx(t,j)||Wx(t',j')|\} = \mathbb{E}\{|Wx(t,j)|\}\mathbb{E}\{|Wx(t',j')|\}$ and these coefficients boil down to the low-moment kurtosis (6.5). For the log-price process x, these coefficients capture long-range non-Gaussian correlation between volatility at different scales j, j' and different times t, t'.

However, matrix $\mathbb{E}\{|Wx|\,|Wx|\}$ contains too many coefficients to be accurately estimated on a single realization $\widetilde{x}$. We again rely on compression properties of wavelets to approximate such matrix by cascading a second wavelet operator W, which yields a quasi-diagonal matrix $\mathbb{E}\{W|Wx|\,W|Wx|^*\}$ where we define generalized increments of volatility as

$$W|Wx|(t,j_1,j_2) = |x \star \psi_{j_1}| \star \psi_{j_2}(t).$$

The non-negligible diagonal coefficients are estimated through an empirical average which yields for $j_1 \leq j_1' < j_2$

$$\Phi_4(x)[j_1,j_1',j_2] = \frac{\langle W|Wx|(t,j_1,j_2)\,W|Wx|(t,j_1',j_2)^*\rangle_t}{\langle |Wx(t,j_1)|^2\rangle_t^{\frac{1}{2}}\,\langle |Wx(t,j_1')|^2\rangle_t^{\frac{1}{2}}}. \tag{6.9}$$

These are multi-scale extensions of the standard low-moment kurtosis. These complex coefficients also capture time-asymmetry through their complex phase. If the log-return process δx is time-reversible then $\operatorname{Im}\Phi_4(x) = 0$.

We therefore define our Scattering Spectra Φ as the collection of (i) estimated average volatility (6.6), (ii) multi-scale skewness (6.8) and (iii) multi-scale kurtosis (6.5,6.9)

$$\Phi(x) = \big(\Phi_1(x), \Phi_2(x), \Phi_3(x), \Phi_4(x)\big). \tag{6.10}$$

In total, Φ consists of $\mathcal{O}(\log_2^3 N)$ order 1 and order 2 statistics for a trajectory of size N and can be estimated with low-variance.

Note that Φ does not rely explicitly on the one-point distribution of increments $\delta_\ell x(t)$. Numerical experiments have shown that slight discrepancies may appear, in particular in order 0 moments $\mathbb{P}(\delta_\ell x(t) > 0)$ which explicitly appear in low-moment smile expansions [Bouchaud, 2013]. We thus complement $\Phi_3(x)$ with the moments

$$\mathbb{P}(\delta_\ell x(t) > 0)$$

for $\ell = 2^j, j = 1, \dots, J$, that are constrained through empirical averages $\langle \text{sigmoid}(\delta_\ell x(t)) \rangle_t$ where $\text{sigmoid}(x) = (1 + e^{-x})^{-1}$. This adds very few coefficients to our Scattering Spectra $\Phi(x)$.

The Scattering Spectra (6.10) thus provide an enriched set of statistics that can be used to quantify model error and interpret any discrepancy. As an example, we revisit through this lens the state-of-the-art, low-parametric Path-Dependent Volatility (PDV) model introduced in a paper by Guyon & Lekeufack[Guyon, 2022]. We show that several stylized facts are actually not accurately reproduced by such a model, see Appendix D.2, Fig. D.3.

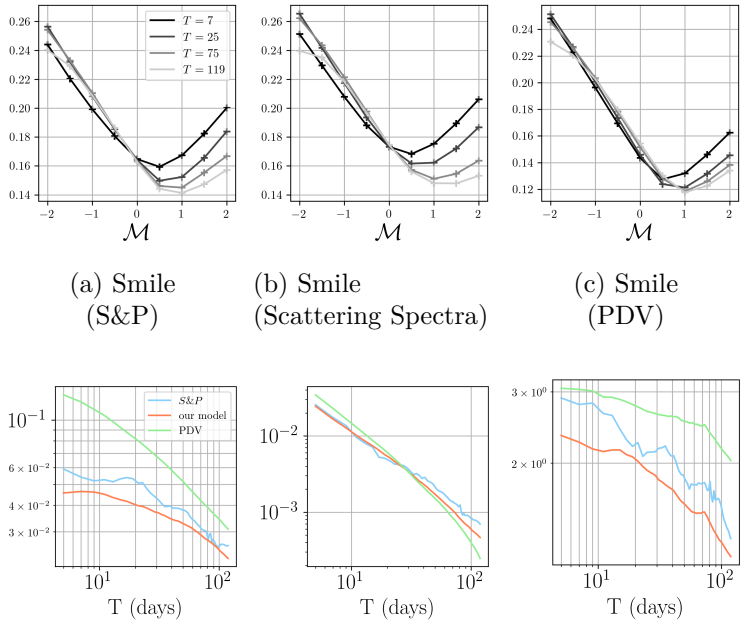
Based on the Scattering Spectra Φ, we have at our disposal a statistical model of financial prices that can be used to generate faithful synthetic time-series (see section 6.2.1). For the S&P time-series $\widetilde{x}$ of size $N = 5827$ days, the Scattering Spectra model contains $248 \approx N/20$ real coefficients, which is the dimension of $\Phi(x)$. Log-return trajectories δx generated from the Scattering Spectra model are shown in Fig. 6.1. Validation of the Scattering Spectra model can be achieved by measuring observables not included in our set $\Phi(x)$ and checking whether or not they are correctly reproduced. Standard statistics such as volatility clustering, leverage effect and structure functions, were indeed shown to be captured by the model, see chapter 2. These are reproduced in Fig. 6.1. While $\Phi(x)$ is composed of order 1 and order 2 moments only, the Scattering Spectra model accurately accounts for up to order 5 moments, which is quite remarkable. Another way to describe the multi-scale statistical properties of price time-series is through maturity dependent option smiles, which we discuss in the next section.

6.3 The average smile as an alternative statistical characterization

In this section we validate the Scattering Spectra model by considering historical option pricing as an alternative, operational way to characterize the multi-scale, non-Gaussian statistics of price time-series. The *average smile* is the unconditional option smile obtained by pricing hedged options using all historical snippets of prices of length equal to the maturity of the option [Potters, 2001 ; Bouchaud, 2013]. Even if real option smiles must be conditioned on a specific past price path [Guyon, 2022] and are therefore almost never equal to the *average smile*, its shape reveals some interesting, non-trivial properties of prices time-series, such as volatility "roughness" (see below).

Option pricing is performed through the Hedged Monte-Carlo method [Potters, 2001], that converts historical probabilities into "risk-neutral" ones. Options are hedged daily, with zero interest rate, on the 6000 price snippets of lengths 150 days available from 2000 to 2023 all initialized at 100. The average implied volatilities $\sigma(T, K)$ are obtained from option prices $\mathcal{C}(T, K)$. Fig. 6.2 compares, for different maturities T, the *average smiles* using observed S&P data and those generated with the Scattering Spectra model. We see that the model indeed reproduces the overall shape of the smile very well. Intuitively, the level of the average smile, its asymmetry, its concavity and its term structure are captured by Φ_2 (6.6), Φ_3 (6.8) and Φ_1 and Φ_4 (6.56.9) [2]. We

2. Appendix D.1 shows in more details the parameterization of the model by studying the sensitivity of the smile to the Scattering Spectra statistics $\Phi(x)$.



(a) Smile (S&P) (b) Smile (Scattering Spectra) (c) Smile (PDV)

(d) ATM skew (abs) (e) ATM kurtosis (f) Skew Stickiness

FIGURE 6.2 – Average option smiles estimated using S&P price data (a), in our Scattering Spectra model (b) and in the path-dependent volatility (PDV) model (c). The Scattering Spectra model qualitatively captures the two regimes of the ATM skew as a function of maturity (d), with a cross-over around 20 days [Delemotte, 2023], as well as the power-law of the ATM kurtosis (e) and the behavior of the skew-stickiness ratio (f). The PDV model, on the other hand, fails to capture the amplitude and term structure of the ATM skew (d).

have also compared the S&P average smiles with the recent Path Dependent Volatility model of [Guyon, 2022], which appear to be too "V-shaped", specially for small maturities.

We now turn to a more refined analysis of the slope and curvature of these average smiles. We denote as $\sigma_{\text{ATM}}(T) = \sigma(T, 100)$ the at-the-money volatility and

$$\mathcal{M} := \frac{\ln(\frac{K}{100})}{\sigma^{\text{ATM}}\sqrt{T}}$$

the *rescaled* log-moneyness. The slope S_T and curvature κ_T of a smile at maturity T are defined by the order 2 expansion around the moneyness $\mathcal{M} = 0$

$$\sigma(\mathcal{M}, T) := \sigma_{\text{ATM}}(T)\Big(1 + \mathrm{S}_T\mathcal{M} + \kappa_T\mathcal{M}^2 + o(\mathcal{M}^2)\Big)$$

In the literature, it is customary to define the ATM skew Skew_T as the slope of the smile as a function of *unscaled* log-moneyness, i.e. $\text{Skew}_T := \mathrm{S}_T/\sqrt{T}$. For most stochastic volatility models, such skew is found to be regular when $T \to 0$, whereas rough volatility models predict a singular behavior $\text{Skew}_T \propto T^{H-1/2}$ where H is the Hurst exponent of volatility, argued to be small, $H \approx 0.1$ [Gatheral, 2018 ; Fukasawa, 2021 ; Bayer, 2016].

Fig. 6.2d shows the absolute value of the ATM skew of the average smile for different maturities. The authors of [Delemotte, 2023] have shown using option market data that the ATM skew

exhibits two power-law regimes pertaining to short and long maturities, with a cutoff between the two regimes around 20 days. Strikingly, we also observe this behavior on the average smile of the S&P, which ony depends on the price process, with no reference to actual option markets. The Scattering Spectra model tracks remarkably well such a behavior. The scaling of log-volatility increments characteristic of rough volatility models [Gatheral, 2018] or multifractal models [Bacry, 2013] is in fact encoded in the model through the coefficients $\mathbb{E}\{|W|Wx|(t, j_1, j_2)|^2\}$ included in $\Phi_4(x)$ (6.9). These consider instantaneous volatility $|Wx(t, j_1)|$ at scale 2^{j_1} and its increments at scales 2^{j_2}.

The Scattering Spectra model furthermore captures two more subtle stylized facts of option smiles :

— The ATM curvature κ_T, related to a low moment kurtosis [Bouchaud, 2013], is well captured and behaves as a power-law of T, both for the S&P and within the Scattering Spectra model. The PDV model, on the other hand, slightly overestimates κ_T for small T and underestimates it a large T (see Fig. 6.2e).

— The instantaneous change of ATM volatility $\sigma_{\text{ATM}}(T)$ induced by a change in underlying price can be linearly regressed on δx. This defines the skew-stickiness ratio R_T through the following regression [Bergomi, 2009 ; Vargas, 2015]

$$\delta\sigma_{\text{ATM}}(T) = -R_T \text{Skew}_T \times \delta x + \epsilon$$

As shown in [Vargas, 2015], R_T has a non-trivial dependence on maturity. Fig. 6.2f shows that the Scattering Spectra model again reproduces quite well such a dependence.

6.4 Path Shadowing Monte-Carlo & volatility predictions

The *average smile* exercise of the previous section is interesting insofar as it allows one to test the ability of various models to capture the distribution of price changes over different maturities. However, it fails to inform us on the power of the model to actually predict, at a given date, the distribution of price changes forward in time. Of course, this is what finance is all about and we now introduce a framework to do precisely that.

We first assume that the real world is at least approximately stationary, in the sense that it can be approximated by a statistical model with fixed, time-independent parameters. Of course, this can only be true if the model is rich enough to generate time-series that superficially appear non-stationary – such as the ones shown in Fig. 6.1, with alternating periods of high and low volatility that are actually described by the *same* model.

If this is the case, then given the past history $\widetilde{x}_{\text{past}}$ at current time t, a model-free method for predicting the unknown future $\widetilde{x}_{\text{future}}$ is to look for occurrences similar to $\widetilde{x}_{\text{past}}$ in the historical realization of log-prices. If such occurrences can be found, what happened thereafter provides some information about the unknown future $\widetilde{x}_{\text{future}}$ at the current time t.

Finding exact occurrences of course happens with vanishing probability. We therefore introduce an embedding $h(\widetilde{x}_{\text{past}})$ that reduces the dimensionality of past trajectories and define

shadowing paths as paths x whose past history $h(x_{\text{past}})$ is close to $h(\widetilde{x}_{\text{past}})$ in a certain sense. Furthermore, instead of scanning the historical past, we propose in this section to scan a long dataset of paths generated using the model presented in section 6.2.

These shadowing paths are then used as inputs of our proposed *Path Shadowing Monte-Carlo* (PS-MC) method, which allows us to obtain state-of-the-art predictions for future realized volatility. The method will be extended in the next section 6.5 to option pricing.

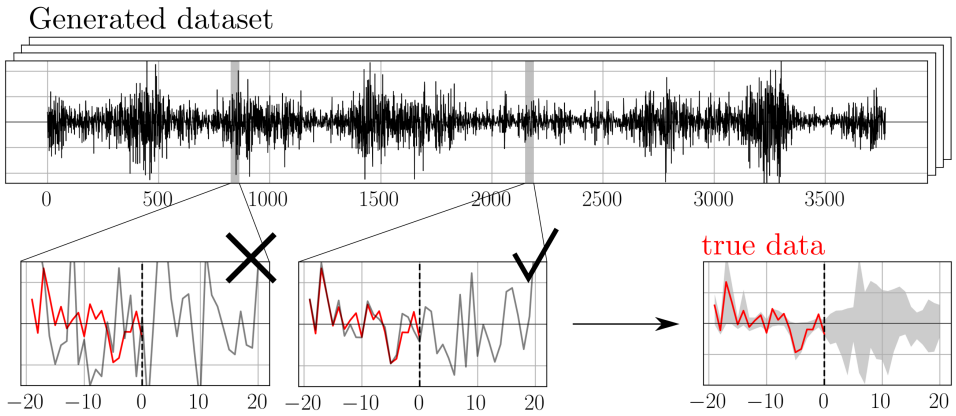


FIGURE 6.3 – Path shadowing. Given current past history $\widetilde{x}_{\text{past}}$ (red), we scan for paths x (black) in a generated dataset whose past history satisfies $x_{\text{past}} \approx \widetilde{x}_{\text{past}}$. Such paths x are said to *shadow* $\widetilde{x}_{\text{past}}$, they provide insights on the future. Predictions are obtained through Monte-Carlo on such shadowing paths.

6.4.1 The Path Shadowing Monte-Carlo method

We first separate a log-price path $x \in \mathbb{R}^N$ into its past $x_{\text{past}} = (x(t), t \leq 0)$ and future $x_{\text{future}} = (x(t), t > 0)$

$$x = (x_{\text{past}}\ ,\ x_{\text{future}})$$

Let $Q(x_{\text{future}}, t)$ be a quantity we want to predict, for example the average realized variance in the next T days $Q(x_{\text{future}}, t) := \sum_{u=t}^{t+T} |\delta x_u|^2/T$. In the following section, we write $Q(x) := Q(x_{\text{future}}, t)$ for simplicity. An optimal prediction of $Q(x)$ for the mean square error as a function of the observed past $\widetilde{x}_{\text{past}}$ is given by the conditional expectation

$$\mathbb{E}\{Q(x) \mid x_{\text{past}} = \widetilde{x}_{\text{past}}\} \tag{6.11}$$

with $\mathbb{E}$ the expectation under the true distribution $p(x)$ of log-prices. The goal is to estimate such conditional expectation.

Let us for a moment omit the conditioning on the past. The standard Monte-Carlo method estimates expectations using a finite number of realizations $x^1, \ldots, x^n$ drawn from $p(x)$ as

$$\frac{1}{n}\sum_{k=1}^{n} Q(x^k) \tag{6.12}$$

which converges to $\mathbb{E}\{Q(x)\}$ as $n \to +\infty$ under independence of $x_1, \dots, x_n$ and integrability of $Q(x)$.

In theory, such method could apply to estimate (6.11), however it would require finding paths x^k such that $x^k_{\text{past}} = \widetilde{x}_{\text{past}}$, which is all but impossible for data of finite size.

To tackle this problem, we relax strict conditioning on $\widetilde{x}_{\text{past}}$ and consider paths x whose past x_{past} is *close* to $\widetilde{x}_{\text{past}}$ in a certain sense. In order to account for possible long-range dependencies, we would like to consider a long past $\widetilde{x}_{\text{past}}$. However, finding paths x_{past} at a given distance from $\widetilde{x}_{\text{past}}$ becomes exponentially difficult as the size of the path grows – this is the curse of dimensionality. In order to control the dimension, we consider a path embedding $h(x_{\text{past}}) \in \mathbb{R}^M$. Given a threshold $\eta > 0$ we define the set of *η-shadowing paths* as

$$H_\eta(\widetilde{x}_{\text{past}}) = \{x \in \mathbb{R}^N \mid \|h(x_{\text{past}}) - h(\widetilde{x}_{\text{past}})\| < \eta\} \tag{6.13}$$

For example, $h(x) = \delta x$ considers past log-returns, and hence log-price paths up to an additive constant. Our choice of h is detailed in the next section. The term *shadowing* is freely inspired by the shadowing principle in chaotic dynamical systems [Anosov, 1969 ; Sinai, 1972 ; Bowen, 1975 ; Hammel, 1987]. The idea is that for a certain small η, paths in $H_\eta(\widetilde{x}_{\text{past}})$ can be considered as true realizations of the process that can be used to faithfully compute observables.

Path Shadowing Monte-Carlo is a Monte-Carlo method on shadowing paths. It is a predictive method since shadowing paths span over the future. Unlike a standard Monte-Carlo method, not all paths should have the same weight since $\|h(x^k_{\text{past}}) - h(\widetilde{x}_{\text{past}})\|$ is not uniform in k. This is to say, certain paths *shadow* more accurately x_{past} than others and should be considered as more *likely* to be extensions of the observed $\widetilde{x}_{\text{past}}$. Path Shadowing Monte-Carlo estimates (6.11) by averaging $Q(x)$ on paths $x^1, \dots, x^n$ with weights $w_1, \dots, w_n$. This yields the following estimator

$$\frac{1}{n} \sum_{k=1}^{n} w_k Q(x^k), \tag{6.14}$$

called the Nadaraya–Watson estimator [Nadaraya, 1964 ; Watson, 1964]. In the following, we choose Gaussian weights, to wit

$$w_k = c \exp\left[-\frac{\|h(x^k_{\text{past}}) - h(\widetilde{x}_{\text{past}})\|^2}{2\eta^2}\right]$$

with c such that $\frac{1}{n}\sum_{k=1}^{n} w_k = 1$. The set of shadowing paths $H_\eta(\widetilde{x}_{\text{past}})$ (6.13) can be defined as the set of all paths that are one standard deviation away from $\widetilde{x}_{\text{past}}$ for the Gaussian kernel.

The following theorem proves the convergence of the estimator (6.14) under standard hypotheses.

Theorem 4 (Path Shadowing Monte-Carlo Method)**.** *If Q is continuous with $\mathbb{E}\{Q(x)\} < +\infty$ and the distribution p of x is continuous with $p(x) > 0$ for all $x \in \mathbb{R}^N$, then given $x^1, \dots, x^n, \dots$ independent realizations of x, the Path Shadowing Monte-Carlo estimator with*

$h(x) = x$ converges almost surely

$$\frac{1}{n}\sum_{k=1}^{n} w_k Q(x^k) \longrightarrow \mathbb{E}\{Q(x) \mid x_{past} = \widetilde{x}_{past}\}$$

for a certain limit $n \to +\infty$ and $\eta \to 0$.

The proof is in Appendix D.3. The continuity assumptions as well as the assumption that $p > 0$ are technical assumptions and can be softened ; note that the p_θ in our model 6.1 satisfies them. We refer the reader to [Hansen, 2008] for convergence theorems under more generic hypotheses.

Path Shadowing Monte-Carlo is a kernel method on log-price paths using a Gaussian kernel. It is a fully non-local method in the sense that the collected paths x^k may be far away in the past from $\widetilde{x}_{\text{past}}$, possibly in a generated dataset of paths, contrary to non-local means [Buades, 2011] that only consider neighborhoods of a patch. Such non-locality is in practice what ensures the independence condition in theorem 4.

6.4.2 Generating shadowing paths

Collecting enough shadowing paths from the historical realization of S&P is unrealistic. The set $H_\eta(\widetilde{x}_{\text{past}})$ will contain almost no paths for reasonable values of η, required to be small for the method to converge.

We thus propose to scan for paths $x^1, \dots x^n$ in a long generated dataset of log-prices, allowing us to take $n \gg 1$ and selective shadowing threshold $\eta \ll 1$. This however immediately introduces a modelling error, due to the fact that we are estimating (6.11) where $\mathbb{E}$ is now the expectation with regard to the model distribution and not the true distribution $p(x)$.

A good model should generate trajectories that capture the long-range dependencies in order for the shadowing paths to have predictive power on the future of $\widetilde{x}_{\text{past}}$. It should also generate a variety of realistic scenarios in order to find enough paths in $H_\eta(\widetilde{x}_{\text{past}})$ for small η. As discussed in section 6.2 the Scattering Spectra based model precisely meets these requirements : it is *realistic*, in the sense that it accurately captures many stylized facts of financial time-series, and it is *versatile*, in the sense that its maximum entropy formulation allows us to generate easily a very large number of representative samples. Furthermore, should our generative algorithm produce occasionally unrealistic paths, such paths would be given a very small weight w and would not contribute to the estimation of $Q(x)$. Shadowing Monte-Carlo is thus robust to outliers in the generated set of paths. Another aspect of our method is that the dataset can be generated once and can be scanned again and again for several prediction dates.

A crucial point for PS-MC to give good results is to understand how the path embedding h affects the notion of path proximity. Such embedding should be chosen adequately. It should pick relevant features of x_{past} to predict the quantity of interest $Q(x)$, while remaining low-dimensional such that enough paths can be found in $H_\eta(\widetilde{x}_{\text{past}})$ for small η. The naive embedding $h(x) = \big(\delta x(t), -M_{\text{past}} + 1 \leq t \leq 0\big) \in \mathbb{R}^{M_{\text{past}}}$ limits the past horizon M_{past} which is also the dimensionality of h.

We propose a representation h that again leverages the scale-invariance of x in the same

way our Scattering Spectra framework does, and incorporates the influence of distant past while remaining low-dimensional. It consists of multi-scale increments in the past with a power-law decaying weight

$$h_{\alpha,\beta}(x) = \Big(\frac{x(t) - x(t-\ell)}{\ell^\beta} \ , \ \ell = \lfloor \alpha^m \rfloor \ , \ m = 1, 2, \dots \Big) \tag{6.15}$$

for a certain $\alpha > 1$ so that the past is progressively coarse-grained, and $\beta \geq 0$ so that the far away past is progressively discounted. For a given $\widetilde{x}_{\text{past}}$ we choose η to be equal to $\bar{\eta}\|h(\widetilde{x}_{\text{past}})\|$ for certain $\bar{h} > 0$, which amounts to normalize the distance $\|h(x_{\text{past}}) - h(\widetilde{x}_{\text{past}})\|$ by $\|h(\widetilde{x}_{\text{past}})\|$. Such h is a discretization of a continuous h that satisfies scaling and dilation equivariance, see Appendix D.4. In practice we truncate the progression $\lfloor \alpha^1 \rfloor, \lfloor \alpha^2 \rfloor, \dots$ in order for the span of h to be bounded by 126 trading days in the past (corresponding to half a year).

The main parameters are thus α, β and $\bar{\eta}$. Parameter α determines the dimensionality of the path embedding $h_{\alpha,\beta}$, small α yields high-dimensional embedding. Parameter β rules the relative importance of distant past to recent past in the selection of shadowing paths. Large $\beta > 0$ means that the recent past bears more weight. Finally, there is a bias-variance trade-off in the choice of $\bar{\eta}$. When $\bar{\eta} \ll 1$ only the closest path will be used for averaging, leading to large variance estimator (6.14). When $\bar{\eta} \gg 1$ then all paths are averaged uniformly, including paths whose past x_{past} has nothing to do with $\widetilde{x}_{\text{past}}$, thus deteriorating the bias of our PS-MC estimator.

6.4.3 Volatility prediction

As a meaningful application of Path Shadowing Monte-Carlo, we consider in this section the prediction of the future daily realized variance over T days, for several values of T :

$$Q_T(x) = \frac{252}{T} \sum_{t=1}^{T} |\delta x_t|^2. \tag{6.16}$$

We consider all 2112 dates from January 2015 to March 2023. For each of them we consider the realized variance over $T = 1, 7, 25, 75, 150$ days.

Our PS-MC method (6.14) uses paths generated from the model presented in section 6.2. We compute the Scattering Spectra statistics (6.10) on observed 3772 S&P log-prices from January 2000 to December 2014, such that all our predictions are *out-of-sample*. From such statistics we generate 32 768 trajectories of same size 3772 (see Fig. D.4 for examples), that represents $n \approx 115$ million paths $x^1, \dots, x^n$ of size 126+150 days. For a given $\widetilde{x}_{\text{past}}$ we scan such dataset and select the 50 000 closest paths in the sense of the distance induced by embedding (6.15), parameterized by α, β. While this scanning step is fastidious, it can be fully parallelized. We then perform a weighted average on those closest paths, parameterized by $\bar{\eta}$.

Parameters $\alpha, \beta, \bar{\eta}$ are calibrated using our generative model itself, in order to avoid any over-fitting on the limited train data from S&P. We choose those parameters such that $Q_T(x)$ is optimally predicted within the model. More precisely, we take 1100 snippets $\widetilde{x}_{\text{past}}$ from the generated dataset, for which we have access to $\widetilde{x}_{\text{future}}$. We obtain an estimate of Q_T for these

1100 dates and $T = 7, 25, 75, 150$. We then choose the best $\alpha, \beta, \bar{\eta}$ so as to maximize the R^2 score between prediction and true values of the Scattering Spectra model itself. This yields the following optimal parameters : $\alpha = 1.15, \beta = 0.9, \bar{\eta} = 0.075$ and a path embedding $h_{\alpha,\beta}$ of dimension 34.

Let us note that these optimal values barely change when predicting realized variance at different maturities. This is because of the scale-invariance of both the model and of the path embedding. Note also that $\alpha = 1.15$ in (6.15) means that the values $\ell = 1, 2, 3, 4$ appear multiple times. This amounts to ascribing an even larger weight to small time lags.

Number of days T	1	7	25	75	150
Benchmark	-0.16	0.43	-0.05	-0.58	-0.79
Path Dependent Vol	**0.37**	**0.56**	0.29	-0.01	-0.08
Path Shadowing MC	0.32	**0.56**	**0.33**	**0.07**	**0.01**

TABLE 6.1 – Prediction of realized daily volatility through Path Shadowing Monte-Carlo (R^2 scores). The PS-MC method based on Scattering Spectra outperforms the recently introduced PDV model [Guyon, 2022] at all time-scales ≥ 7 days, and upholds predictive power up to a period of ≈ 150 days. For $T = 1$ day, however, the PDV model performs best. The benchmark is simply the realized variance on the T previous days.

The prediction quality is measured through the R^2 score on future volatility estimates and are shown in Table 6.1 for different maturities T. As a simple benchmark we consider the realized variance on the T previous days as a predictor of $Q_T(x)$. As a second, more challenging, benchmark we consider the recent path dependent volatility (PDV) model introduced in [Guyon, 2022], which reads

$$\sqrt{Q_T(x)} = \beta_0 + \beta_1 F_{1,t} + \beta_2 \sqrt{F_{2,t}} \tag{6.17}$$

$$\text{with} \quad F_{1,t} = k_1 \star \delta x(t) \ , \ F_{2,t} = k_2 \star |\delta x|^2(t),$$

where k_1 and k_2 are two power-law kernels acting on past returns and past absolute returns. We take the very same kernels as specified in [Guyon, 2022] but optimize the regression coefficients $\beta_0(T), \beta_1(T), \beta_2(T)$ for each maturity T separately, on the same train period as for PS-MC, i.e. from January 2000 to December 2014. [3]

Using the very same shadowing paths for all maturities, our Path Shadowing Monte-Carlo method outperforms both the naive benchmark and the PDV model for all maturities from $T > 7$ days, and ties with PDV for $T = 7$ days, see Table 6.1. In particular, our method upholds predictive power up to ≈ 150 days, which none of the two other methods are capable of. This is, we believe, due to the scale-invariance of both the Scattering Spectra generative model and our choice of path embedding (6.15). Again, we insist on the fact that the PDV model parameters are refitted for each maturity T whereas the Scattering Spectra model is calibrated once and for all.

3. Note that the authors of Ref. [Guyon, 2022] estimate realized daily variance using 5-min ticks for better estimation, but the scores we obtain with daily ticks are actually similar for the longest maturities $T = 3$ and $T = 5$ that were tested in their study. Hence, we do not think that using 5-min ticks would substantially change the conclusions reached below for $T \geq 7$ days.

These results vindicate both the generative model presented in section 6.2 and the PS-MC method of the present section. In particular the Scattering Specra generative model, based on 182 parameters, is *not* over-fitting the training dataset.

6.5 Option pricing & trading games

In section 6.3, we have used the Hedged Monte-Carlo method to price options *unconditionally* within the Scattering Spectra model, i.e. by averaging over all possible price paths of a given length T. This allowed us to obtain *average smiles* as a function of maturity, which we compared to those obtained using the same procedure but with real S&P trajectories.

Now, at a given date, option prices reflect anticipations of the market, conditioned on present market conditions – in particular the past price trajectory – and any available information about the future, such as earning announcements, dividends, political events, etc. Of course, such events cannot be directly captured by a purely statistical model, however faithful it might be. Still, it is interesting to run the exercise of pricing option smiles that anticipates the future solely based on the past of underlying price process.

In this section we investigate this question by combining the Scattering Spectra generative model presented in section 6.2 with Path Shadowing introduced in section 6.4. Option prices are then obtained by upgrading the Hedged Monte-Carlo method [Potters, 2001] with, as an input, shadowing paths generated by the model. The overall level of the resulting smiles at time t is nothing but the prediction of the future realized volatility for $t' \in [t, t+T]$, which was already investigated in the previous section. We now extend such prediction to the full shape of the smile. We assess the quality of our smiles by trading a buy-sell signal on options whenever the model option smile is telling us that the option is under-priced or over-priced compared to another smile model, or of the option market itself.

6.5.1 Path Shadowing hedged Monte-Carlo

Hedged Monte-Carlo (HMC) [Potters, 2001] enables one to use time-series of prices to compute the option prices. It iteratively determines the optimal price and hedging strategy by minimizing the expected financial risk of a portfolio containing the option to be priced and its hedge, at all times $t = T-1, T-2, \ldots, 0$. The expected risk is computed as an average over paths, which in the present context are the shadowing paths, based on the notion of distance induced by the path embedding h, (6.15). This defines the Path Shadowing Hedged Monte-Carlo (PS-HMC) that can be used in a versatile way to price any derivative contract. We use the same Gaussian weights given by Eq. (6.14) and the very same parameters α, β, η detailed in section 6.4.3, that are optimal for volatility prediction within the model itself.

Fig. 6.4 shows the resulting smiles as a function of rescaled log-moneyness, for different maturities and at 3 typical dates. As one would have hoped, the level, but also the slope and the curvature of those smiles do depend on the chosen date, and more precisely on the actual path trajectory of the price before that date.

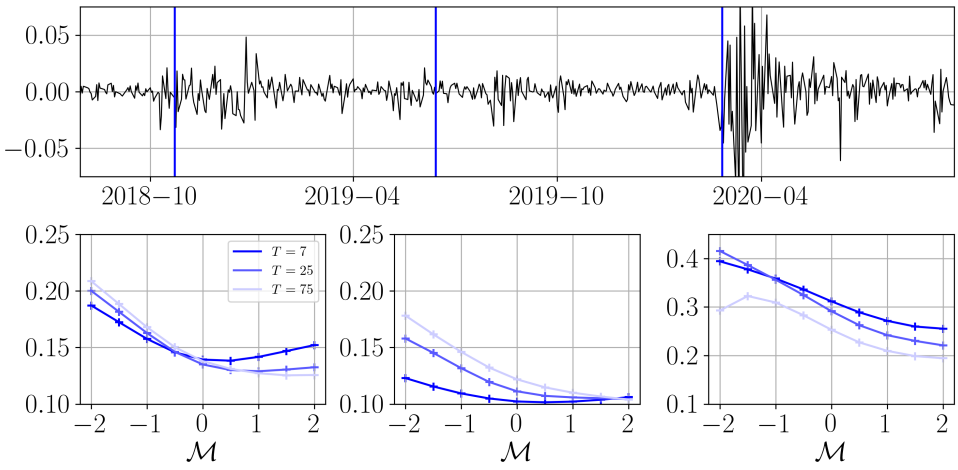


FIGURE 6.4 – Conditional smiles obtained from hedged Monte-Carlo on shadowing paths generated using the Scattering Spectra model at 3 different dates, 2018-10-23, 2019-06-14 and 2020-02-26. Note that the level, the slope and the curvature of those smiles strongly depend on the chosen date.

6.5.2 Validation through trading game

In order to assess the quality of the smiles predicted by the Scattering Spectra model, we set up the following trading game. We trade call options at several dates t on the option market. We neglect bid-ask spread and consider the option price $\mathcal{C}^{\mathrm{MKT}}(t,T,K)$ to be the quoted mid-price. We denote $\sigma^{\mathrm{MKT}}(t,T,K)$ the observed implied volatility and $\sigma^{\mathrm{model}}(t,T,K)$ the implied volatility computed within the model that we decide to trade with.

We then test the following trading strategy : buy the corresponding option from the market whenever we deemed it under-priced, i.e. $\sigma^{\mathrm{MKT}}(t,T,K) < \sigma^{\mathrm{model}}(t,T,K)$ or sell it if we deem it over-priced $\sigma^{\mathrm{MKT}}(t,T,K) > \sigma^{\mathrm{model}}(t,T,K)$. We then follow the hedged option until maturity and register the corresponding profit or loss associated to the trade.

The buy-sell signal of such strategy is thus given by

$$\epsilon_t = \begin{cases} +1 & \text{if } \sigma^{\mathrm{MKT}}(t,T,K) < \sigma^{\mathrm{model}}(t,T,K), \\ -1 & \text{if } \sigma^{\mathrm{MKT}}(t,T,K) > \sigma^{\mathrm{model}}(t,T,K). \end{cases}$$

The un-hedged forward $\mathrm{P\&L}_t(T,K)$ of one transaction is obtained as

$$\mathrm{P\&L}_t(T,K) = v_t \epsilon_t \Big((e^{x_{t+T}} - K)_+ - \mathcal{C}^{\mathrm{MKT}}(t,T,K) \Big) \tag{6.18}$$

where v_t is the volume of option traded. In order to remove non-stationary effect due to the long-term change in the value of the underlying, we take $v_t = 100/S_t = 100e^{-x_t}$ which amounts to trade options on percentage of variation of the underlying rather than the underlying itself.

To reduce the variance of the strategy, we hedge the option using a simple Black-Scholes delta-hedge with a constant volatility 0.2. Such delta-hedge gives zero profit on average but reduces greatly (although not optimally, see [Potters, 2001]) the variance of $\mathrm{P\&L}_t$.

In the following, we will consider the model smile σ^{model} to be given either by the smile computed in the Scattering Spectra model using PS-HMC or the smile computed in a Path Dependent model [Guyon, 2022], both using HMC. The two models are calibrated using the same data in the train period i.e. January 2000-December 2014. As in the previous section, the PDV model parameters are furthermore optimized for each maturity T independently, whereas the Scattering Spectra model is parameterized once and for all with the Scattering Spectra determined in the train period.

The trading game is then in both cases played out-of-sample, for all 2112 dates t from January 2015 to May 2023. We choose 5 maturities $T = 8, 25, 50, 75, 150$ and 9 strikes at constant rescaled log-moneyness $\mathcal{M} = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$.

Detailed P&Ls over 3 different periods of 3 year each are shown in Fig. 6.5. Their variance across dates t is shown in Fig. D.5 in Appendix D.5. For most maturities and strikes, the trading game using the Scattering Spectra model yields positive P&Ls and clearly outperforms the trading game using the PDV model. In fact, one can directly play the Scattering Spectra model against the PDV model without any reference to the actual option market, fully confirming that the Scattering Spectra outperforms PDV for almost all maturities and strikes, see Appendix D.5 and in particular Fig. D.9.

Since the P&Ls are of the same order of magnitude across different strikes and maturities, we average them over all the maturities and strikes. Table 6.2 shows such grand averages and reveals that the trading game using the Scattering Spectra model is significantly more profitable than using the PDV model, with furthermore less variance across different periods. This is confirmed by the aggregated P&Ls over time, see Fig. 6.6.

	full period	2015-2017	2018-2020	2021-2023
PDV	0.03 ± 0.05	**0.15** ± 0.03	-0.12 ± 0.08	0.07 ± 0.04
Scattering Spectra	**0.13** ± 0.05	0.14 ± 0.03	**0.14** ± 0.07	**0.11** ± 0.04

TABLE 6.2 – Average P&L of the trading game against the S&P market using the PDV model or using the Scattering Spectra model. Detailed P&Ls are shown in Fig. 6.5.

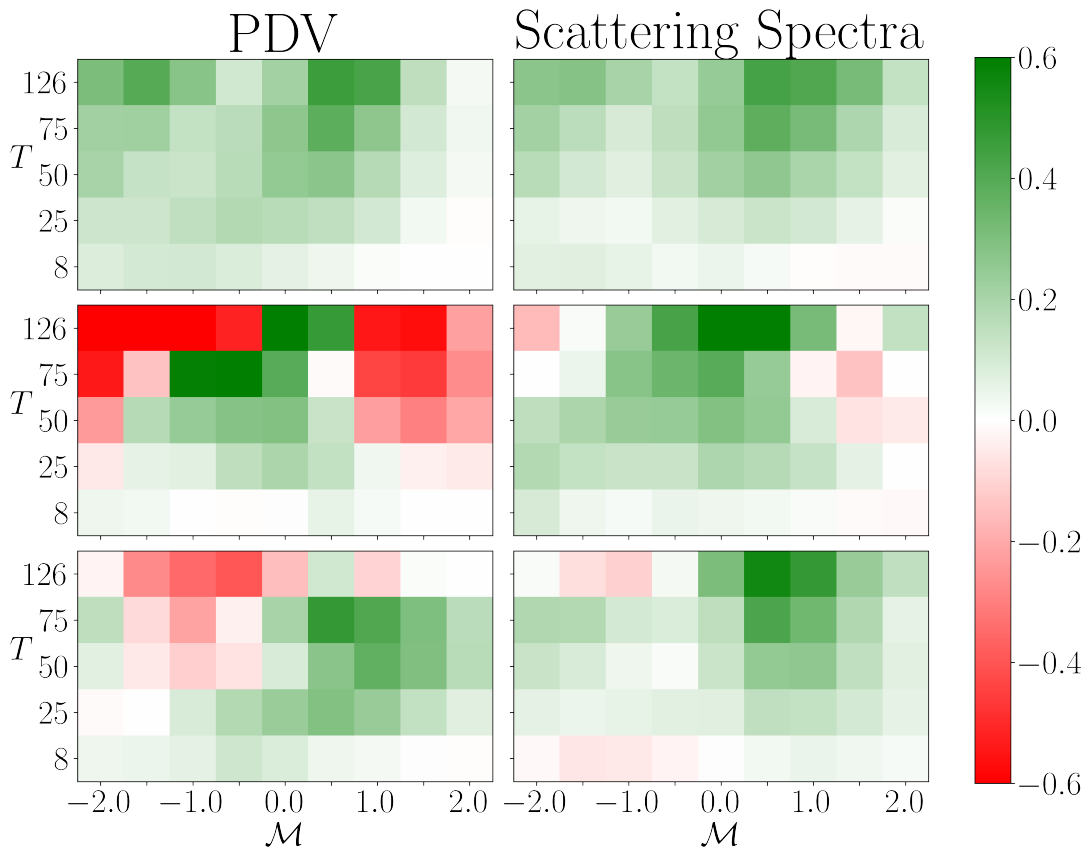


FIGURE 6.5 – P&Ls of the trading game against the S&P market with a PDV model or with the Scattering Spectra model. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

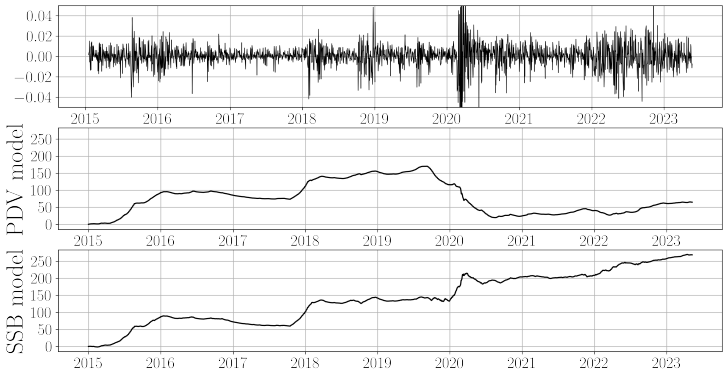


FIGURE 6.6 – Aggregated P&Ls of the trading game against the S&P market with a PDV model or with the Scattering Spectra model. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

6.6 Conclusion

We presented a statistical model of financial prices based on the Scattering Spectra introduced in chapter 2. Scattering Spectra are multi-scale extensions of the standard skewness and kurtosis. Such a model achieves a tradeoff between accuracy and diversity. It captures main statistical properties of prices as well as recently discovered scaling properties of option smiles. As a maximum entropy model with a small number of constraints, our Scattering Spectra model also avoids over-fitting.

We then introduced Path Shadowing Monte-Carlo (PS-MC) which enables building models of forward looking probabilities $p(x|x_{\text{past}})$ from any generative model of $p(x)$. Combined with our statistical model of prices, PS-MC provides state-of-the-art volatility predictions. Shadowing paths can also be used to obtain option smiles that depend only on the distribution of the price process. A trading game allowed us to show that the Scattering Spectra model better anticipates future price movements than other recently introduced models.

One limitation of PS-MC is that it requires to scan a large dataset of generated paths for delivering good performances. This scanning step could be done more efficiently. In particular, could one find "typical" price paths that should be frequently selected for prediction in order to save intensive scanning efforts ?

Beyond prediction, we believe that Path Shadowing is a way of tackling other burning questions in Finance, such as *typicality* : how typical or atypical is a given sequence of price changes ? Another highly relevant extension is towards the description of multivariate time-series. We hope to address these issues in the near future using the methods introduced in this work.

Chapitre 7

Conclusion

This thesis introduces models of multi-scale processes that can be estimated using limited data. These models serve multiple purposes, including data generation, source separation, and prediction.

7.1 Summary of contributions

7.1.1 Scattering Spectra

The first important problem tackled in this thesis is to define a vector of statistics that specifies key properties of multi-scale, stationary, ergodic, non-Gaussian processes x which are commonly encountered in various fields such as Finance and Physics, and that can be estimated on limited data.

While state-of-the-art representations, based on phase-harmonics [Portilla, 2000 ; Zhang, 2021], introduced non-linear correlations that capture scale dependencies, which are crucial to model non-Gaussianity, they exhibit too many coefficients resulting in large model variance.

A main contribution of this thesis is to introduce the Scattering Spectra, a reduced representation of scale dependencies which leverages the multi-scale nature of the data. It is achieved through a diagonal approximation, after a second wavelet transform, of the joint correlation across time and scales of wavelet coefficients and their modulus. It extends the standard scattering moments of order 1 [Mallat, 2012 ; Bruna, 2013], by computing correlation across separate scale channels.

When incorporated into a source separation framework [Regaldo-Saint Blancard, 2021], this representation demonstrates the ability to effectively separate transient events in Mars seismic data known as glitches and to clean observed Marsquakes which are essential to the study of the interior of Mars.

Maximum entropy models based on Scattering Spectra are shown to provide accurate models of financial and turbulent time-series. They can be extended to two-dimensional fields and provide a compact characterization of dependencies between oriented scales.

Multi-channel models of time-series can be constructed by constraining the Scattering Spec-

tra of certain factors. Such models are shown to reproduce important non-linear dependencies across financial stocks, including joint time-asymmetry.

7.1.2 Wide-Sense Self-Similarity

Among multi-scale processes self-similar processes exhibit some form of scale invariance. An important problem is to define and detect this scale regularity on limited data.

The main characterizations of self-similarity are based on structure functions, which involve high-order moments on wavelet coefficients, and assume they have a power-law scaling. Multifractal analysis then relates the scaling exponents to properties of the signal. While this can be used successfully to discriminate and evidence non-trivial scaling behaviors in the data, this provides a relatively weak characterization of self-similarity in the extent that the structure functions do not pick up important non-Gaussian properties such as sign-asymmetry or time-asymmetry.

A main contribution of this thesis is to introduce a wide-sense definition of self-similarity similar to the widely adopted wide-sense definition of time-stationarity in signal processing. This definitions posits that the joint correlation of wavelet coefficients and their modulus across times and scales is invariant to scale shift, up to a power-spectrum normalization.

Such characterization can be tested numerically and can be used to evidence self-similarity in a financial time-series of prices from the scale of a few minutes to the scale of a decade.

This wide-sense characterization also provides a way of reducing models of self-similar processes, or processes with scale regularity in a broader sense. This enables building models of multi-scale physical fields from very few statistics.

7.1.3 Path Shadowing Monte-Carlo

Another problem we tackle in this thesis is the prediction of multi-scale time-series from limited data. This involves estimating conditional expectation of future quantities conditioned on a given observed past history.

Important reference methods include linear regression and kernel methods. The lack of data constrains linear model to simplistic relations between the past time-series and the quantities to predict. Non-local kernel methods average predictions on “close” data points with a proximity notion given by a kernel. However, when dealing with financial processes characterized by high entropy and significant noise, finding ’close’ paths from limited observed data becomes infeasible.

Inspired by non-local methods, we introduced Path Shadowing Monte-Carlo, which proposes to average predictions over generated data from our Scattering Spectra model. This method exploits the ability of our model to generate a variety of paths with accurate long-range dependencies. We show that it yields state-of-the-art volatility prediction in Finance as well as option pricing through a trading game.

7.2 Perspectives

7.2.1 Multivariate self-similarity

The characterization of joint self-similarity across multiple time-series remains a relatively unexplored area of research. One of the reasons for this is the absence, to date, of a natural extension in a generic setting for the standard structure function tools that underlie the multifractal formalism. Interesting attempts have been made to establish the existence of a multivariate multifractal formalism in specific cases under synchronicity hypotheses [Jaffard, 2019].

For sake of simplicity, let us consider the case of two time-series x and y. A correlation-based representation $\mathbb{E}\{\rho W x\,(\rho W x)^T\}$ can be naturally extended to the case of multi-channel processes by considering $\mathbb{E}\{\rho W x\,(\rho W y)^T\}$. This matrix characterizes important non-linear dependencies, as it was demonstrated in the case of physical fields [Régaldo-Saint Blancard, 2023]. Can we extend the wide-sense self-similarity introduced in chapter 2 as an invariance to scale shift on the matrix $\mathbb{E}\{\rho W x\,(\rho W y)^T\}$? Can we identify a low-dimensional structure on $\mathbb{E}\{\rho W x\,(\rho W y)^T\}$ as a consequence of joint self-similarity ?

7.2.2 Optimal directions in a maximum entropy factor model

Chapter 4 introduced a model of a multi-channel time process $x(t) = (x_1(t), \ldots, x_C(t))$ by choosing a small number of directions $w^1, \ldots, w^r \in \mathbb{R}^C$ for which the time-structure of the process projected on these directions $\langle w, x\rangle$ is essential to capture the joint time-structure of x. For that we chose two types of directions, well-known PCA directions and sparse directions, obtained by dictionary learning. These directions are defined independently of the problem of modeling x.

A question remains : what are "optimal" directions $w^1, \ldots, w^r \in \mathbb{R}^N$ across channels whose projections $\langle w, x\rangle$ drive the joint process ? Which criteria should be used to pose the optimization problem ? Do they coincide with known bases e.g. PCA or dictionary bases ? If such optimal directions exist, what information do they reveal on the structure across channels ?

7.2.3 Towards a mathematical understanding of Transformers

Transformer models have been widely used in the recent years for a number of tasks, including prediction on time-series or images [Vaswani, 2017 ; Ranftl, 2021 ; Wen, 2022]. They provide state-of-the-art results when trained on large, if not huge, amounts of data. One of the key distinguishing features of Transformers is the use of attention mechanisms, which efficiently capture the influence of potentially distant past information. The attention layer learns associations between 'keys' and 'queries' mapped to corresponding 'values' using a kernel-defined mechanism. Multiple attention layers are then cascaded within the Transformer architecture.

Drawing an analogy with Path Shadowing Monte-Carlo, our method establishes a similar association using generated predefined keys, a query based on past history, and the value to be predicted. This is achieved through a kernel method that incorporates information from distant paths via a multi-scale embedding. Unlike the attention mechanism in Transformers, our Path

Shadowing Monte-Carlo approach only necessitates a single realization of limited size. Can we develop a simplified model of attention layer that remains efficient even with limited data ? Additionally, what role does the depth play in Transformer architectures ?

7.2.4 Typicality of an observed realization

The main limitation of Path Shadowing Monte-Carlo introduced in chapter 6 is that it requires to find predictive paths that are close to the observed past history, thus confronting directly the entropy of the process. This algorithm is inefficient because many scanned paths are discarded.

From another perspective, this raises the question of identifying "typical" paths i.e. which are frequently selected by the algorithm for prediction at a given date. Beyond improving computational efficiency of Path Shadowing Monte-Carlo, we believe that path shadowing could be used to define a notion of typicality with respect to a maximum entropy model.

Annexe A

Appendices for Chapter 2

A.1 Wavelet transform properties

We impose that the wavelet ψ satisfies the following energy conservation law called Littlewood-Paley equality

$$\forall \omega > 0 \quad , \quad \sum_{j=-\infty}^{+\infty} |\widehat{\psi}(2^j \omega)|^2 = 1. \tag{A.1}$$

A Battle-Lemarié wavelet [Battle, 1987 ; Lemarié, 1988] is an example of such wavelet. The wavelet transform is computed up to a largest scale 2^J which is smaller than the signal size N. The signal lower frequencies in $[-2^{-J}\pi, 2^{-J}\pi]$ are captured by a low-pass filter $\varphi_J(t)$ whose Fourier transform is

$$\widehat{\varphi}_J(\omega) = \Big(\sum_{j=J+1}^{+\infty} |\widehat{\psi}(2^j \omega)|^2 \Big)^{1/2}. \tag{A.2}$$

One can verify that it has a unit integral $\int \varphi_J(t)\, \mathrm{d}t = 1$. To simplify notations, we write this low-pass filter as a last scale wavelet $\psi_{J+1} = \varphi_J$, and $Wx(t, J+1) = x \star \psi_{J+1}(t)$. By applying the Parseval formula, we derive from (A.1) that for all x with $\|x\|^2 = \int |x(t)|^2\, dt < \infty$

$$\|Wx\|^2 = \sum_{j=-\infty}^{J+1} \|x \star \psi_j\|^2 = \|x\|^2.$$

The wavelet transform W preserves the norm and is therefore invertible, with a stable inverse.

Properties of signal increments are carried over to wavelet coefficients by observing that wavelet coefficients are obtained by filtering signal increments $\delta_j x(t) = x(t) - x(t - 2^j)$ with a dilated integrable filter :

$$x \star \psi_j(t) = \delta_j x \star \theta_j(t) \quad \text{where} \quad \theta_j(t) = 2^{-j} \theta(2^{-j} t), \tag{A.3}$$

where filter θ is obtained from ψ through $\widehat{\theta}(\omega) = \widehat{\psi}(\omega) \,/\, (1 - e^{-i\omega})$. This is because $1 - e^{i2^j\omega}$ is the Fourier transform of $\delta(t) - \delta(t - 2^j)$, the filter that creates increments. From (A.3) we get that if $\delta_j x(t)$ is stationary then $x \star \psi_j(t)$ is also stationary.

A.2 Proof that strong distribution self-similarity implies weak moment self-similarity.

Let x be a stationary process that is self-similar according to (2.1). For $q \in \mathbb{R}$, marginal moments are written $S(q,j) = \mathbb{E}\{|\delta_j x(t)|^q\}$. They do not depend upon t. For all $\ell \geq 0, j \leq J$, self-similarity implies that marginal distributions are equal : $\delta_j x(2^\ell t) \stackrel{d}{=} A_\ell \, \delta_{j-\ell} x(t)$. On order q moments, since A_ℓ is independent from x this yields

$$S(q,j) = \mathbb{E}\{A_\ell^q\} S(q, j-\ell).$$

Since the factors $(A_\ell)_\ell$ are log-infinitely divisible, for all $\ell_1, \ell_2 \geq 0$ $\mathbb{E}\{A^q_{\ell_1+\ell_2}\} = \mathbb{E}\{A^q_{\ell_1}\}\mathbb{E}\{A^q_{\ell_2}\}$. This implies that $\log \mathbb{E}\{A_\ell^q\}$ is linear in ℓ which means there exists ζ_q such that $\mathbb{E}\{A_\ell^q\} = 2^{j\zeta_q}$. By defining $\widetilde{S}(q,j) = 2^{-j\zeta_q} S(q,j)$, we obtain $\widetilde{S}(q,j) = \widetilde{S}(q, j-\ell)$ for all $j \leq J, \ell \geq 0$, which implies that $\widetilde{S}(q,j)$ is equal to a constant $\widetilde{c}_q$ which does not depend on j

$$S(q,j) = \mathbb{E}\{|\delta_j x(t)|^q\} = \widetilde{c}_q 2^{j\zeta_q}. \tag{A.4}$$

Let us now establish the same property for wavelet coefficients. According to the self-similarity property (2.1), wavelet coefficients satisfy (2.8). Indeed, one has :

$$\begin{aligned} x \star \psi_j(2^\ell t) &= \delta_j x \star \theta_j(2^\ell t) \;\; \text{thanks to} \;\; (A.3) \\ &= \delta_j x(2^\ell \cdot) \star \theta_{j-\ell}(t) \;\; \text{as } \theta_j \text{ are dilated filters} \\ &\stackrel{d}{=} A_\ell \, \delta_{j-\ell} x \star \theta_{j-\ell}(t) \;\; \text{by self-similarity} \;\; (2.1) \\ &= A_\ell \, x \star \psi_{j-\ell}(t) \;\; \text{thanks to} \;\; (A.3) \end{aligned}$$

Increments $\delta_j x(t)$ are a special case where $\psi_j = \delta(t) - \delta(t-2^j)$. For general wavelets ψ_j the same proof than for increments holds, there exists c_q such that for the same ζ_q :

$$\mathbb{E}\{|x \star \psi_j(t)|^q\} = c_q 2^{j\zeta_q}.$$

A.3 Proof that strong distribution self-similarity implies wide-sense self-similarity

Let x be a process with stationary increments that is self-similar and thus satisfies (2.8). For $q = 1$ and $q = 2$, A.2 proves that $\mathbb{E}\{|x \star \psi_j(t)|\} = c_1 \, 2^{j\zeta_1}$ and

$$\sigma_W^2(j) = \mathbb{E}\{|x \star \psi_j(t)|^2\} = c_2 \, 2^{j\zeta_2}, \tag{A.5}$$

which proves (2.15) and (2.16).
The equality in distribution (2.8) implies that for τ, j, a fixed we have

$$\Big(x \star \psi_j(t),\, x \star \psi_{j-a}(t - 2^j\tau)\Big) \stackrel{d}{=} A_\ell \Big(x \star \psi_{j-\ell}(2^{-\ell}t),\, x \star \psi_{j-a}(2^{-\ell}t - 2^{j-\ell}\tau)\Big) \tag{A.6}$$

Applying ρ and taking expected value gives

$$\mathbb{E}\{\rho W x(t,j)\,\rho W x(t-2^j\tau, j-a)\} = 2^{\ell\zeta_2}\mathbb{E}\{\rho W x(t,j-\ell)\,\rho W x(t-2^{j-\ell}\tau, j-a-\ell)\}, \quad \text{(A.7)}$$

because of stationarity and $\mathbb{E}\{A_\ell^2\} = 2^{\ell\zeta_2}$. Normalized correlations $C_{\rho W}(\tau; j, a)$ are obtained by dividing by $\sigma_W(j)\sigma_W(j-a)$. It results from (A.5) that

$$2^{\ell\zeta_2}\sigma_W(j)^{-1}\sigma_W(j-a)^{-1} = \sigma_W(j-\ell)^{-1}\sigma_W(j-a-\ell)^{-1}$$

and hence

$$C_{\rho W}(\tau; j, a) = C_{\rho W}(\tau, j-\ell, a).$$

Taking $j = \ell$ proves (2.17).

A.4 Proof of proposition 3 and theorem 2

Let x be a Gaussian process with stationary increments and assume that $\widehat{\psi}_j\widehat{\psi}_{j-a} = 0$. Then for any τ, $x \star \psi_j(t)$ and $x \star \psi_{j-a}(t-\tau)$ are decorrelated because their power spectra do not overlap. Since x is Gaussian these are also Gaussian. It implies that the processes $x \star \psi_j(t)$ and $x \star \psi_{j-a}(t)$ are independent. In particular, $|x \star \psi_j| \star \psi_{j-b}(t)$ and $|x \star \psi_{j-a}| \star \psi_{j-b}(t)$ are independent. Their correlation is thus zero and $C_S(j, a, b) = 0$.

Let x be a time-reversible process with stationary increments. Thanks to proposition 2 we know that $C_{\rho W}$ has the Hermitian symmetry : $C_{\rho W}(\tau; j, a) = C_{\rho W}(-\tau; j, a)^*$. Let us write $C_{|W|}$ the subblock of this matrix composed of the normalized modulus auto-correlation

$$C_{|W|}(\tau; j, a) = \frac{\mathbb{E}\{|x \star \psi_j(t)|\,|x \star \psi_{j-a}(t-2^j\tau)|\}}{\sigma_W(j)\sigma_W(j-a)}$$

We derive from (2.20) that the scattering cross-spectrum C_S satisfies :

$$C_S(j, a, b) = \int_\tau C_{|W|}(\cdot\,; j, a) \star \psi_{-b}(\tau)\ \psi_{-b}(\tau)^* \mathrm{d}\tau. \quad \text{(A.8)}$$

With $Rx(t) = x(-t)$, the Hermitian symmetry of $C_{|W|}$ implies that

$$\begin{aligned}
C_S(j, a, b) &= \int_\tau RC_{|W|}(\cdot\,; j, a)^* \star \psi_{-b}(\tau)\ \psi_{-b}(\tau)^* \mathrm{d}\tau \\
&= \int_\tau C_{|W|}(\cdot\,; j, a)^* \star R\psi_{-b}(-\tau)\ \psi_{-b}(\tau)^* \mathrm{d}\tau \\
&= \int_\tau C_{|W|}(\cdot\,; j, a)^* \star \psi_{-b}^*(-\tau)\ \psi_{-b}(\tau)^* \mathrm{d}\tau \\
&= \int_\tau C_{|W|}(\cdot\,; j, a)^* \star \psi_{-b}^*(-\tau)\ \psi_{-b}(-\tau) \mathrm{d}\tau \\
&= C_S^*(j, a, b)
\end{aligned}$$

because $R\psi_{-b} = \psi_{-b}^*$. It proves that $\operatorname{Im} C_S(j, a, b) = 0$ which proves proposition 3.

Under scale invariance (2.17) $C_{\rho W}(\tau; j, a) = C_{\rho W}(\tau; 0, a)$. As a subblock, one has also $C_{|W|}(\tau; j, a) = C_{|W|}(\tau; 0, a)$. In particular, the equation (A.8) that expresses C_S from $C_{|W|}$ implies $C_S(j, a, b) = C_S(0, a, b)$ which proves theorem 2.

A.5 Financial data preprocessing

We use a standard preprocessing on S&P data which accounts for missing values, overnight period, intraday seasonality and tick effect. It is performed on 5min increments of S&P from January 3rd 2000 to October 10th 2018 which represents 751 116 values.

Missing values, 17 956 5min increments, are replaced by independent Gaussian values with zero mean and standard deviation observed at this time of the day.

Intraday seasonality is the fact that the volatility is larger at certain typical hours of the day : it is a non-stationary effect. It is removed by dividing 5min increments by the average volatility profile over all days.

The overnight period corresponds to the first bin at the beginning of each day. The corresponding increments are generally larger than other 5min increments. That again creates a non-stationarity effect that can be attenuated by dividing each overnight increment by their average volatility on all days.

Prices of S&P are present on a grid with certain tick size. Hence, 5min increments are discrete with many values equal to 0 or to plus/minus the tick size. This tick effect is present only for high-frequency increments and hence breaks the scale invariance property at high frequency. To remove it we apply a low-pass filter to x that amounts to a moving average on small windows of 15 minutes.

A.6 Microcanonical sampling

Given n observed samples $\widetilde{x}_1, \dots, \widetilde{x}_n \in \mathbb{R}^N$ of a process, e.g. time-series $N = T$, multi-channel time-series $N = C \times T$, d-dimensional field $N = L^d$, with possibly $n = 1$, a microcanonical set is defined as follows :

$$\Omega_\epsilon = \Big\{ x_1, \dots, x_m \in \mathbb{R}^N \mid \| \left\langle \Phi(x_j) \right\rangle_j - \left\langle \Phi(\widetilde{x}_i) \right\rangle_i \|_2 < \epsilon \Big\}. \tag{A.9}$$

Where $x_1, \dots, x_m$ are multiple realizations considered simultaneously, enabling parallel generation. Microcanonical models are maximum entropy distributions over Ω_ϵ. Due to the average over j in $\left\langle \Phi(x_j) \right\rangle_j$ the $x_1, \dots, x_m$ share the same distribution. When choosing Φ to be the Scattering Spectra, the set Ω_ϵ is compact, thus a microcanonical model has a uniform distribution over this set. Increasing the number of samples n, depending on data availability, reduces the variance of $\left\langle \Phi(\widetilde{x}_i) \right\rangle_i$ which concentrates around $\mathbb{E}\{\Phi(x)\}$. This reduces the information about a specific realization which is contained in $\left\langle \Phi(\widetilde{x}_i) \right\rangle_i$, thus limiting over-fitting.

Sampling from the microcanonical model amounts to drawing a realization from a uniform distribution in Ω_ϵ. We approximate this sampling with a gradient descent algorithm studied in [Bruna, 2019]. This algorithm progressively transports a white Gaussian noise distribution,

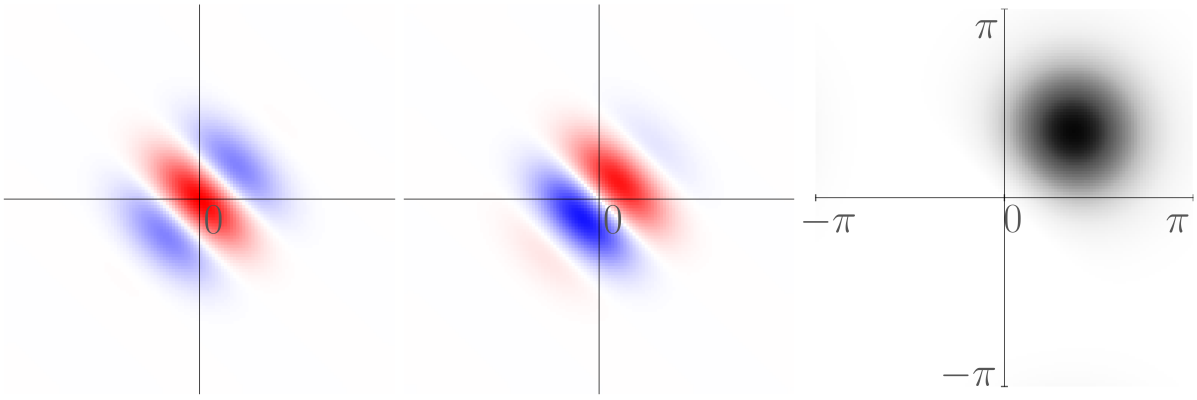
which has a higher entropy than the microcanonical model, into distributions supported in Ω_ϵ. This is done with a gradient descent on $\ell(y_1, \dots, y_m) = \| \langle \Phi(y_j) \rangle_j - \langle \Phi(\widetilde{x}_i) \rangle_i \|^2$, where the y_j are initialized as independent realizations of white noises. At each iteration, the y_i are updated with the L-BFGS-B algorithm, which is a quasi-Newton method that uses an estimate of the Hessian matrix. In practice, we perform 200 gradient descent steps which yield a typical error $\epsilon \approx 10^{-4}$.

It is proved in [Bruna, 2019] that this algorithm converges to a distribution that has the same symmetries as $\Phi(x)$, similarly to the microcanonical one. However, it has been shown that this algorithm recovers a maximum entropy distribution in Ω_ϵ only under appropriate conditions and that such gradient descent models may differ, in general, from maximum entropy ones. Nevertheless, these algorithms provide powerful sampling methods to approximate large classes of high-dimensional stationary processes, while being much faster and computationally tractable than alternative MCMC algorithms.

Annexe B

Appendices for Chapter 3

B.1 Wavelets in $\mathbb{R}^d$ and scattering covariances



(a) Real ψ (b) Imag ψ (c) $\widehat{\psi}$

FIGURE B.1 – Real and imaginary parts of a Morlet wavelet $\psi(u)$ and its Fourier transform $\widehat{\psi}(\omega)$, used in numerical calculations.

A Morlet wavelet ψ defined on $\mathbb{R}^d$ is the product of a Gaussian envelope with a sinusoidal wave

$$\psi(u) = g_\sigma(u)(e^{i\langle \xi, u\rangle} - c) \quad \text{with} \quad g_\sigma(u) = \frac{1}{(\sigma\sqrt{2\pi})^d} e^{-\frac{\|u\|^2}{2\sigma^2}},$$

where c is chosen so that $\int \psi(u)\,\mathrm{d}u = 0$ [1]. Such a wavelet recovers variations around scale 2^j in the direction of ξ. It is invariant to any rotation of $\mathbb{R}^d$ that fixes ξ. In practice we choose $\xi = (3\pi/4, 0, \dots, 0)$ and $\sigma = 0.8$. For simplifying equations in appendices we assume $\|\xi\| = 1$, without loss of generality. To recover variations at other scales and in other directions we define the wavelet filters

$$\psi_\lambda(u) = 2^{-jd}\psi(2^{-j}r^{-1}u) \quad \text{with} \quad \lambda = 2^{-j}r^{-1}\xi$$

1. In practice, the envelope g_σ is an elliptical Gaussian window to increase the angular resolution of ψ, but this does not virtually modify our discussion.

for $(j, r) \in \mathbb{R} \times SO(d)$. In Fourier, $\widehat{\psi}_\lambda$ is a Gaussian centered in λ substracted by a Gaussian centered in 0 so that $\widehat{\psi}_\lambda(0) = 0$

$$\widehat{\psi}_\lambda(\omega) = \widehat{\psi}(2^{-j} r^{-1}\omega) \quad \text{with} \quad \widehat{\psi}(\omega) = e^{-\frac{\sigma^2}{2}\|\omega-\xi\|^2} - c\, e^{-\frac{\sigma^2}{2}\|\omega\|^2}$$

We shall restrict the scales 2^j to dyadic scales, hence taking j integer, and restrict the rotations to a discrete subgroup Γ of $SO(d)$ of order $2^d - 1$ [Y Meyer, 1992]. In dimension $d = 2$ such a group can be parameterized by one angle, in dimension $d = 3$ it can be parameterized by 2 angles. We write $\Lambda = \mathbb{Z} \times \Gamma$ the group that defines filters ψ_λ from ψ.

To guarantee that the wavelet transform W (defined in (3.7)) is invertible and satisfies an energy conservation, we impose that the ψ_λ satisfy the following Littlewood-Paley inequality for $0 < \delta < 1$

$$\forall \omega \neq 0 \quad , \quad 1 - \delta \leq \sum_{\lambda \in \Lambda} |\widehat{\psi}_\lambda(\omega)|^2 \leq 1 + \delta.$$

For fields defined on a cubic d-dimensional lattice of length L, the wavelets ψ_λ are discretized accordingly. The wavelet transform is computed up to the largest scale 2^J which is smaller than length L so as to achieve a reasonable estimate of low-frequency moments, even on a single realization. The lower frequencies of x in the ball $|\omega| \leq 2^J$ are captured by a low-pass filter ψ_0 which is a Gaussian centered in $\omega = 0$ in Fourier $\widehat{\psi}_0(\omega) = c_0 \exp(-\sigma_0^2\|\omega\|^2/2)$ with $\sigma_0 = \sigma\, 2^{J-1}$. The Littlewood-Paley inequality now reads :

$$\forall \omega \neq 0 \quad , \quad 1 - \delta \leq |\psi_0(\omega)|^2 + \sum_{|\lambda|^{-1} \leq 2^J} |\widehat{\psi}_\lambda(\omega)|^2 \leq 1 + \delta.$$

By applying the Parseval formula we derive that for all x

$$(1 - \delta)\|x\|^2 \leq \|Wx\|^2 \leq (1 + \delta)\|x\|^2$$

which insures that W preserves the norm of x, up to a relative error of δ, and is therefore invertible, with stable inverse. For the wavelet used for syntheses of physical fields in chapter 3, we have $\delta \approx 0.8$.

Covariance of wavelet coefficients $Wx(u, \lambda)$ can be written from the power spectrum $P(\omega)$ of x

$$\mathbb{E}\{Wx(u, \lambda) Wx(u', \lambda')^*\} = \frac{1}{2\pi} \int P(\omega) \widehat{\psi}_\lambda(\omega) \widehat{\psi}_{\lambda'}(\omega) e^{i\langle u - u', \omega\rangle} \mathrm{d}\omega.$$

It implies that this correlation is zero if the supports of $\widehat{\psi}_\lambda$ and $\widehat{\psi}_{\lambda'}$ do not overlap. For the specified wavelets, as soon as $\lambda \neq \lambda'$, these supports barely overlap and $\mathbb{E}\{Wx(u, \lambda)\, Wx(u', \lambda')^*\} \approx 0$. Moreover, since x is stationary, the covariance $\mathbb{E}\{Wx(u, \lambda)\, Wx(u', \lambda')^*\}$ only depends on $u - u'$ and have a fast decay when the power spectrum $P(\omega)$ is regular. Thus, even if dependencies across separate scales may exist, they are not captured by correlation.

Taking the modulus of wavelet coefficients removes complex phase oscillations and thus recenter the frequency support of $Wx(u, \lambda)$. Indeed, the power spectrum $P_\lambda(\omega)$ of $x \star \psi_\lambda$ is mostly supported in a ball $\|\omega - \lambda\| \leq 2^{-j}\sigma^{-1}$ which does not overlap with the Fourier support

of the power spectrum $P_{\lambda'}(\omega)$ of $x \star \psi_{\lambda'}$. Taking a modulus on $x \star \psi_{\lambda'}$ eliminates the phase which oscillates at the central frequency λ'. As a consequence, the power spectrum of $|x \star \psi_{\lambda'}|$ is centered at $\omega = 0$ and its energy is mostly concentrated in $\|\omega\| \leq 2^{-j}\sigma^{-1}$ which now may overlap with the support of $P_\lambda(\omega)$ as can be seen in Fig. 3.2. The power spectra of $|Wx(u,\lambda)|$ and $|Wx(u,\lambda')|$, both centered at zero, also overlap.

We now justify taking $u = u'$ in order 3 moments given by (3.9). The cross spectrum $P_{\lambda,\lambda'}$ between $Wx(u,\lambda)$ and $|Wx(u,\lambda')|$ is assumed regular for the fields considered in chapter 3. In that case one can approximate such cross-spectrum using wavelets, which gives the moments $\mathbb{E}\{WWx(u,\lambda,\gamma)\, W|Wx|(u,\lambda',\gamma)\}$. However, the left-hand-side $WWx(u,\lambda,\gamma)$ is negligible when $\lambda \neq \gamma$ because Fourier support of wavelets ψ_λ and ψ_γ barely overlap. The resulting coefficients

$$\mathbb{E}\{WWx(u,\lambda,\lambda)\, W|Wx|(u,\lambda',\lambda)\} = \frac{1}{2\pi}\int P_{\lambda,\lambda'}(\omega)|\widehat{\psi}_\lambda|^2 \mathrm{d}\omega$$

average $P_{\lambda,\lambda'}(\omega)$ in a ball $\|\omega\| \leq 2^{-j}\sigma^{-1}$ through $|\widehat{\psi}_\lambda|^2$. However, $P_{\lambda,\lambda'}(\omega)$ is already concentrated in this ball. We thus remove $|\widehat{\psi}_\lambda|^2$ which yields $\mathbb{E}\{Wx(u,\lambda)\,|Wx|(u,\lambda')\}$.

The following proposition shows that Scattering Spectra reveal non-Gaussianity in a field x.

Proposition 4. *Let x be a stationary process.*

1. *If x is Gaussian then for any separate scales λ, λ', meaning that $\widehat{\psi}_\lambda \widehat{\psi}_{\lambda'} = 0$*

$$\mathbb{E}\{\bar{S}_1(x)\} = \frac{\pi}{4},$$

$$\mathbb{E}\{\bar{S}_3(x)[\lambda,\lambda']\} = 0 \quad \textit{and} \quad \mathbb{E}\{\bar{S}_4(x)[\lambda,\lambda',\gamma]\} = 0.$$

2. *If x is symmetric i.e. $p(-x) = p(x)$ then*

$$\mathbb{E}\{\bar{S}_3(x)\} = 0.$$

3. *If x is invariant by rotation of angle π i.e. $p(x(-u)) = p(x(u))$ then*

$$\operatorname{Im}\mathbb{E}\{\bar{S}_3(x)\} = 0 \quad \textit{and} \quad \operatorname{Im}\mathbb{E}\{\bar{S}_4(x)\} = 0.$$

Démonstration. If x is Gaussian then $Wx(u,\lambda)$ is also Gaussian and the ratio between its first and second order moment is $\pi/4$. If $\widehat{\psi}_\lambda \widehat{\psi}_{\lambda'} = 0$ then $Wx(u,\lambda)$ and $Wx(u,\lambda')$ are decorrelated, since $(Wx(u,\lambda), Wx(u,\lambda'))$ is Gaussian, this implies that $Wx(u,\lambda)$ and $Wx(u,\lambda')$ are independent. Thus, $Wx(u,\lambda)$ and $|Wx(u,\lambda')|$ are independent, so are $W|Wx|(u,\lambda,\gamma)$ and $W|Wx|(u,\lambda',\gamma)$ which proves 1. Point 2. is proved by observing that $S_3(-x) = S_3(x)$ and point 3. by observing that $S_3(x(-u)) = S_3(x)^*$ and $S_4(x(-u)) = S_4(x)^*$. □

For the physical fields studied in chapter 3, such coefficients are non-zero, thus revealing their non-Gaussianity Fig.B.3.

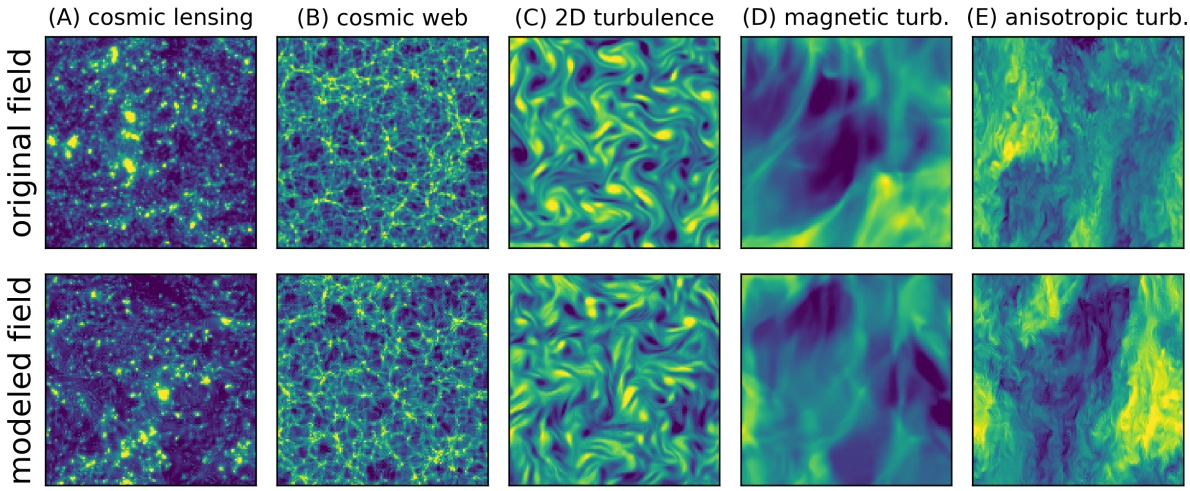


FIGURE B.2 – Visual assessment of our model based on $\bar{S}$ with 11 641 coefficients estimated on a single realization (top). Generated fields (bottom) show very good visual quality.

B.2 Equivariance and invariance to rotations and scaling

The Scattering Spectra are computed from wavelet transforms, which are equivariant to rotations and scalings. We show that Scattering Spectra inherit these equivariance properties. If $p(x)$ is isotropic or self-similar, then one can build isotropic or self-similar maximum entropy models by averaging renormalized Scattering Spectra over rotations or scales, which reduces both the variance and dimensionality of $\bar{S}$.

To avoid discretization and boundary issues for rotations and scaling, we consider fields $x(u)$ defined over continuous variables $u \in \mathbb{R}^d$, and establish the mathematical results in this framework. For this purpose, the sum in the wavelet transform defined in (3.7) is replaced by an integral over $\mathbb{R}^d$. Wavelets are dilated by 2^j for $j \in \mathbb{Z}$ and rotated by r in a rotation group G of cardinal R. In dimension $d = 2$, these rotations have an angle $2\pi\ell/R$.

Proposition 5. *For $r \in G$ with $x_r(u) = x(r^{-1}u)$ one has*

$$S(x_r)[\lambda, \lambda', \gamma] = S(x)[r\lambda, r\lambda', r\gamma]. \tag{B.1}$$

For $j \in \mathbb{Z}$ with $x_j(u) = x(2^{-j}u)$ one has

$$S(x_j)[\lambda, \lambda', \gamma] = S(x)[2^j\lambda, 2^j\lambda', 2^j\gamma]. \tag{B.2}$$

Proof. It follows from the equivariance of wavelet coefficients, $Wx_r(u, \lambda) = Wx(r^{-1}u, r\lambda)$ and $Wx_j(u, \lambda) = Wx(2^{-j}u, 2^j\lambda)$.

Isotropic fields x have a distribution that is invariant to rotation $x_r \overset{d}{=} x$ for all $r \in G$. Self-similar fields x have a distribution that is invariant to scaling, up to random multiplicative factors $x_j \overset{d}{=} A_j x$ for all $j \geq 0$ [Mandelbrot, 1997]. For such fields, we show that the expected Scattering Spectra exhibit invariance to rotation or scaling of their indices, and thus have a lower-dimensional structure. For that purpose we used normalized Scattering Spectra coefficient

$\bar{S}(x)$ defined (3.16), where the normalization is done by $\sigma^2[\lambda] = \mathbb{E}\{|Wx(u,\lambda)|^2\}$.

Proposition 6. *If x is isotropic then for any $r \in G$*

$$\mathbb{E}\{\bar{S}(x)[r\lambda, r\lambda', r\gamma]\} = \mathbb{E}\{\bar{S}(x)[\lambda, \lambda', \gamma]\}. \tag{B.3}$$

If x is self-similar at scales $2^j \leq 2^J$ then

$$\mathbb{E}\{S_1(x)[\lambda]\} = c_1|\lambda|^{-\zeta_1} \quad , \quad \mathbb{E}\{S_2(x)[\lambda]\} = c_2|\lambda|^{-\zeta_2} \tag{B.4}$$

$$\mathbb{E}\{\bar{S}_3(x)[2^j\lambda, 2^j\lambda']\} = \mathbb{E}\{\bar{S}_3(x)[\lambda, \lambda']\} \tag{B.5}$$

$$\mathbb{E}\{\bar{S}_4(x)[2^j\lambda, 2^j\lambda', 2^j\gamma]\} = \mathbb{E}\{\bar{S}_4(x)[\lambda, \lambda', \gamma]\} \tag{B.6}$$

Proof. Let us assume x is isotropic $x_r \overset{d}{=} x$. It implies that $\mathbb{E}\{S(x_r)\} = \mathbb{E}\{S(x)\}$. Thanks to the equivariance property of (B.1) one gets the invariance property on S : $\mathbb{E}\{S(x)[r\lambda, r\lambda', r\gamma]\} = \mathbb{E}\{S(x)[\lambda, \lambda', \gamma]\}$. We obtain (B.3) by dividing this equation by $\mathbb{E}\{S_2(x)[r\lambda]\} = \mathbb{E}\{S_2(x)[\lambda]\}$. Let us assume x is self-similar, $x_j \overset{d}{=} A_j x$. In that case one has $A_{j+j'} \overset{d}{=} A_j A_{j'}$, taking order 1 and order 2 moments, this implies $\mathbb{E}\{A_j\} = 2^{-j\zeta_1}$ and $\mathbb{E}\{A_j^2\} = 2^{-j\zeta_2}$ for certain power-law exponents ζ_1, ζ_2. Now from self-similarity and equivariance property given by (B.2) one has $\mathbb{E}\{S_1(x)[2^j\lambda]\} = \mathbb{E}\{A_j\}\mathbb{E}\{S_1(x)[\lambda]\} = 2^{-j\zeta_1}\mathbb{E}\{S_1(x)[\lambda]\}$. Taking $2^{-j} = |\lambda|$ one obtains $\mathbb{E}\{S_1(x)[\lambda]\} = c_1|\lambda|^{-\zeta_1}$ with $c_1 = \mathbb{E}\{S_1(x)[|\lambda|^{-1}\lambda]\}$ independent on $|\lambda|$. With the same reasoning on S_2 we obtain (B.4). From self-similarity and equivariance property given (B.2), we get similarly : $\mathbb{E}\{S_3(x)[2^j\lambda, 2^j\lambda']\} = 2^{-j\zeta_2}\mathbb{E}\{S_3(x)[\lambda, \lambda']\}$. Dividing by $\mathbb{E}\{S_2(x)[\lambda]\} = c_2|\lambda|^{-\zeta_2}$ yields (B.5). We obtain (B.6) similarly, which proves the proposition.

The wavelet coefficient renormalization is necessary to ensure that the Scattering Spectra are invariant to scaling. As explained in [Marchand, 2022], it is directly related to Wilson renormalization, which yields macrocanonical parameters (physical couplings) that remain constant across scales (fixed point) at phase transitions, where the field becomes self-similar.

If x is isotropic, then (B.3) implies that

$$\left\langle \bar{S}(x)[r\lambda, r\lambda', r\gamma] \right\rangle_{r \in G} \tag{B.7}$$

is an unbiased estimator of $\mathbb{E}\{\bar{S}(x)[\lambda, \lambda', \gamma]\}$ with lower variance than $\bar{S}(x)[\lambda, \lambda', \gamma]$. Choosing $\Phi(x) = \left\langle \bar{S}(x)[r\lambda, r\lambda', r\gamma] \right\rangle_{r \in G}$ also reduces the dimension of our model by a factor R. Since this representation is invariant to rotations of x in G, the macrocanonical and microcanonical models defined from it are also invariant to these rotations.

Similarly, if x is self similar on a range of scales $2^j \leq 2^J$, then (B.5) and (B.6) implies that

$$\left\langle \bar{S}_3(x)[2^j\lambda, 2^j\lambda'] \right\rangle_j \quad , \quad \left\langle \bar{S}_4(x)[2^j\lambda, 2^j\lambda', 2^j\gamma] \right\rangle_j \tag{B.8}$$

where the average is taken on all scales j such that $(2^j|\lambda|)^{-1} \leq 2^J, (2^j|\lambda'|)^{-1} \leq 2^J, (2^j|\gamma|)^{-1} \leq 2^J$, are unbiased estimators of $\mathbb{E}\{\bar{S}_3(x)[\lambda, \lambda']\}$ and $\mathbb{E}\{\bar{S}_4(x)[\lambda, \lambda', \gamma]\}$ with lower variance than

$\bar{S}_3(x)$ and $\bar{S}_4(x)$. Choosing $\Phi(x) = (\bar{S}_1(x), \bar{S}_2(x), \left\langle \bar{S}_3(x) \right\rangle_j, \left\langle \bar{S}_4(x) \right\rangle_j)$ reduces the dimension of our model by at most a factor $\log L$. The resulting maximum entropy model is not necessarily self-similar due to the presence of scale-dependent moments $\mathbb{E}\{\bar{S}_1(x)\}$ and $\mathbb{E}\{\bar{S}_2(x)\}$. However, if $\bar{S}_1(x)[\lambda]$ and $\bar{S}_2(x)[\lambda]$ have a power-law decay along λ our model becomes self-similar.

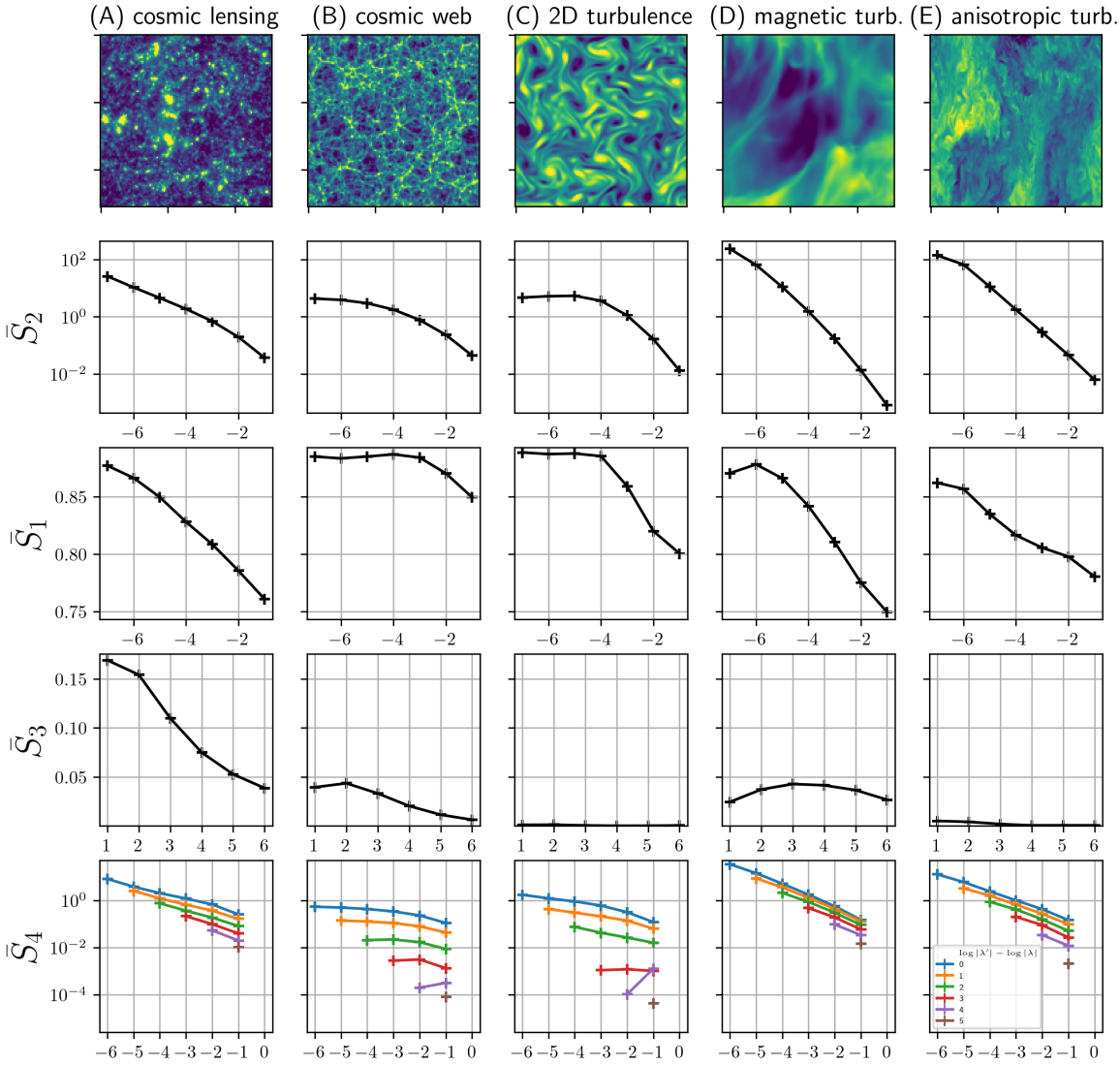


FIGURE B.3 – Visualization of Scattering Spectra $\bar{S}$ for different physical fields. Power-spectrum S_2 and sparsity factors $\bar{S}_1$ are averaged along all angles (amount to taking the 0-th angle Fourier harmonic). We only show the 0-th angle Fourier harmonic and 0-th scale Fourier harmonic for order 3 and order 4-moment estimators $\bar{S}_3$ and $\bar{S}_4$. Thus, the quantities that are shown are invariant to the rotation of the field, and the last two rows ($\bar{S}_3, \bar{S}_4$ are furthermore invariant to scaling). Non-zero coefficients $\bar{S}_3$ show that the cosmic lensing and cosmic web fields are not invariant to sign flip. This is due to the presence of high positive peaks on the former and filaments on the latter. The large amplitude of envelope coefficients $\bar{S}_4$ on the last 2 fields indicate long-range spatial dependencies as evidenced by the presence of structures at the level of the map.

B.3 Dimension reduction with Fourier thresholding

We give in this appendix the details of the dimensional reduction of $\bar{S}$ into $P\bar{S}$, which is done by Fourier projectors of $\bar{S}(x)$ along rotations and scales, estimated by thresholding. This dimensional reduction based on regular variations of the dependence of $\bar{S}$ on different scales, allows for a representation of lower variance, bringing the microcanonical and macrocanonical models closer together.

We concentrate on the two-dimensional case $d = 2$ corresponding to numerical applications. The rotation group is then Abelian and defined by a single angle parameter, which simplifies the Fourier transform calculation. However, the same approach applies to non-commutative groups G of rotations in $\mathbb{R}^d$ for $d > 2$, with their Fourier transform. Each wavelet frequency is defined in (3.6) by $\lambda = 2^{-j} r_\ell \xi$, where r_ℓ is a rotation of angle $2\pi\ell/R$. To guaranty that the Scattering Spectra frequencies satisfy $|\lambda| \leq |\lambda'| < |\gamma|$, we write

$$\lambda = 2^{-j_1} r_{\ell_1} \xi \ , \ \lambda' = 2^{-j_1-a} r_{\ell_2} \xi \ , \ \gamma = 2^{-j_1-b} r_{\ell_3} \xi$$

with $0 \leq a < b \leq J - j_1$ and $J < \log L$. It leads to a scale and angle reparametrization of the Scattering Spectra :

$$\bar{S}(x)[\lambda, \lambda', \gamma] = \bar{S}(x)[j_1, a, b, \ell_1, \ell_2, \ell_3].$$

If $\bar{S}(x)$ has regular variations as a function of rotations then its three-dimensional Fourier transform along the (ℓ_1, ℓ_2, ℓ_3) has coefficients of negligible amplitude at high frequencies, which can thus be eliminated. One can also take advantage of regularities along scales. Since $1 \leq j_1 \leq J$ varies on an interval without periodicity, the Fourier transform is replaced by a cosine transform along j_1 for a and b fixed. We could also perform a cosine transform along the scale shift a and b, but this is not done in numerical applications because their range of variations is small and j-dependent. The Fourier transforms along j_1 is however sufficient to identify scale-invariance, since one then expects $\bar{S}$ to only depend on a and b, see appendix B.2. We write $F\,\bar{S}(x)$ the Fourier transform of $\bar{S}(x)$ along (ℓ_1, ℓ_2, ℓ_3) and its cosine transform along j_1.

Since F is unitary, it preserves the estimator variance :

$$\sigma^2_{\bar{S}} = \mathbb{E}\{\| \left\langle F\bar{S}(x_i) \right\rangle_i - \mathbb{E}\{F\bar{S}(x)\}\|^2\}. \tag{B.9}$$

Ideally, the estimation error of $\mathbb{E}\{F\bar{S}(x)\}$ is reduced by eliminating its coefficients whose squared amplitude is smaller than the variance of the empirical estimation error. It amounts to suppressing all coefficients having a variance that is larger than the bias resulting from their elimination. However, we can not implement this optimal "oracle" decision because we do not know $\mathbb{E}\{F\bar{S}(x)\}$. In chapter 3, we instead apply an approximate thresholding algorithm, which eliminates small amplitude coefficients of $\bar{S}(x)$ below a threshold proportional to their standard deviation, as discussed in the main text. This thresholding algorithm is adaptive and the selected coefficients vary from one process to another. For each process studied, an ensemble of between 20 to 100 samples x_i were used to empirically estimate the average and variance of $F\bar{S}$, called $\mu(F\bar{S})$ and $\sigma(F\bar{S})$. The coefficients which have been kept are those that individually

verify $\mu(F\bar{S}) > 2\sigma(F\bar{S})$.

A projected Scattering Spectra

$$\Phi(x) = P\,\bar{S}(x)$$

is computed with a linear Fourier projection P which eliminates all coefficients of $F\bar{S}(x)$ corresponding to coefficients of $\left\langle F\bar{S}(x_i)\right\rangle_i$ below their threshold. The efficiency of this projected scattering is the variance reduction ratio $\sigma^2_{P\bar{S}}/\sigma^2_{\bar{S}}$ with

$$\sigma^2_{P\bar{S}} = \mathbb{E}\{\| \left\langle P\bar{S}(x_i)\right\rangle_i - \mathbb{E}\{P\bar{S}(x)\}\|^2\}. \tag{B.10}$$

If $p(x)$ is isotropic or self-similar then we expect that P is a low-frequency projector along global rotations (which act similarly on all l_i coordinates) or scalings (which act on j), which corresponds to the averages described in (B.7) and (B.8). The Fourier projection P is however much more general and can adapt to unknown regularities of $p(x)$ along rotations and scales.

B.4 Number of coefficients for shell binned polyspectra

For a 2D field, there are originally $O(L^4)$ bispectrum coefficients in total, as there are two independent frequencies in the bispectrum and each has two dimensions. If we take N_{bin} linear frequency bins along each side of L lattice points, the coefficients to be estimated is reduced to $O(N^4_{\text{bin}})$. A rotation and parity average will further reduce and better estimate the bispectrum coefficients, which eliminates one dimension and leads to $\sim \frac{1}{A_3^3}\cdot\frac{1}{2}N_{\text{bin}}\cdot\frac{3}{4}N^2_{\text{bin}}\cdot\frac{1}{2} = \frac{1}{32}N^3_{\text{bin}}$ binned coefficients, where $1/A_3^3 = \frac{1}{6}$ is the repeated counting of the three-frequency combinations in bispectrum, $\frac{1}{2}N_{\text{bin}}$ is the number of choice of k_1, given rotation invariance, $\frac{3}{4}$ is the number of choice of k_2 given the requirement that each k is within the $L\times L$ lattice in Fourier space and $k_1+k_2+k_3=0$, and the factor $\frac{1}{2}$ comes from parity average.

For 2D fields, the shell-binned bispectrum is essentially a fast way to compute the rotation and parity average of the bispectrum. It does not mix very different configurations, because a given set of $|k_1|,|k_2|,|k_3|$ combined with the condition $k_1+k_2+k_3=0$ uniquely set the configuration up to free rotations. The number of coefficients is of the order $\sim \frac{1}{8}N^3_{\text{bin}}$ (the scaling power is 3 rather than $2d=4$ because the orientation average eliminates one degree of freedom). For our choice of $N_{\text{bin}}=10$, there are 151 shell-binned bispectrum coefficients. Similarly, the shell-binned trispectrum $\bar{T}$ has 651 $\sim \frac{1}{16}N^4_{\text{bin}}$ coefficients. Note that the shell-binning for trispectrum is more aggressive, because in 2D the same set of $|k_1|,|k_2|,|k_3|,|k_4|$ may come from different combinations k_1,k_2,k_3,k_4 even if the condition $k_1+k_2+k_3+k_4=0$ is applied.

The ordering of $\bar{B}$ and $\bar{T}$ shown in Fig. 3.4 is determined in a nested way. The frequency annuli are labeled by i from small to large $|k|$. To remove redundant coefficients, we require $i_1\le i_2\le i_3(\le i_4)$ and order them first by i_1 in increasing order ; when two binning configurations have the same i_1, they are then ordered by i_2 and so on.

Annexe C

Appendices for Chapter 5

C.1 Source Separation Guarantees

We prove theorem 1, discuss its assumptions for the deglitching example applied to data from Mars, and show how our implementation relates to these assumptions. For sake of simplicity we take $a_1 = 1$.

Proof. Part I. One can prove that there exists a unique process n that maximises entropy under moment constraint $\mathbb{E}\{\Phi(n)\}$, its distribution takes the form $p_n(\cdot) = Z_\theta^{-1} e^{-\langle \theta, \Phi(\cdot) \rangle}$ for certain Lagrange multipliers $\theta \in \mathbb{R}^M$ where M is the dimension of Φ. Assumptions 1, 2, 3 imply that n and $\bar{n}$ are the same unique process, meaning $p_n = p_{\bar{n}}$.
Part II. Due to the independence of s_1, n and $\bar{s}_1, \bar{n}$ (4) we have $p_x = p_{s_1} \star p_n$ and $p_x = p_{\bar{s}_1} \star p_{\bar{n}}$. Since $p_{\bar{n}} = p_n$ we get $p_{s_1} \star p_n = p_{\bar{s}_1} \star p_{\bar{n}}$. This is a measure deconvolution problem. Taking the Fourier transform on measures yields

$$(\widehat{p}_{s_1} - \widehat{p}_{\bar{s}_1})\, \widehat{p}_n = 0.$$

Under assumption 5 we get $p_{\bar{s}_1} = p_{s_1}$, which proves the theorem. □

Assumption 1 is the main assumption. It implies that the processes n is fully determined by the values $\mathbb{E}\{\Phi(n)\}$, since there is a unique distribution satisfying 1. A maximum entropy process n under correlation constraints $\mathbb{E}\{nn^T\}$ is a Gaussian process. A wavelet Scattering Covariance captures non-linear correlations, assumption 1 tells us that process n is a non-Gaussian noise fully characterized by $\mathbb{E}\{\Phi(n)\}$. Now, the Scattering Covariance $\mathbb{E}\{\Phi(n)\}$ was shown to characterize a wide range of non-Gaussian noises, see chapter 2. In our case, the Mars seismic background noise n may not be fully characterized by its Scattering Covariance $\mathbb{E}\{\Phi(n)\}$, so that assumption 1 is only verified approximately, depending on the descriptive power of the representation $\mathbb{E}\{\Phi(n)\}$ for n.

Assumption 2 is approximately verified, requiring the entropy of x to be close to the entropy of n, which is typically the case of time-localized signals such as glitch, of comparable amplitude

than n. The gradient descent algorithm implements 2, reconstructed $\bar{n}$ is initialized to x and is updated until $\Phi(x)$ matches the $\Phi(n_k)$.

Assumption 3 is imposed through the loss term $\mathcal{L}_{\text{prior}}$, up to estimation error of $\Phi(n)$ on a finite number of realizations.

Assumption 4 relates to the loss term $\mathcal{L}_{\text{cross}}$ that imposes statistical independence up to the cross-Scattering Covariance.

Assumption 5 is a technical assumption satisfied for a Gaussian noise n for which the Fourier transform of p_n is a Gaussian. A non-Gaussian noise n satisfying 1 has a distribution of the form $p_n(\cdot) = Z_\theta^{-1} e^{-\langle \theta, \Phi(\cdot) \rangle}$. Apart from the coefficients $\text{Ave}(S(n))$, the scattering covariance Φ is quadratic in n, thus we may assume 5 is still satisfied.

C.2 Baseline method

The glitch detection algorithm that we use as baseline is developed by Scholz et al. [Scholz, 2020] and consists of several processing steps applied to seismic data :

— Decimation : The data is downsampled to a uniform rate of two samples per second to ensure consistent parameterization and improve computational efficiency ;

— Deconvolution and band-pass filtering : Instrument response is removed from each component, transforming the data into acceleration. Additional band-pass filtering is also applied to highlight the significant features of acceleration ;

— Time derivative calculation : The time derivative of the filtered acceleration data is computed, resulting in acceleration steps becoming impulse-like signals ;

— Glitch detection : A constant threshold is applied to the time derivative to identify glitches. A window length is introduced to avoid false triggers on subsequent samples that are part of the same glitch event, serving as a safeguard against spurious detections.

After glitch detection, removal is based on obtaining a model (template) for the glitch signatures, followed by a separation techniques that assumes the observed data as a linear combination of the glitch and the glitch spike. To characterize each detected glitch, a glitch model is employed, consisting of three parameters : an amplitude scaling factor, an offset, and a linear trend parameter. The modeling process entails solving a nonlinear least squares data fitting problem to determine these parameters. Subsequently, the deglitched data is obtained by subtracting the fitted glitch (excluding the offset and linear trend) from the original data.

In comparison to our approach, the glitch modeling step in the mentioned method could be a significant limitation. Unlike their method, we do not make any assumptions about the functional form of the glitch or the unknown source. Instead, we focus on learning the wavelet scattering covariance statistics of the background noise. This allows us to overcome the potential limitations associated with explicitly modeling the glitches.

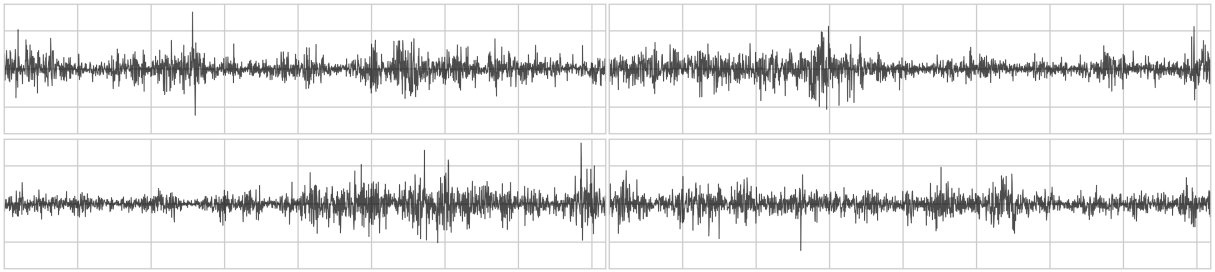


FIGURE C.1 – Realizations of increments of the multifractal random walk process.

C.3 Multifractal random Walk realizations

Figure C.1 shows realizations of the multifractal random walk process used in the stylized example.

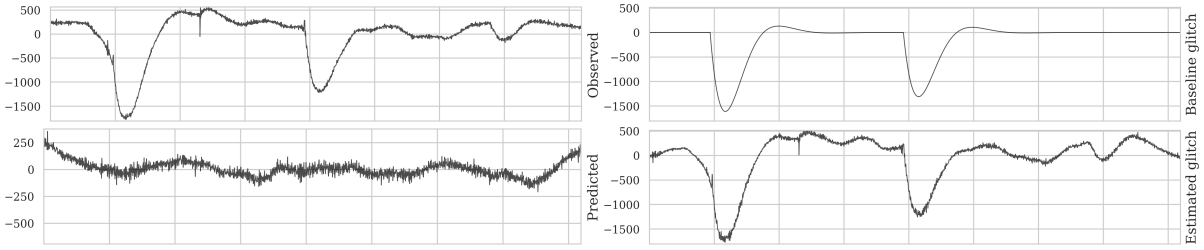


FIGURE C.2 – Unsupervised source separation for glitch removal.

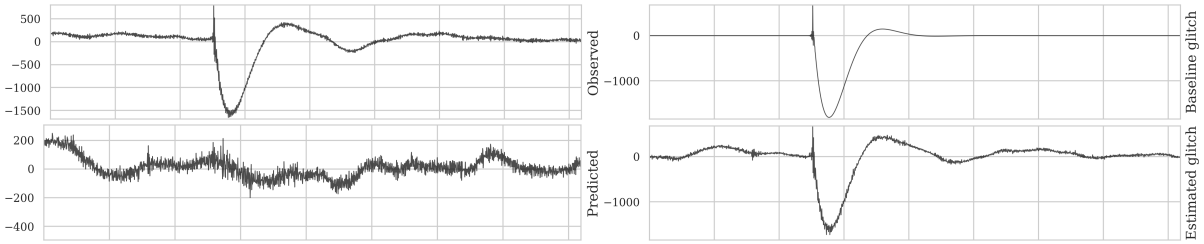


FIGURE C.3 – Unsupervised source separation for glitch removal.

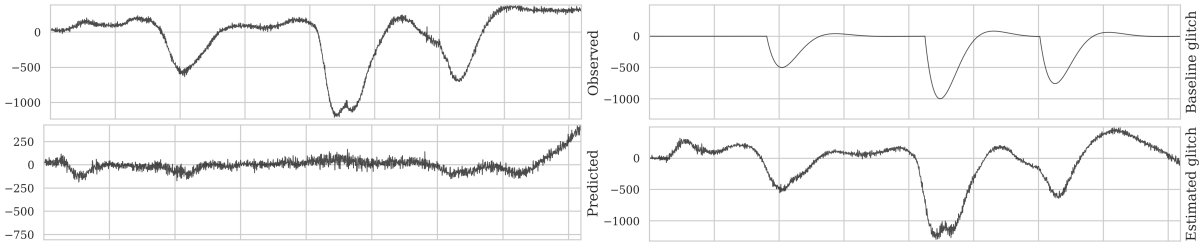


FIGURE C.4 – Unsupervised source separation for glitch removal.

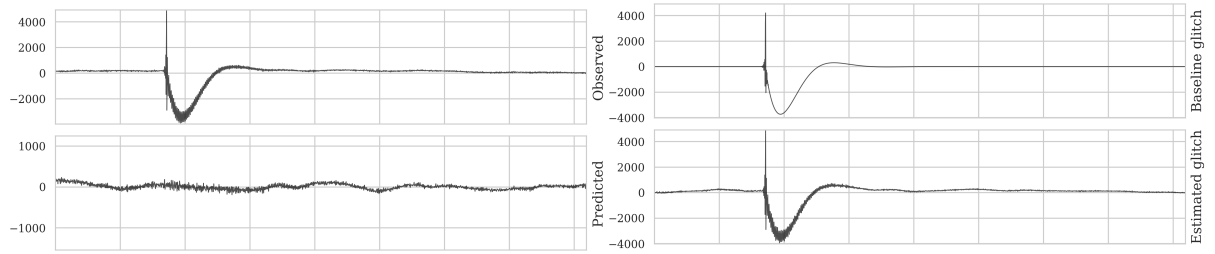


FIGURE C.5 – Unsupervised source separation for glitch removal.

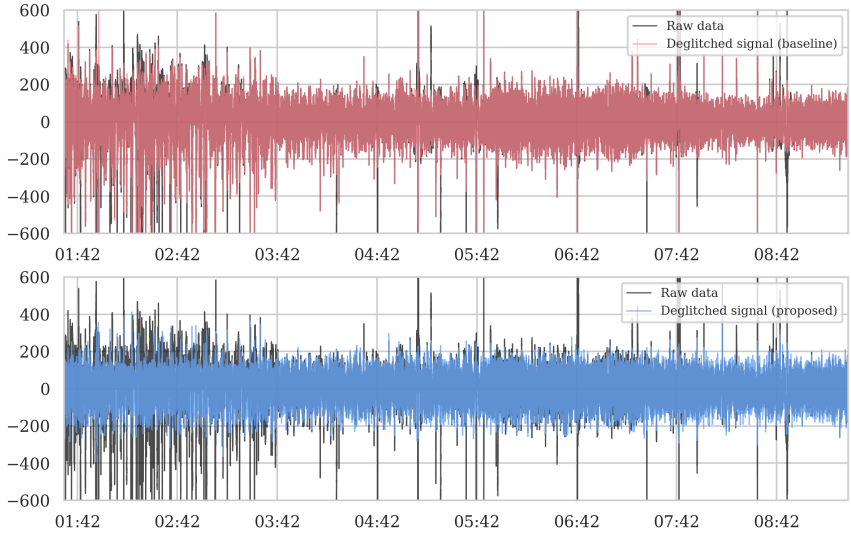


FIGURE C.6 – Unsupervised separation of glitches from seismic data recorded during sol 187 (June 6, 2019) from 17 :08 to 00 :55 Martian local time (the horizontal axis is in UTC time zone). The raw data is depicted in black, with the predicted deglitched data overlaid, represented by the baseline method in red and the proposed method in blue. The high-amplitude "spikes" observed in the raw waveform correspond to glitches. A successful deglitching outcome should exclude these spikes. Our deglitching results effectively separate a significant number of these high-amplitude events, whereas the baseline method fails to address a considerable portion of them.

C.4 Additional glitch separation results

Here we provide more results regarding separating glitches from the seismic data recorded during the NASA InSight mission. Figures C.2–C.5 provide glitch removal results for a more diverse set of glitches using the same setup as described in section 5.6.2.1.

We provide more comprehensive deglitching results by applying our approach to perform glitch separation on the U component for the nighttime (17 :08–00 :55 LMST) during sol 187 (June 6, 2019), as the glitches during the day are often obscured by daytime noise. We used a set of 50 snippets with window size of 204.8 s and solved the source separation optimization problem using 200 L-BFGS iterations.

Our results indicate that the baseline method appears to overlook several anomalies in the U component that we believe to be glitches. In contrast, our method not only detects all the glitches identified by the baseline method, but it also recognizes a significant number of additional

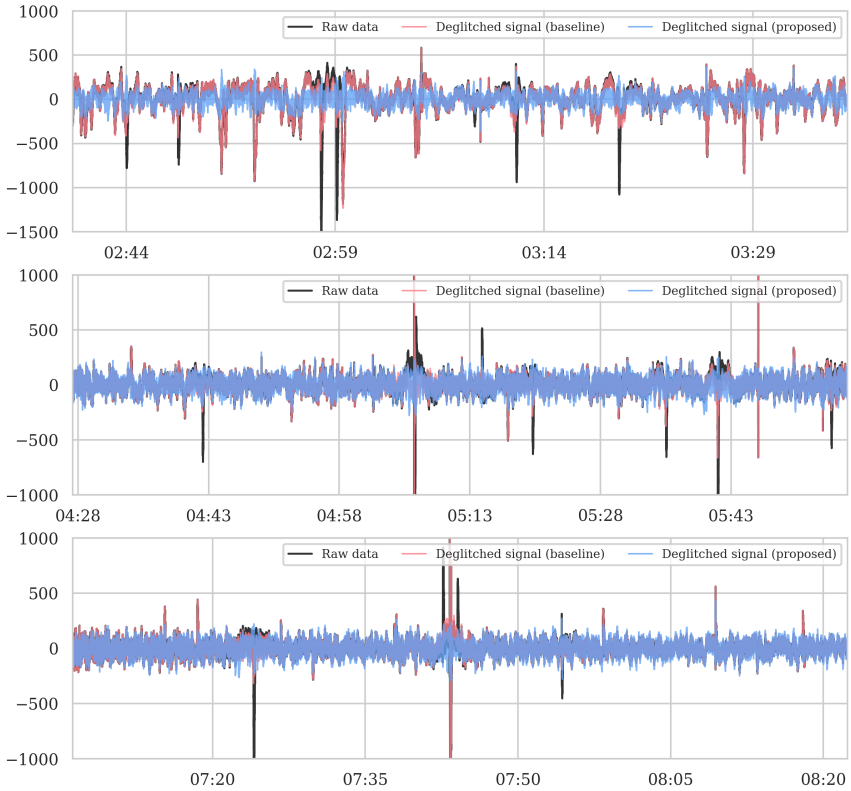


FIGURE C.7 – Three zoomed-in time intervals from Figure C.6 to facilitate a detailed performance comparison between the baseline (red) and the proposed deglitching results. Both outcomes are overlaid on the raw waveform shown in black. The glitches manifest as high-amplitude one-sided pulses in the raw waveform, which we intend to separate. Within each of the aforementioned time intervals, it is evident that the baseline approach falls short in effectively separating several glitches. The horizontal axis represents the UTC time zone.

glitches. Although it is true that our method appears to detect more glitches than the baseline, we must recognize that the baseline is the only dependable reference for identifying glitches and further verification by InSight experts is necessary to confirm the legitimacy of the identified events as glitches.

C.5 Additional marsquake background noise separation results

We present additional results on the separation of marsquake background noise and glitches, showcasing different marsquake characteristics. The first example pertains to a marsquake recorded on January 2, 2022 [InSight Marsquake Service, 2023]. This particular marsquake exhibits a larger amplitude and a longer coda wave compared to the one presented in Figure 5.8. Although the background noise appears negligible and is not readily visible in the raw waveform, this provides an opportunity to demonstrate the effectiveness of our unsupervised source separation method when one source (the marsquake in this case) dominates in amplitude.

To achieve the separation of background noise, we selected approximately 36 hours of detrended raw data from the U component with a sampling rate of 20Hz. This ensured an accurate

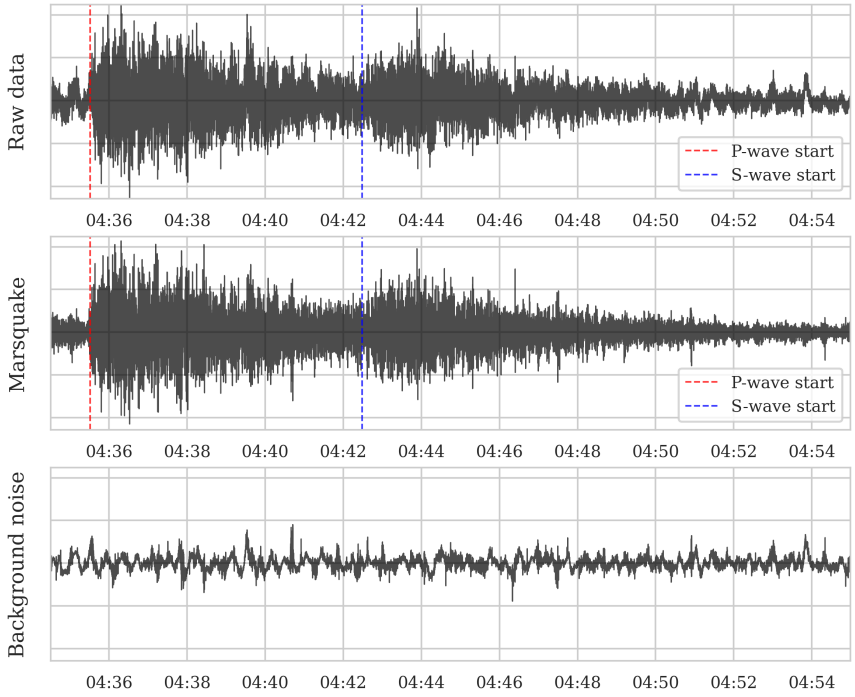


FIGURE C.8 – Unsupervised separation of background noise and glitches from a marsquake recorded by the InSight lander's seismometer on January 2, 2022 [InSight Marsquake Service, 2023]. Approximately 36 hours of raw data from the U component were used without any additional prior knowledge of marsquakes or glitches. The horizontal axis is in UTC time zone.

estimation of the wavelet scattering covariance statistics. The network architecture used is the same as in previous examples, and we employed a window size of 204.8,s. By solving the optimization problem outlined in equation (5.9) with 200 L-BFGS iterations, we obtained the results depicted in Figure C.8. Notably, glitches occurring just before the P-wave arrival and towards the end of the marsquake were successfully separated. Moreover, the separated background noise exhibits a stationary characteristic, which is desirable as it indicates minimal leakage of the marsquake signal.

The final example involves a marsquake recorded on July 26, 2019 [InSight Marsquake Service, 2023]. Separating the background noise in this case proves more challenging, as the P-wave arrival is barely discernible in the raw waveform shown in the top panel of Figure C.9. Furthermore, the presence of background noise masks the detection of the S-wave, as well as the secondary PP- and SS-wave arrivals. To address these complexities and achieve accurate separation of the marsquake while minimizing signal leakage, we require 95 hours of detrended raw data from the U component. A window size of 409.6 s is used, and the optimization problem in equation (5.9) is solved with 200 L-BFGS iterations. The results are depicted in Figure C.9, where the separated marsquake is distinctly delineated. The accuracy of our approach is further confirmed by the independently picked arrival times by the InSight team [Scholz, 2020], shown as dotted lines in Figure C.9. The alignment between their picked arrival times and our separated marsquake serves as validation for the accuracy of our method.

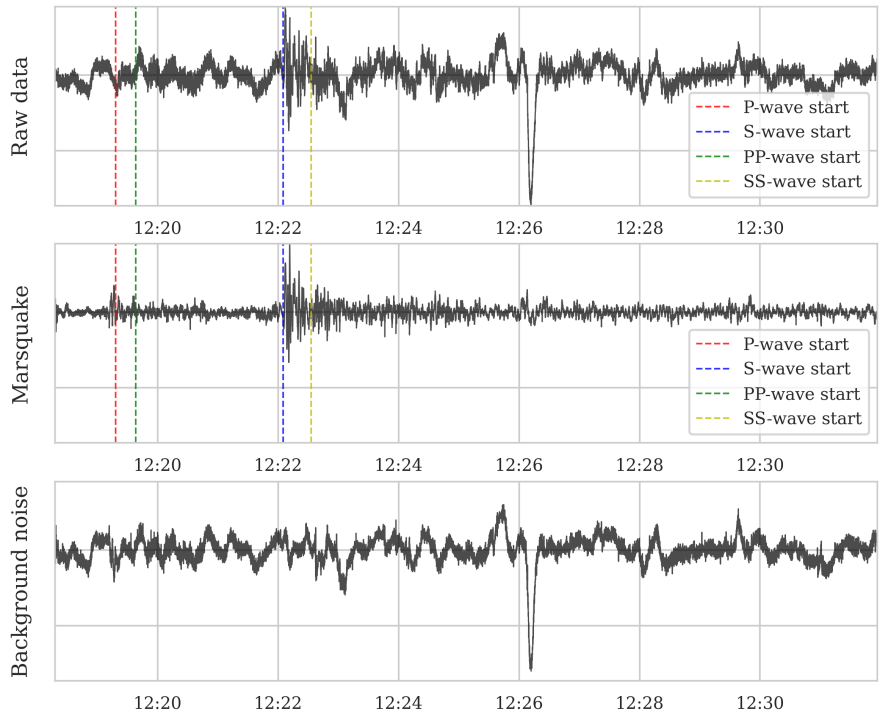


FIGURE C.9 – Unsupervised separation of background noise and glitches from a marsquake recorded by the InSight lander's seismometer on July 26, 2019 [InSight Marsquake Service, 2023]. Approximately 95 hours of raw data from the U component were used without any additional prior knowledge of marsquakes or glitches. The horizontal axis is in UTC time zone.

Annexe D

Appendices for Chapter 6

D.1 Smile sensitivity in the Scattering Spectra model

The Scattering Spectra model defined in (6.1) depends only on the estimated values $\Phi(\widetilde{x})$ of the Scattering Spectra estimated on a single realization $\widetilde{x}$ of S&P. This models gives an unconditional smile shown in Fig. 6.2b.

We are interested in the change in this unconditional smile in the case where the statistics $\Phi(\widetilde{x})$ change significantly. Thanks to the interpretation of Scattering Spectra coefficients, we can see what happens to the smile if the market is "more skewed" or "more kurtic" for example.

Here we focus on the shape of the smiles. Of course, changing the amplitude of Φ_2, which is equivalent to increasing the overall volatility of the model, only moves the overall level of the smile up or down.

These sensitivities in general can be seen as amplifying or reducing the departure of the price process from a Gaussian process x_{Gaussian}, also called Black-Scholes model, with the same average volatility, meaning the same value of $\Phi_2(\widetilde{x})$.

$$\text{new statistics} = (1-\lambda)\Phi(x_{\text{Gaussian}}) + \lambda\Phi(\widetilde{x}).$$

If $\lambda < 1$ the corresponding model should be "closer" to a Black-Scholes model, if $\lambda > 1$ the corresponding model should become less "Gaussian". Fig. D.1 shows 4 directions of change that are detailed below.

Skewness $|\Phi_3(x)|$. The skewness coefficients $\Phi_3(x_{\text{Gaussian}})$ should be zero for a Gaussian process. We consider a model of x with modified statistics

$$\Phi_3(x) = \lambda\Phi_3(\widetilde{x})$$

for 3 values of λ. For $\lambda = 0$, the modeled process is not skewed, meaning that an increment trajectory δx is equally likely as a trajectory $-\delta x$. Unsurprisingly, we get smiles that are symmetrical at $\mathcal{M} = 0$. For $\lambda = 1$ we get the same smile as in the Scattering Spectra model of S&P. For $\lambda = 1.3$, the smile has a higher downward slope, as expected.

Time-asymmetry $\operatorname{Im}\Phi(x)$. Our representation Φ is complex-valued. While Φ_1, Φ_2 are real, skewness Φ_3 and kurtosis Φ_4 may have non-zero imaginary parts that were shown to characterize certain types of time-asymmetry. We consider a model x whose statistics are

$$\Phi(x) = \operatorname{Re}\Phi(\widetilde{x}).$$

It is thus time-reversible, i.e. a trajectory $x(t)$ is equally likely as $x(-t)$. We notice that its smiles are symmetrical, but compared to the previous case, these are not symmetrical around $\mathcal{M} = 0$ but around values $\mathcal{M} > 0$ depending on the maturity. This is consistent since the process still has non-zero skewness $|\Phi_3(x)|$.

Kurtosis Φ_1**.** For a Gaussian process, $\Phi_1(x_{\text{Gaussian}}) = \pi/4$. We consider a model x with modified statistics

$$\Phi_1(x) = (1-\lambda)\frac{\pi}{4} + \lambda\Phi_1(\widetilde{x}).$$

For $\lambda = 0.5$, the model is less kurtic than the S&P, it shows smiles that tend to flatten around a straight line with negative slope. For $\lambda = 1.75$, the model is more kurtic and the smiles have more curvature, as expected.

Kurtosis Φ_4**.** We consider a model x with modified statistics

$$\Phi_4(x) = (1-\lambda)\Phi_4(x_{\text{Gaussian}}) + \lambda\Phi_4(\widetilde{x}).$$

The change in smiles for two different values $\lambda = 0.5, 1.5$ seem small compared to the other effects presented, however such changes impacts a lot the trajectories.

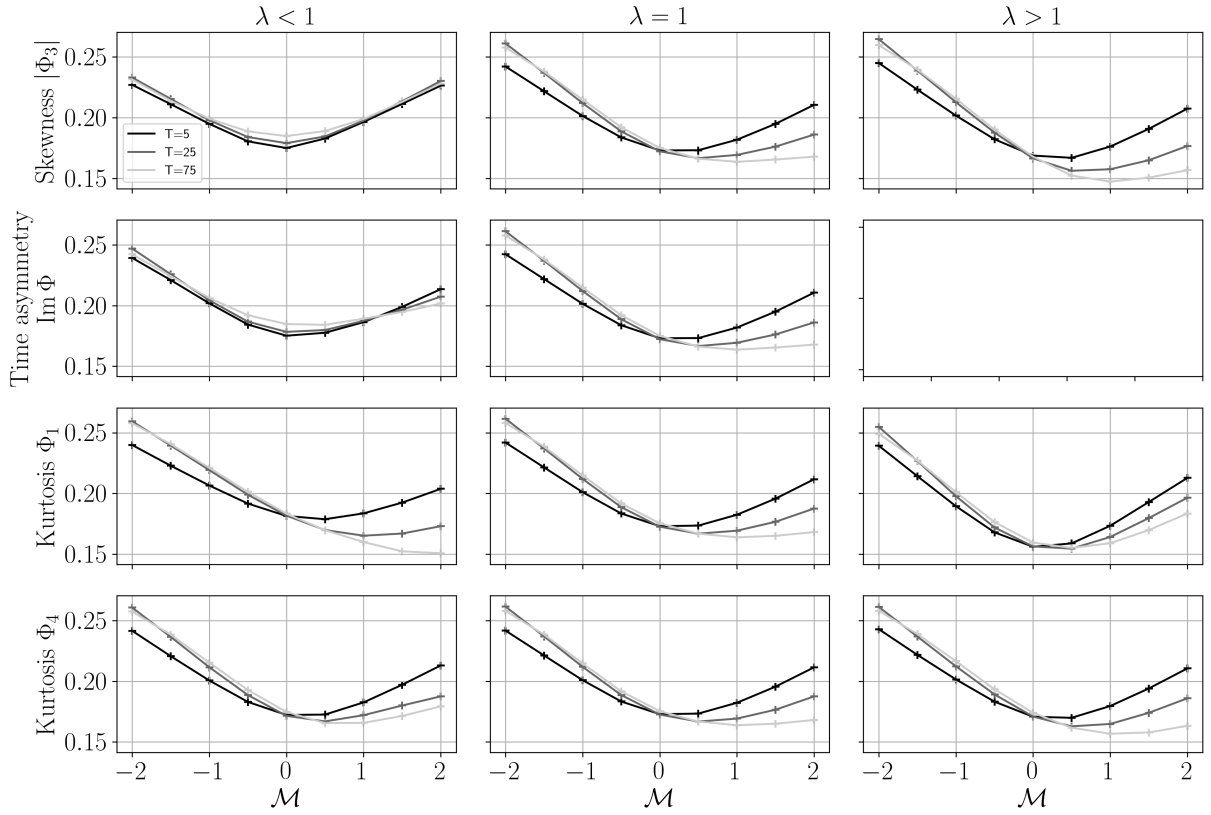


FIGURE D.1 – Smile sensitivity to change in Scattering Spectra statistics $\Phi(\widetilde{x})$, decomposed as changes in skewness $|\Phi_3(x)|$, time-asymmetry $\mathrm{Im}\,\Phi(x)$, kurtosis $\Phi_1(x)$ or kurtosis $|\Phi_4(x)|$. A value $\lambda < 1$ indicates respectively, no skewness, no time-asymmetry, less kurtosis. A factor $\lambda = 1$ does not change the statistics $\Phi(\widetilde{x})$ estimated on S&P. Besides well-known influence of the skewness and kurtosis on the shape of the smile, the Scattering Spectra $\Phi(x)$ also decompose the contribution of time-asymmetry in the shape of the smile.

D.2 The Path Dependent Volatility Model

The path-dependent volatility (PDV) model introduced in [Guyon, 2022] consists of a 4-factor Markovian model with 9 parameters. Writing the price process as $S_t = S_0 e^{x_t}$, it assumes that

$$\begin{aligned}
\frac{dS_t}{S_t} &= \sigma_t dW_t, \\
\sigma_t &= \sigma(R_{1,t}, R_{2,t}), \\
\sigma(R_1, R_2) &= \beta_0 + \beta_1 R_1 + \beta_2 R_2 \\
R_{1,t} &= \int_{-\infty}^{t} K_1(t-u)\frac{dS_u}{S_u} \\
R_{2,t} &= \int_{-\infty}^{t} K_2(t-u)\left(\frac{dS_u}{S_u}\right)^2
\end{aligned}$$

Among these parameters, 6 are used to parameterize the kernels K_1, K_2 both being a linear combination of exponentials, and 3 are the regression coefficients $\beta_0, \beta_1, \beta_2$. The 6 kernel parameters

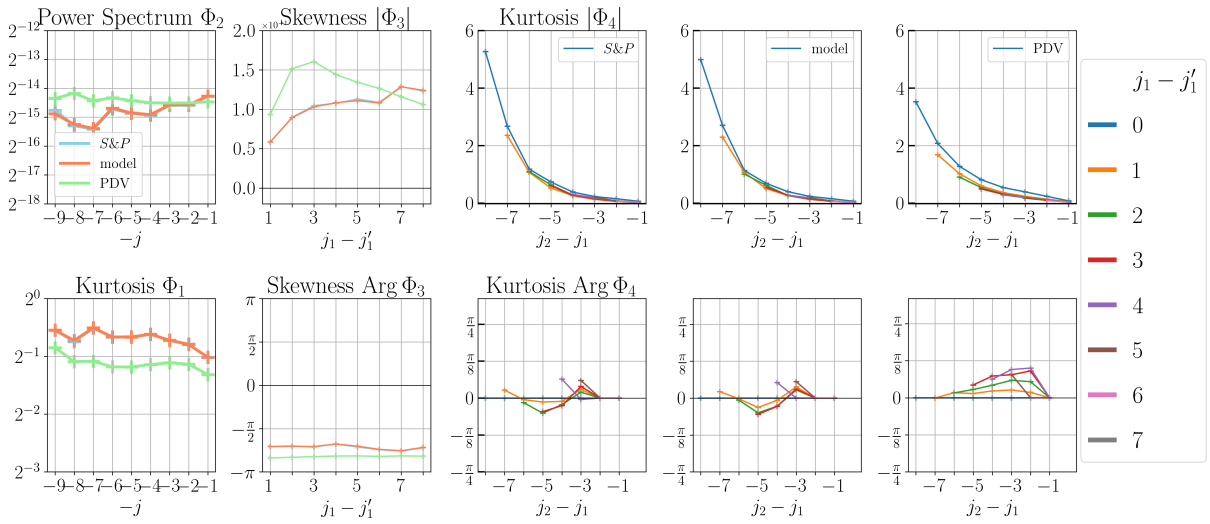


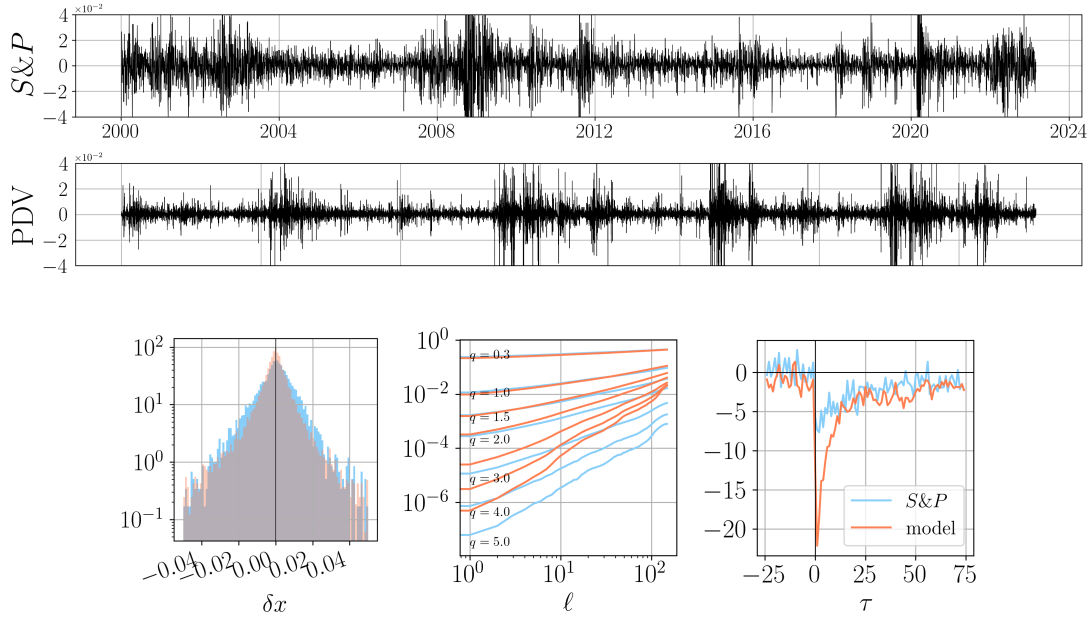
FIGURE D.2 – The Scattering Spectra : a statistical dashboard for financial prices. We compare these statistics in a Scattering Spectra model and path-dependent volatility (PDV) model to the one estimated on the S&P500.

are set to the optimal values presented in [Guyon, 2022] (parameter set 1, table 8, "Realistic paths").

To obtain unconditional smiles in the PDV model, the 3 regression coefficients $\beta_0, \beta_1, \beta_2$ are set to the optimal values presented in [Guyon, 2022] so that the PDV model we present here is exactly the same as in [Guyon, 2022]. The process is evolved with 10 steps per day until it reaches a stationary regimes. In such regime, the Scattering Spectra $\Phi(x)$ are shown in Fig.D.2 and are compared to the ones estimated in a Scattering Spectra model and on the S&P500 log-price time series $\widetilde{x}$ consisting of $N = 5827$ days from January 2000 to April 2023.

It shows that kurtosis $\Phi_1(x)$ and skewness $\Phi_3(x)$ have a significant mismatch with S&P500 data. Looking at the log-return trajectories shown in Fig. D.3 we indeed notice clear qualitative discrepancies, in particular we notice abnormal negative values, which can also be observed on price trajectories. In [Guyon, 2022] we can indeed see that these trajectories exhibit price drops that are more abrupt than those of the S&P500. In Fig. D.3 we see that structure functions (b) and, quite strikingly, the leverage effect (c) are poorly reproduced.

To obtain good conditional smiles, to be used in trading games (see section 6.5), we had to recalibrate the parameters $\beta_0(T), \beta_1(T), \beta_2(T)$ for each maturity T independently, this in order to provide the best prediction of the future realized variance, that is the overall level of the smile.



(a) Histogram (b) Struct. functions (c) Leverage

FIGURE D.3 – Standard statistics in a low parametric model (orange) of the S&P (blue). We chose a path-dependent volatlity model PDV. (a) Histogram of daily log-returns δx. (b) Structure functions $\mathbb{E}\{|\delta_\ell x(t)|^q\}$. (c) Leverage correlation $\mathbb{E}\{\delta x(t-\tau)|\delta x(t)|^2\}$.

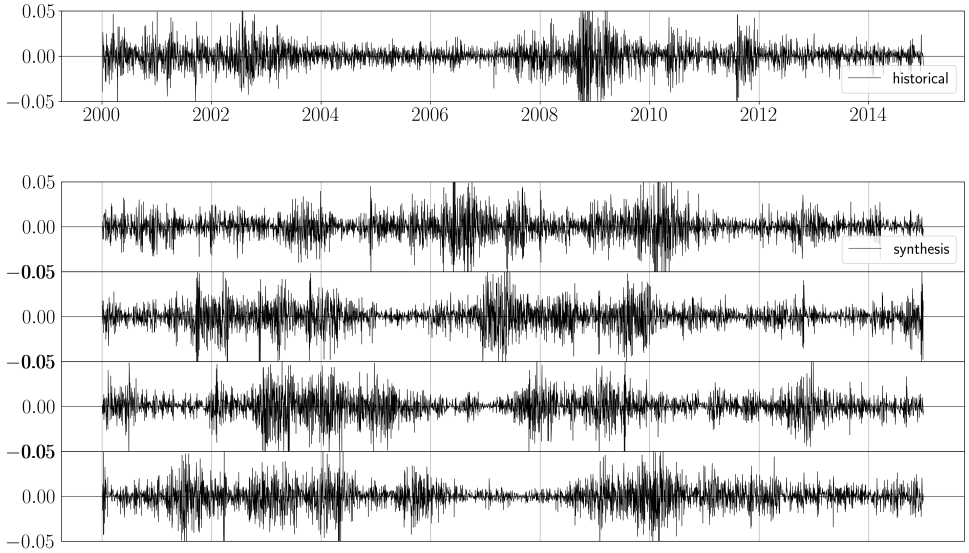


FIGURE D.4 – (Top) Historical S&P log-returns from 2000 to 2014. (Bottom) Generated syntheses from the Scattering Spectra model, which are scanned for shadowing paths.

D.3 Proof of theorem 4

Let us write $w_k = c_n g_\eta(x^k - \widetilde{x})$ where

$$g_\eta(x) = (\eta\sqrt{2\pi})^{-N'} e^{-\frac{1}{2\eta^2}\|x_{\text{past}}\|^2}$$

is a Gaussian kernel with N' being the dimension of x_{past} and c_n is such that $\frac{1}{n}\sum w_k = 1$. We write $\bar{Q}$ the estimator

$$\bar{Q} = \frac{1}{n}\sum_{k=1}^{n} w_k Q(x^k).$$

We prove the convergence of $\bar{Q}$ to $\mathbb{E}\{Q(x) \mid x_{\text{past}} = \widetilde{x}_{\text{past}}\}$ almost surely by first taking the limit $n \to +\infty$ and then $\eta \to 0$.

Limit $n \to \infty$. One can calculate $c_n^{-1} = \frac{1}{n}\sum g_\eta(x^k)$ so that one has

$$\bar{Q} = \frac{\frac{1}{n}\sum_{k=1}^{n} g_\eta(x^k - \widetilde{x})Q(x^k)}{\frac{1}{n}\sum_{k=1}^{n} g_\eta(x^k - \widetilde{x})}.$$

Since g_η is bounded one has $\mathbb{E}\{g_\eta(x-\widetilde{x})|Q(x)|\} < +\infty$ and $\mathbb{E}\{g_\eta(x-\widetilde{x})\} < +\infty$. From the law of large numbers, knowing that $\mathbb{E}\{g_\eta(x-\widetilde{x})\} > 0$, it follows

$$\bar{Q} \underset{n\to+\infty}{\longrightarrow} \frac{\mathbb{E}\{g_\eta(x-\widetilde{x})Q(x)\}}{\mathbb{E}\{g_\eta(x-\widetilde{x})\}}.$$

Limit $\eta \to 0$. We will make use of the following lemma of approximation by convolution, proved in [Evans, 2022].

Lemma 1. *If $f \in \mathcal{C}^0 \cap \boldsymbol{L}^1$ then for all $\widetilde{x} \in \mathbb{R}^N$*

$$g_\eta \star f(\widetilde{x}) \underset{\eta\to 0}{\longrightarrow} \int f(\widetilde{x}_{past}, x_{future})dx_{future}.$$

Let us notice that $\mathbb{E}\{g_\eta(x-\widetilde{x})\} = \int g_\eta(\widetilde{x}-x)p(x)dx = g_\eta \star p(\widetilde{x})$, and $\mathbb{E}\{g_\eta(x-\widetilde{x})Q(x)\} = g_\eta \star (Qp)(\widetilde{x})$. Since $\mathbb{E}\{Q(x)\} < +\infty$ one has $Qp \in \mathbf{L}^1$, p being a probability distribution one also has $p \in \mathbf{L}^1$. From the lemma we get :

$$\frac{\mathbb{E}\{g_\eta(x-\widetilde{x})Q(x)\}}{\mathbb{E}\{g_\eta(x-\widetilde{x})\}} \underset{\eta\to 0}{\longrightarrow} \frac{\int Q(\widetilde{x}_{\text{past}}, x_{\text{future}})p(\widetilde{x}_{\text{past}}, x_{\text{future}})dx_{\text{future}}}{\int p(\widetilde{x}_{\text{past}}, x_{\text{future}})dx_{\text{future}}},$$

where the denominator is non-zero because $p(x) > 0$ for all $x \in \mathbb{R}^N$. The former term being $\mathbb{E}\{Q(x) \mid x_{\text{past}} = \widetilde{x}_{\text{past}}\}$, this proves the theorem.

D.4 Choice of representation h

The following proposition shows that the choice of h in chapter 6 (6.15) induces equivariance properties on the set of shadowing paths $H_\eta(\widetilde{x}_{\text{past}})$. We recall that we chose $\eta = \widehat{\eta}\|h(\widetilde{x}_{\text{past}})\|$ for a fixed $\widehat{\eta}$. These equivariance properties are proved on continuously sampled paths $x_{\text{past}} = (x(t), t < 0)$ with $t \in \mathbb{R}$. In that case, we still write $h_{\alpha,\beta}$ the continuously sampled analogue

$$h_{\alpha,\beta}(x) = \Big(\frac{x(0) - x(-\ell)}{\ell^\beta}\ ,\ \ell = \alpha^m\ ,\ m \in \mathbb{Z}\Big)$$

which is now of infinite dimension.

Proposition 7. *For $h = h_{\alpha,\beta}$ with $\alpha > 0$ and $\beta > 1$ one has*
1. (Multiplication equivariance) for $\lambda > 0$

$$H_\eta(\lambda\, \widetilde{x}_{past}) = \lambda . H_\eta(\widetilde{x}_{past})$$

2. (Dilation equivariance) writing $\Gamma x(t) = x(\alpha t)$

$$H_\eta(\Gamma \widetilde{x}_{past}) = \Gamma . H_\eta(\widetilde{x}_{past})$$

Proof. The first equivariance follows directly from the fact that $h_{\alpha,\beta}$ is itself equivariant to multiplication $h_{\alpha,\beta}(\lambda x_{\rm past}) = \lambda\, h_{\alpha,\beta}(x_{\rm past})$. For dilation, one has

$$\begin{aligned} h_{\alpha,\beta}(\Gamma x_{\rm past}) &= \Big(\frac{x(0) - x(-\alpha\ell)}{\ell^\beta} \ ,\ \ell = \alpha^m \ ,\ m \in \mathbb{Z} \Big) \\ &= \alpha^\beta \Big(\frac{x(0) - x(-\ell)}{\ell^\beta} \ ,\ \ell = \alpha^{m+1} \ ,\ m \in \mathbb{Z} \Big) \end{aligned}$$

This means that $h_{\alpha,\beta}(\Gamma x_{\rm past})$ is equal to $h_{\alpha,\beta}(x_{\rm past})$ up to a shift in indices and up to a multiplicative constant. It follows that

$$\|h_{\alpha,\beta}(\Gamma x_{\rm past}) - h_{\alpha,\beta}(\Gamma \widetilde{x}_{\rm past})\| = \alpha^\beta \|h_{\alpha,\beta}(x_{\rm past}) - h_{\alpha,\beta}(\widetilde{x}_{\rm past})\|.$$

Now, normalizing by $\|h_{\alpha,\beta}(\Gamma \widetilde{x}_{\rm past})\| = \alpha^\beta \|h_{\alpha,\beta}(\widetilde{x}_{\rm past})\|$ yields $H_\eta(\Gamma \widetilde{x}_{\rm past}) = \Gamma . H_\eta(\widetilde{x}_{\rm past})$. □

D.5 Additional trading game statistics

In addition to the P&Ls and aggregated P&Ls of a trading game played against the option market shown in Figs. 6.5,6.6, we show in Figs. D.5,D.6 the standard deviation and winning rate, defined as the average number of times payoff of the hedged-option exceeds its initial price.

In the remaining figures, we also show the statistical results of the trading game between PDV and Scattering Spectra. These results do not require option market data (but require actual price series of the underlying) and directly test the relative quality of purely statistical price models. As seen in Fig. D.9, trading game unequivocally favours the Scattering Spectra model framework over the PDV model.

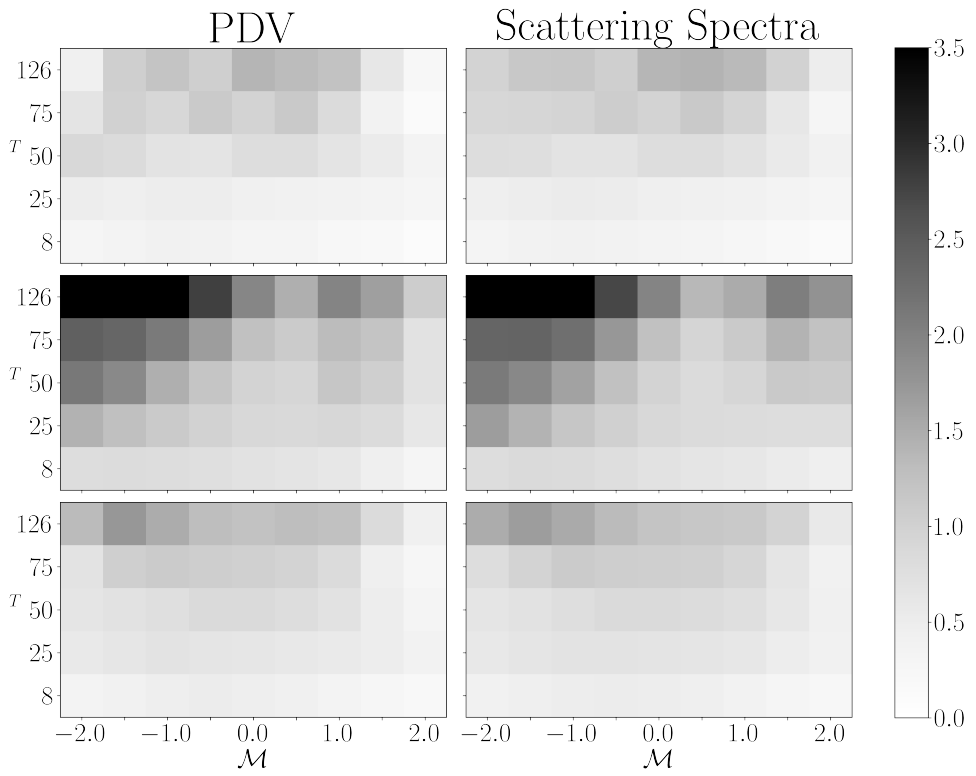


FIGURE D.5 – Standard deviation of P&Ls when playing the Scattering Spectra model vs. S&P or PDV model vs. S&P. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

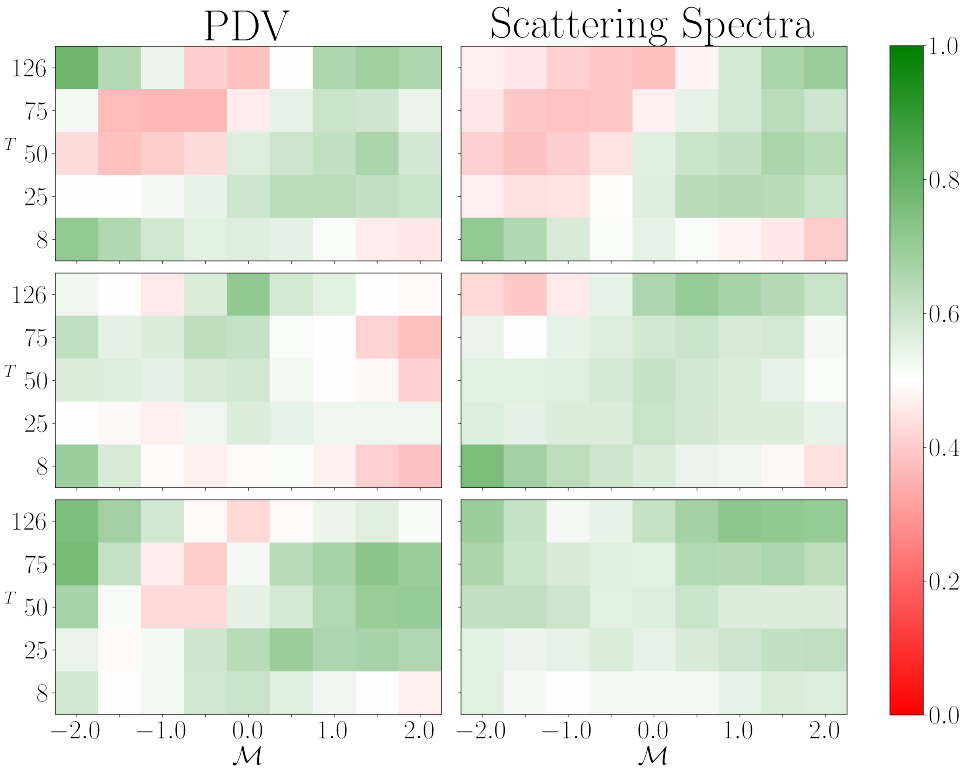


FIGURE D.6 – Rate of winning trades when playing the Scattering Spectra model vs. S&P or PDV model vs. S&P. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

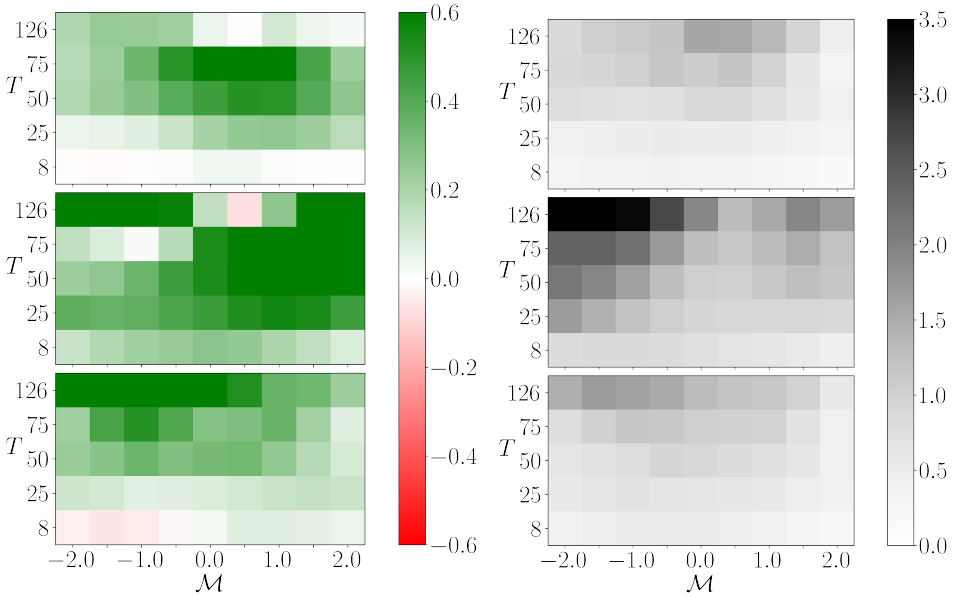


FIGURE D.7 – P&Ls average (left) and standard deviation (right) when playing the Scattering Spectra model against PDV model. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

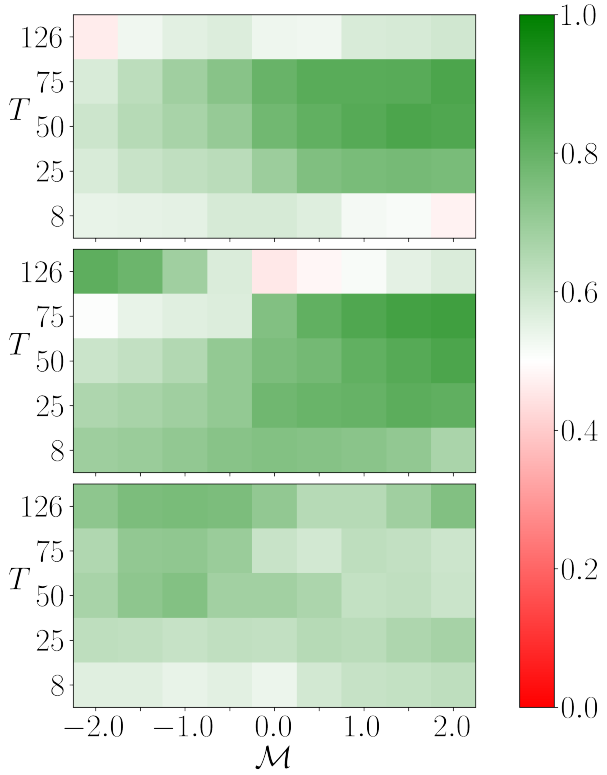


FIGURE D.8 – Rate of winning trades when playing the Scattering Spectra model against PDV model. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

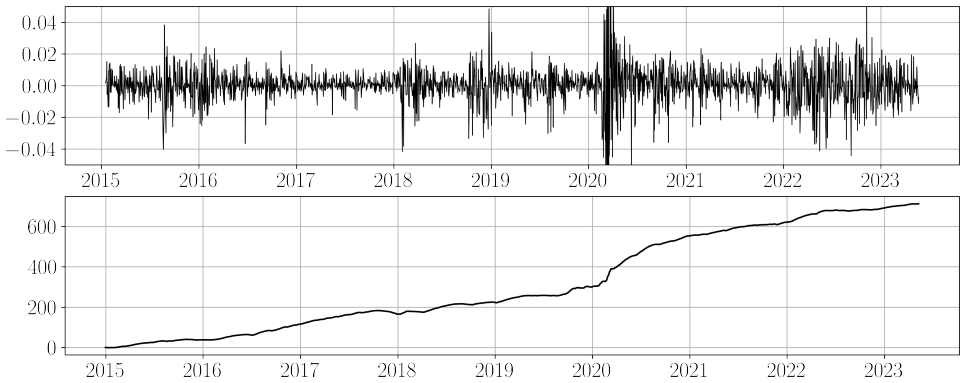


FIGURE D.9 – Aggregated P&L when playing the Scattering Spectra model vs. PDV model. Each heat-map corresponds to a 3 years period, from top to bottom (2015-2017, 2018-2020, 2021-2023).

Bibliography

[Abry, 2000] Patrice Abry, Patrick Flandrin, Murad S Taqqu, Darryl Veitch. « Wavelets for the analysis, estimation, and synthesis of scaling data ». *Self-Similar Network Traffic and Performance Evaluation* (2000), p. 39-88 (cf. p. 37).

[Abry, 2010] Patrice Abry, Herwig Wendt, Stéphane Jaffard, Hannes Helgason, Paulo Gonçalves, Edmundo Pereira et al. « Methodology for multifractal analysis of heart rate variability : From LF/HF ratio to wavelet leaders ». *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, p. 106-109 (cf. p. 5, 22).

[Adali, 2015] Tülay Adali, Yuri Levin-Schwartz, Vince D. Calhoun. « Multimodal Data Fusion Using Source Separation : Application to Medical Imaging ». *Proceedings of the IEEE* 103.9 (2015), p. 1494-1506 (cf. p. 100).

[Alden, 2022] Andrew Alden, Carmine Ventre, Blanka Horvath, Gordon Lee. « Model-Agnostic Pricing of Exotic Derivatives Using Signatures ». *Proceedings of the Third ACM International Conference on AI in Finance*. 2022, p. 96-104 (cf. p. 115).

[Allys, 2019] Erwan Allys, F Levrier, S Zhang, C Colling, B Regaldo-Saint Blancard, F Boulanger, P Hennebelle, S Mallat. « The RWST, a comprehensive statistical description of the non-Gaussian structures in the ISM ». *Astronomy & Astrophysics* 629 (2019), A115 (cf. p. 64, 72, 75).

[Allys, 2020] Erwan Allys, Tanguy Marchand, Jean-François Cardoso, Francisco Villaescusa-Navarro, Shirley Ho, Stéphane Mallat. « New interpretable statistics for large-scale structure analysis and generation ». *Physical Review D* 102.10 (2020), p. 103506 (cf. p. 11, 27, 64).

[Andén, 2018] Joakim Andén, Vincent Lostanlen, Stéphane Mallat. « Classification with joint time-frequency scattering ». *arXiv preprint arXiv :1807.08869* (2018) (cf. p. 5, 23).

[Andreux, 2018] Mathieu Andreux. « Foveal autoregressive neural time-series modeling ». Thèse de doct. Université Paris sciences et lettres, 2018 (cf. p. 16, 33).

[Anosov, 1969] Dmitrij V Anosov. *Geodesic flows on closed Riemann manifolds with negative curvature*. 89-90. American Mathematical Society, 1969 (cf. p. 125).

[Aubrun, 2021] C Aubrun, M Benzaquen, J.-P. Bouchaud. « On Hawkes Processes with Infinite Mean Intensity ». *arXiv !* (2021), p. 2112.14161 (cf. p. 54).

[Aubrun, 2024] Cecilia Aubrun, Rudy Morel, Michael Benzaquen, Jean-Philippe Bouchaud. « Riding wavelets : A method to discover new classes of price jumps ». *arXiv preprint arXiv :2404.16467* (2024) (cf. p. 17, 34).

[Auclair, 2023] Constant Auclair, Erwan Allys, François Boulanger, Matthieu Béthermin, Athanasia Gkogkou, Guilaine Lagache et al. « Separation of dust emission from the Cosmic Infrared Background in Herschel observations with Wavelet Phase Harmonics ». *arXiv preprint arXiv :2305.14419* (2023) (cf. p. 64).

[Bacry, 2001a] E. Bacry, J. Delour, J. F. Muzy. « Multifractal random walk ». *Phys. Rev. E* 64 (2 2001), p. 026103 (cf. p. 48, 53, 55, 105).

[Bacry, 2001b] Emmanuel Bacry, Jean Delour, Jean-François Muzy. « Modelling financial time series using multifractal random walks ». *Physica A : statistical mechanics and its applications* 299.1-2 (2001), p. 84-92 (cf. p. 54, 55).

[Bacry, 2013] Emmanuel Bacry, Alexey Kozhemyak, Jean-François Muzy. « Log-normal continuous cascade model of asset returns : aggregation properties and estimation ». *Quantitative Finance* 13.5 (2013), p. 795-818 (cf. p. 115, 116, 123).

[Bacry, 2015] Emmanuel Bacry, Iacopo Mastromatteo, Jean-François Muzy. « Hawkes processes in finance ». *Market Microstructure and Liquidity* 1.01 (2015), p. 1550005 (cf. p. 54, 55).

[Bacry, 2014] Emmanuel Bacry, Jean-François Muzy. « Hawkes model for price and trades high-frequency dynamics ». *Quantitative Finance* 14.7 (2014), p. 1147-1166 (cf. p. 54).

[Bacry, 1993] Emmanuel Bacry, Jean-François Muzy, Alain Arneodo. « Singularity spectrum of fractal signals from wavelet analysis : Exact results ». *Journal of statistical physics* 70.3 (1993), p. 635-674 (cf. p. 5, 22, 37).

[Bak, 1987] Per Bak, Chao Tang, Kurt Wiesenfeld. « Self-organized criticality : An explanation of the 1/f noise ». *Phys. Rev. Lett.* 59 (4 1987), p. 381-384 (cf. p. 67).

[Barkaoui, 2021] Salma Barkaoui, Philippe Lognonné, Taichi Kawamura, Éléonore Stutzmann, Léonard Seydoux, Maarten de Hoop et al. « Anatomy of continuous Mars SEIS and pressure data from unsupervised learning ». *Bulletin of the Seismological Society of America* 111.6 (2021), p. 2964-2981 (cf. p. 101, 105).

[Barriga, 2003] Eduardo S. Barriga, Paul W. Truitt, Marios S. Pattichis, Dan T'so, Young H. Kwon M.D., Randy H. Kardon M.D., Peter Soliz. « Blind source separation in retinal videos ». *Medical Imaging 2003 : Image Processing*. Sous la dir. de Milan Sonka, J. Michael Fitzpatrick. T. 5032. International Society for Optics et Photonics. SPIE, 2003, p. 1591-1601 (cf. p. 100).

[Battle, 1987] Guy Battle. « A block spin construction of ondelettes. Part I : Lemarié functions ». *Communications in Mathematical Physics* 110.4 (1987), p. 601-615 (cf. p. 40, 118, 138).

[Bayer, 2016] Christian Bayer, Peter Friz, Jim Gatheral. « Pricing under rough volatility ». *Quantitative Finance* 16.6 (2016), p. 887-904 (cf. p. 122).

[Beghein, 2022] C. Beghein, J. Li, E. Weidner, R. Maguire, J. Wookey, V. Lekić, P. Lognonné, W. Banerdt. « Crustal Anisotropy in the Martian Lowlands From Surface Waves ». *Geophysical Research Letters* 49.24 (2022), e2022GL101508. eprint : `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL101508` (cf. p. 101).

[Bekaert, 2000] Geert Bekaert, Guojun Wu. « Asymmetric volatility and risk in equity markets ». *The review of financial studies* 13.1 (2000), p. 1-42 (cf. p. 54, 55, 58, 88).

[Bergomi, 2009] Lorenzo Bergomi. « Smile dynamics IV ». *Available at SSRN 1520443* (2009) (cf. p. 117, 123).

[Betancourt, 2017] Michael Betancourt. « A conceptual introduction to Hamiltonian Monte Carlo ». *arXiv preprint arXiv :1701.02434* (2017) (cf. p. 2, 19).

[Bickel, 2008] Peter J. Bickel, Elizaveta Levina. « Covariance regularization by thresholding ». *The Annals of Statistics* 36.6 (2008), p. 2577-2604 (cf. p. 73).

[Billingsley, 2013] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013 (cf. p. 3, 21).

[Bingham, 2000] Ella Bingham, Aapo Hyvärinen. « A FAST FIXED-POINT ALGORITHM FOR INDEPENDENT COMPONENT ANALYSIS OF COMPLEX VALUED SIGNALS ». *International Journal of Neural Systems* 10.01 (2000), p. 1-8 (cf. p. 100).

[Blanc, 2017] Pierre Blanc, Jonathan Donier, J-P Bouchaud. « Quadratic Hawkes processes for financial prices ». *Quantitative Finance* 17.2 (2017), p. 171-188 (cf. p. 54, 55).

[Bobin, 2007] Jérôme Bobin, Jean-Luc Starck, Jalal Fadili, Yassir Moudden. « Sparsity and morphological diversity in blind source separation ». *IEEE Transactions on Image Processing* 16.11 (2007), p. 2662-2674 (cf. p. 13, 30, 100).

[Borland, 2005] Lisa Borland, Jean-Philippe Bouchaud, Jean-François Muzy, Gilles Zumbach. « The Dynamics of Financial Markets–Mandelbrot's multifractal cascades, and beyond ». *arXiv preprint cond-mat/0501292* (2005) (cf. p. 116).

[Bormetti, 2013] Giacomo Bormetti, Lucio Maria Calcagnile, Michele Treccani, Fulvio Corsi, Stefano Marmi, Fabrizio Lillo et al. « Modelling systemic cojumps with Hawkes factor models ». *arXiv preprint arXiv :1301.6141* (2013) (cf. p. 90).

[Bouchaud, 2013] Jean-Philippe Bouchaud, Lorenzo De Leo, Vincent Vargas, Stefano Ciliberti. « Smile in the low moments ». *Risk* 26.7 (2013), p. 56 (cf. p. 118, 120, 121, 123).

[Bouchaud, 1990] Jean-Philippe Bouchaud, Antoine Georges. « Anomalous diffusion in disordered media : Statistical mechanisms, models and physical applications ». *Physics Reports* 195.4 (1990), p. 127-293 (cf. p. 67).

[Bouchaud, 2001] Jean-Philippe Bouchaud, Andrew Matacz, Marc Potters. « Leverage effect in financial markets : The retarded volatility model ». *Physical review letters* 87.22 (2001), p. 228701 (cf. p. 54, 55, 58, 88).

[Bouchaud, 2003] Jean-Philippe Bouchaud, Marc Potters. *Theory of financial risk and derivative pricing : from statistical physics to risk management*. Cambridge university press, 2003 (cf. p. 85).

[Bougeret, 1995] J. -L. Bougeret, M. L. Kaiser, P. J. Kellogg, R. Manning, K. Goetz, S. J. Monson et al. « Waves : The Radio and Plasma Wave Investigation on the Wind Spacecraft ». 71.1-4 (1995), p. 231-263 (cf. p. 63).

[Bowen, 1975] Rufus Bowen. « ω-limit sets for axiom A diffeomorphisms ». *Journal of differential equations* 18.2 (1975), p. 333-339 (cf. p. 125).

[Brillinger, 1965] David R Brillinger. « An introduction to polyspectra ». *The Annals of mathematical statistics* (1965), p. 1351-1374 (cf. p. 37, 67).

[Brochard, 2022] Antoine Brochard, Sixin Zhang, Stéphane Mallat. « Generalized rectifier wavelet covariance models for texture synthesis ». *arXiv preprint arXiv :2203.07902* (2022) (cf. p. 7, 11, 27, 28).

[Bruna, 2013] Joan Bruna, Stéphane Mallat. « Invariant scattering convolution networks ». *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), p. 1872-1886 (cf. p. 5, 23, 37, 64, 70, 71, 82, 134).

[Bruna, 2019] Joan Bruna, Stéphane Mallat. « Multiscale sparse microcanonical models ». *Mathematical Statistics and Learning* 1.3 (2019), p. 257-315 (cf. p. 2, 3, 20, 37, 38, 141, 142).

[Bruna, 2015] Joan Bruna, Stéphane Mallat, Emmanuel Bacry, Jean-François Muzy. « Intermittent process analysis with scattering moments » (2015) (cf. p. 5, 23, 52).

[Buades, 2011] Antoni Buades, Bartomeu Coll, Jean-Michel Morel. « Non-local means denoising ». *Image Processing On Line* 1 (2011), p. 208-212 (cf. p. 15, 32, 126).

[Cai, 2011] Tony Cai, Weidong Liu. « Adaptive Thresholding for Sparse Covariance Matrix Estimation ». *Journal of the American Statistical Association* 106.494 (2011), p. 672-684. eprint : `https://doi.org/10.1198/jasa.2011.tm10560` (cf. p. 73).

[Cardoso, 1989] J.-F. Cardoso. « Source separation using higher order moments ». *International Conference on Acoustics, Speech, and Signal Processing,* 1989, 2109-2112 vol.4 (cf. p. 13, 30, 100).

[Cardoso, 1998] J.-F. Cardoso. « Blind signal separation : statistical principles ». *Proceedings of the IEEE* 86.10 (1998), p. 2009-2025 (cf. p. 13, 30, 100).

[Ceylan, 2022] Savas Ceylan, John F. Clinton, Domenico Giardini, Simon C. Stähler, Anna Horleston, Taichi Kawamura et al. « The marsquake catalogue from InSight, sols 0–1011 ». *Physics of the Earth and Planetary Interiors* 333 (2022), p. 106943 (cf. p. 101).

[Chanal, 2000] O Chanal, B Chabaud, B Castaing, B Hébral. « Intermittency in a turbulent low temperature gaseous helium jet ». *The European Physical Journal B-Condensed Matter and Complex Systems* 17.2 (2000), p. 309-317 (cf. p. 56).

[Chang, 2000] S Grace Chang, Bin Yu, Martin Vetterli. « Adaptive wavelet thresholding for image denoising and compression ». *IEEE transactions on image processing* 9.9 (2000), p. 1532-1546 (cf. p. 73).

[Chen, 2001] Scott Shaobing Chen, David L Donoho, Michael A Saunders. « Atomic decomposition by basis pursuit ». *SIAM review* 43.1 (2001), p. 129-159 (cf. p. 92).

[Cheng, 2021a] Sihao Cheng, Brice Ménard. « How to quantify fields or textures ? A guide to the scattering transform ». *arXiv preprint arXiv :2112.01288* (2021) (cf. p. 64, 70, 79).

[Cheng, 2021b] Sihao Cheng, Brice Ménard. « Weak lensing scattering transform : dark energy and neutrino mass sensitivity ». *Monthly Notices of the Royal Astronomical Society* (2021) (cf. p. 64).

[Cheng, 2024] Sihao Cheng, Rudy Morel, Erwan Allys, Brice Ménard, Stéphane Mallat. « Scattering spectra models for physics ». *PNAS nexus* 3.4 (2024), pgae103 (cf. p. 17, 34).

[Cheng, 2020] Sihao Cheng, Yuan-Sen Ting, Brice Ménard, Joan Bruna. « A new approach to observational cosmology using the scattering transform ». 499.4 (2020), p. 5902-5914. arXiv : `2006.08561 [astro-ph.CO]` (cf. p. 64).

[Chevreuil, 2014] Antoine Chevreuil, Philippe Loubaton. « Chapter 4 - Blind Signal Separation for Digital Communication Data ». *Academic Press Library in Signal Processing : Volume 2.* Sous la dir. de Nicholas D. Sidiropoulos, Fulvio Gini, Rama Chellappa, Sergios Theodoridis. T. 2. Academic Press Library in Signal Processing. Elsevier, 2014, p. 135-186 (cf. p. 100).

[Chicheportiche, 2015] Rémy Chicheportiche, J-P Bouchaud. « A nested factor model for non-linear dependencies in stock returns ». *Quantitative Finance* 15.11 (2015), p. 1789-1804 (cf. p. 85).

[Chicheportiche, 2014a] Rémy Chicheportiche, Jean-Philippe Bouchaud. « The fine-structure of volatility feedback I : Multi-scale self-reflexivity ». *Physica A : Statistical Mechanics and its Applications* 410 (2014), p. 174-195 (cf. p. 55, 59).

[Chicheportiche, 2014b] Rémy Chicheportiche, Anirban Chakraborti. « Copulas and time series with long-ranged dependencies ». *Physical Review E* 89.4 (2014), p. 042117 (cf. p. 12, 29, 85, 94, 95).

[Christiansen, 2012] Charlotte Christiansen, Maik Schmeling, Andreas Schrimpf. « A comprehensive look at financial volatility prediction by economic variables ». *Journal of Applied Econometrics* 27.6 (2012), p. 956-977 (cf. p. 14, 31).

[Chua, 2016] Jiawen Chua, Ganlong Wang, W. Bastiaan Kleijn. « Convolutive blind source separation with low latency ». *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. 2016, p. 1-5 (cf. p. 100).

[Coles, 1991] Peter Coles, Bernard Jones. « A lognormal model for the cosmological mass distribution. » 248 (1991), p. 1-13 (cf. p. 67, 79).

[Delemotte, 2023] Jules Delemotte, Stefano De Marco, Florent Segonne. « Yet another analysis of the SP500 at-the-money skew : crossover of different power-law behaviours ». *Available at SSRN 1520443* (2023) (cf. p. 116, 122).

[Delouis, 2022] J-M Delouis, E Allys, Edouard Gauvrit, F Boulanger. « Non-Gaussian modelling and statistical denoising of Planck dust polarisation full-sky maps using scattering transforms ». *Astronomy & Astrophysics* 668 (2022), A122 (cf. p. 64, 82).

[Delouis J-M, 2022] Delouis, J.-M., Allys, E., Gauvrit, E., Boulanger, F. « Non-Gaussian modelling and statistical denoising of Planck dust polarisation full-sky maps using scattering transforms ». *A&A* 668 (2022), A122 (cf. p. 101, 103).

[Denton, 2022] Tom Denton, Scott Wisdom, John R. Hershey. « Improving Bird Classification with Unsupervised Sound Separation ». *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, p. 636-640 (cf. p. 13, 30, 100).

[Desnianskii, 1974] VN Desnianskii, EA Novikov. « Evolution of turbulence spectra toward a similarity regime ». *Akademiia Nauk SSSR, Izvestiia, Fizika Atmosfery i Okeana* 10 (1974), p. 127-136 (cf. p. 10, 27).

[Donoho, 2006] David L Donoho. « For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution ». *Communications on Pure and Applied Mathematics : A Journal Issued by the Courant Institute of Mathematical Sciences* 59.6 (2006), p. 797-829 (cf. p. 92).

[Donoho, 1994] David L Donoho, Iain M Johnstone. « Ideal spatial adaptation by wavelet shrinkage ». *Biometrika* 81.3 (1994), p. 425-455. eprint : `https://academic.oup.com/biomet/article-pdf/81/3/425/26079146/81.3.425.pdf` (cf. p. 73).

[Drude, 2019] Lukas Drude, Daniel Hasenklever, Reinhold Haeb-Umbach. « Unsupervised training of a deep clustering model for multichannel blind source separation ». *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, p. 695-699 (cf. p. 13, 30, 100).

[Dudok de Wit, 2004] T. Dudok de Wit. « Can high-order moments be meaningfully estimated from experimental turbulence measurements ? » *Phys. Rev. E* 70 (5 2004), p. 055302 (cf. p. 67).

[Eckerli, 2021] Florian Eckerli. « Generative Adversarial Networks in finance : an overview ». *Available at SSRN 3864965* (2021) (cf. p. 56).

[El Euch, 2019a] Omar El Euch, Jim Gatheral, Mathieu Rosenbaum. « Roughening heston ». *Risk* (2019), p. 84-89 (cf. p. 55).

[El Euch, 2019b] Omar El Euch, Mathieu Rosenbaum. « The characteristic function of rough Heston models ». *Mathematical Finance* 29.1 (2019), p. 3-38 (cf. p. 55).

[Evans, 2022] Lawrence C Evans. *Partial differential equations*. T. 19. American Mathematical Society, 2022 (cf. p. 164).

[Fama, 1965] Eugene F Fama. « The behavior of stock-market prices ». *The journal of Business* 38.1 (1965), p. 34-105 (cf. p. 90).

[Fama, 1993] Eugene F Fama, Kenneth R French. « Common risk factors in the returns on stocks and bonds ». *Journal of financial economics* 33.1 (1993), p. 3-56 (cf. p. 85).

[Fan, 2013] Jianqing Fan, Yuan Liao, Martina Mincheva. « Large covariance estimation by thresholding principal orthogonal complements ». *Journal of the Royal Statistical Society. Series B, Statistical methodology* 75.4 (2013) (cf. p. 73).

[Févotte, 2009] Cédric Févotte, Nancy Bertin, Jean-Louis Durrieu. « Nonnegative Matrix Factorization with the Itakura-Saito Divergence : With Application to Music Analysis ». *Neural Computation* 21.3 (2009), p. 793-830. eprint : `https://direct.mit.edu/neco/article-pdf/21/3/793/820289/neco.2008.04-08-771.pdf` (cf. p. 13, 30, 100).

[Flandrin, 1992] Patrick Flandrin. « Wavelet analysis and synthesis of fractional Brownian motion ». *IEEE Transactions on information theory* 38.2 (1992), p. 910-917 (cf. p. 40).

[Frisch, 1991] U Frisch. « From global scaling, a la Kolmogorov, to local multifractal scaling in fully developed turbulence ». *Proceedings of the Royal Society of London. Series A : Mathematical and Physical Sciences* 434.1890 (1991), p. 89-99 (cf. p. 10, 27, 56).

[Frisch, 1985] Uriel Frisch, Giorgio Parisi. « Fully developed turbulence and intermittency ». *Proceedings of the International Summer School on Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics* (1985), p. 84-88 (cf. p. 37).

[Fukasawa, 2017] Masaaki Fukasawa. « Short-time at-the-money skew and rough fractional volatility ». *Quantitative Finance* 17.2 (2017), p. 189-198 (cf. p. 115, 117).

[Fukasawa, 2021] Masaaki Fukasawa. « Volatility has to be rough ». *Quantitative Finance* 21.1 (2021), p. 1-8 (cf. p. 122).

[Gallagher, 2013] Isabelle Gallagher, Laure Saint-Raymond, Benjamin Texier. *From Newton to Boltzmann : hard spheres and short-range potentials*. European Mathematical Society Zürich, Switzerland, 2013 (cf. p. 3, 20).

[Gatheral, 2018] Jim Gatheral, Thibault Jaisson, Mathieu Rosenbaum. « Volatility is rough ». *Quantitative finance* 18.6 (2018), p. 933-949 (cf. p. 55, 115, 116, 122, 123).

[Gatys, 2015] Leon Gatys, Alexander S Ecker, Matthias Bethge. « Texture synthesis using convolutional neural networks ». *Advances in neural information processing systems* 28 (2015) (cf. p. 6, 23, 60).

[Gay, 2012] Steven L Gay, Jacob Benesty. *Acoustic signal processing for telecommunication*. T. 551. Springer Science & Business Media, 2012 (cf. p. 100).

[Geman, 1984] Stuart Geman, Donald Geman. « Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images ». *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), p. 721-741 (cf. p. 65).

[Ghysels, 2006] Eric Ghysels, Pedro Santa-Clara, Rossen Valkanov. « Predicting volatility : getting the most out of return data sampled at different frequencies ». *Journal of Econometrics* 131.1-2 (2006), p. 59-95 (cf. p. 14, 31).

[Giardini, 2020] Domenico Giardini, Philippe Lognonné, W Bruce Banerdt, William T Pike, Ulrich Christensen, Savas Ceylan et al. « The seismicity of Mars ». *Nature Geoscience* 13.3 (2020), p. 205-212 (cf. p. 101).

[Golombek, 2020] M Golombek, NH Warner, JA Grant, Ernst Hauber, V Ansan, CM Weitz et al. « Geology of the InSight landing site on Mars ». *Nature communications* 11.1 (2020), p. 1-11 (cf. p. 101).

[Grais, 2014] Emad M. Grais, Mehmet Umut Sen, Hakan Erdogan. « Deep neural networks for single channel source separation ». *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, p. 3734-3738 (cf. p. 100).

[Greig, 2022] Bradley Greig, Yuan-Sen Ting, Alexander A. Kaurov. « Exploring the cosmic 21-cm signal from the epoch of reionization using the wavelet scattering transform ». 513.2 (2022), p. 1719-1741. arXiv : `2204.02544 [astro-ph.CO]` (cf. p. 64).

[Gupta, 2018] Arushi Gupta, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman. « Non-Gaussian information from weak lensing data via deep learning ». *Physical Review D* 97.10 (2018), p. 103515 (cf. p. 75).

[Guyon, 2022] Julien Guyon, Jordan Lekeufack. « Volatility is (mostly) path-dependent ». *Volatility Is (Mostly) Path-Dependent (July 27, 2022)* (2022) (cf. p. 14, 32, 115, 116, 121, 122, 128, 131, 161, 162).

[Ha, 2021] Wooseok Ha, Chandan Singh, Francois Lanusse, Srigokul Upadhyayula, Bin Yu. « Adaptive wavelet distillation from neural networks through interpretations ». *Advances in Neural Information Processing Systems* 34 (2021), p. 20669-20682 (cf. p. 63).

[Hammel, 1987] Stephen M Hammel, James A Yorke, Celso Grebogi. « Do numerical orbits of chaotic dynamical processes represent true orbits ? » *Journal of Complexity* 3.2 (1987), p. 136-145 (cf. p. 15, 32, 125).

[Hansen, 2008] Bruce E Hansen. « Uniform convergence rates for kernel estimation with dependent data ». *Econometric Theory* 24.3 (2008), p. 726-748 (cf. p. 15, 32, 126).

[Hasan, 2018] Ahmed M. Hasan, Ali Melli, Khan A. Wahid, Paul Babyn. « Denoising Low-Dose CT Images Using Multiframe Blind Source Separation and Block Matching Filter ». *IEEE Transactions on Radiation and Plasma Medical Sciences* 2.4 (2018), p. 279-287 (cf. p. 100).

[Hershey, 2016] John R. Hershey, Zhuo Chen, Jonathan Le Roux, Shinji Watanabe. « Deep Clustering : Discriminative Embeddings for Segmentation and Separation ». *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China : IEEE Press, 2016, p. 31-35 (cf. p. 100).

[Heston, 1993] Steven Heston. « A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options ». *Review of Financial Studies* 6 (1993), p. 327-343 (cf. p. 115, 116).

[Horleston, 2022] Anna C. Horleston, John F. Clinton, Savas Ceylan, Domenico Giardini, Constantinos Charalambous, Jessica C. E. Irving et al. « The Far Side of Mars : Two Distant Marsquakes Detected by InSight ». *The Seismic Record* 2.2 (2022), p. 88-99. eprint : `https://pubs.geoscienceworld.org/ssa/tsr/article-pdf/2/2/88/5592848/tsr-2022007.1.pdf` (cf. p. 101).

[Huber, 1981] Peter J. Huber. *Robust statistics*. Wiley Ser. Probab. Math. Stat. John Wiley & Sons, Hoboken, NJ, 1981 (cf. p. 67).

[Ibrahim, 2014] Amr Ibrahim, Mauricio D. Sacchi. « Simultaneous source separation using a robust Radon transform ». *GEOPHYSICS* 79.1 (2014), p. V1-V11 (cf. p. 100).

[InSight Marsquake Service, 2023] InSight Marsquake Service. *Mars Seismic Catalogue, InSight Mission ; V13 2023-01-01*. 2023 (cf. p. 101, 110, 111, 156-158).

[Jaffard, 2004] Stephane Jaffard. *Wavelet techniques in multifractal analysis*. Rapp. tech. PARIS UNIV (FRANCE), 2004 (cf. p. 5, 22, 37, 39, 40, 78).

[Jaffard, 2006] Stéphane Jaffard, Bruno Lashermes, Patrice Abry. « Wavelet leaders in multifractal analysis ». *Wavelet analysis and applications*. Springer, 2006, p. 201-246 (cf. p. 5, 22, 37).

[Jaffard, 2019] Stéphane Jaffard, Stéphane Seuret, Herwig Wendt, Roberto Leonarduzzi, Patrice Abry. « Multifractal formalisms for multivariate analysis ». *Proceedings of the Royal Society A* 475.2229 (2019), p. 20190150 (cf. p. 136).

[Jang, 2003] Gil-Jin Jang, Te-Won Lee. « A maximum likelihood approach to single-channel source separation ». *The Journal of Machine Learning Research* 4 (2003), p. 1365-1392 (cf. p. 100).

[Jaynes, 1957] E. T. Jaynes. « Information Theory and Statistical Mechanics ». *Physical Review* 106.4 (1957), p. 620-630 (cf. p. 2, 3, 19, 20, 63).

[Jeffrey, 2022] Niall Jeffrey, François Boulanger, Benjamin D. Wandelt, Bruno Regaldo-Saint Blancard, Erwan Allys, François Levrier. « Single frequency CMB B-mode inference with realistic foregrounds from a single training image ». 510.1 (2022), p. L1-L6. arXiv : `2111.01138 [astro-ph.CO]` (cf. p. 64, 101).

[Jutten, 2004] Christian Jutten, Massoud Babaie-Zadeh, Shahram Hosseini. « Three easy ways for separating nonlinear mixtures ? » *Signal Processing* 84.2 (2004), p. 217-229 (cf. p. 13, 30, 100).

[Jutten, 1991] Christian Jutten, Jeanny Herault. « Blind separation of sources, part I : An adaptive algorithm based on neuromimetic architecture ». *Signal Processing* 24.1 (1991), p. 1-10 (cf. p. 13, 30, 100).

[Kameoka, 2019] Hirokazu Kameoka, Li Li, Shota Inoue, Shoji Makino. « Supervised Determined Source Separation with Multichannel Variational Autoencoder ». *Neural Computation* 31.9 (2019), p. 1891-1914 (cf. p. 100).

[Ke, 2020] Shanfa Ke, Ruimin Hu, Xiaochen Wang, Tingzhao Wu, Gang Li, Zhongyuan Wang. « Single Channel Multi-Speaker Speech Separation Based on Quantized Ratio Mask and Residual Network ». *Multimedia Tools Appl.* 79.43–44 (2020), p. 32225-32241 (cf. p. 100).

[Kello, 2010] Christopher T. Kello, Gordon D.A. Brown, Ramon Ferrer-i-Cancho, John G. Holden, Klaus Linkenkaer-Hansen, Theo Rhodes, Guy C. Van Orden. « Scaling laws in cognitive sciences ». *Trends in Cognitive Sciences* 14.5 (2010), p. 223-232 (cf. p. 67).

[Khan, 2021] Amir Khan, Savas Ceylan, Martin van Driel, Domenico Giardini, Philippe Lognonné, Henri Samuel et al. « Upper mantle structure of Mars from InSight seismic data ». *Science* 373.6553 (2021), p. 434-438. eprint : `https://www.science.org/doi/pdf/10.1126/science.abf2966` (cf. p. 110).

[Khosravy, 2020] Mahdi Khosravy, Neeraj Gupta, Nilesh Patel, Nilanjan Dey, Naoko Nitta, Noboru Babaguchi. « Probabilistic Stone's Blind Source Separation with application to channel estimation and multi-node identification in MIMO IoT green communication and multimedia systems ». *Computer Communications* 157 (2020), p. 423-433 (cf. p. 100).

[Knapmeyer-Endrun, 2020] Brigitte Knapmeyer-Endrun, Taichi Kawamura. « NASA's InSight mission on Mars—first glimpses of the planet's interior from seismology ». *Nature Communications* 11.1 (2020), p. 1-4 (cf. p. 101).

[Knapmeyer-Endrun, 2021] Brigitte Knapmeyer-Endrun, Mark P. Panning, Felix Bissig, Rakshit Joshi, Amir Khan, Doyeon Kim et al. « Thickness and structure of the martian crust from InSight seismic data ». *Science* 373.6553 (2021), p. 438-443. eprint : `https://www.science.org/doi/pdf/10.1126/science.abf8966` (cf. p. 110).

[Kolmogorov, 1962] A. N. Kolmogorov. « A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number ». *Journal of Fluid Mechanics* 13.1 (1962), p. 82-85 (cf. p. 10, 27, 56).

[Kolmogorov, 1 941] Andrej Nikolaevich Kolmogorov. « On degeneration (decay) of isotropic turbulence in an incompressible viscous liquid ». *Dokl. Akad. Nauk SSSR*. T. 31. 1 941, p. 538-540 (cf. p. 10, 27, 55).

[Kolmogorov, 1941a] Andrey Nikolaevich Kolmogorov. « Dissipation of energy in the locally isotropic turbulence ». *Dokl. Akad. Nauk SSSR A*. T. 32. 1941, p. 16-18 (cf. p. 10, 27, 55).

[Kolmogorov, 1941b] Andrey Nikolaevich Kolmogorov. « The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers ». *Cr Acad. Sci. URSS* 30 (1941), p. 301-305 (cf. p. 10, 27, 55, 78).

[Kumar, 2015] Rajiv Kumar, Haneet Wason, Felix J. Herrmann. « Source separation for simultaneous towed-streamer marine acquisition — A compressed sensing approach ». *GEOPHYSICS* 80.6 (2015), WD73-WD88. eprint : `https://doi.org/10.1190/geo2015-0108.1` (cf. p. 100).

[Lakhal, 2023] Samy Lakhal, Laurent Ponson, Michael Benzaquen, Jean-Philippe Bouchaud. « Wrapping and unwrapping multifractal fields ». *arXiv preprint arXiv :2310.01927* (2023) (cf. p. 10, 26).

[Landau, 2013] Lev Davidovich Landau, Evgenii Mikhailovich Lifshitz. *Statistical Physics : Volume 5*. T. 5. Elsevier, 2013 (cf. p. 63).

[Lanford, 1975] O Lanford. « Time evolution of large classical systems ». *In Dynamical systems, theory and applications* (1975), p. 1-111 (cf. p. 3, 20).

[Lemarié, 1988] Pierre-Gilles Lemarié. « Ondelettes à localisation exponentielle ». *J. Math. Pures Appl.* 67 (1988), p. 227-236 (cf. p. 40, 118, 138).

[Leonarduzzi, 2018] Roberto Leonarduzzi, Patrice Abry, Herwig Wendt, Stéphane Jaffard, Hugo Touchette. « A generalized multifractal formalism for the estimation of nonconcave multifractal spectra ». *IEEE Transactions on Signal Processing* 67.1 (2018), p. 110-119 (cf. p. 37).

[Leonarduzzi, 2019] Roberto Leonarduzzi, Gaspar Rochette, Jean-Phillipe Bouchaud, Stéphane Mallat. « Maximum-entropy scattering models for financial time series ». *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, p. 5496-5500 (cf. p. 7, 58).

[Levin, 2017] David A Levin, Yuval Peres. *Markov chains and mixing times*. T. 107. American Mathematical Soc., 2017 (cf. p. 2, 20).

[Li, 2008] Yi Li, Eric Perlman, Minping Wan, Yunke Yang, Charles Meneveau, Randal Burns et al. « A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence ». *Journal of Turbulence* 9, N31 (2008), N31. arXiv : `0804.1703 [physics.flu-dyn]` (cf. p. 75).

[Lin, 2011] Zhouchen Lin, Risheng Liu, Zhixun Su. « Linearized alternating direction method with adaptive penalty for low-rank representation ». *Advances in neural information processing systems* 24 (2011) (cf. p. 92).

[Liu, 1989] Dong C. Liu, Jorge Nocedal. « On the limited memory BFGS method for large scale optimization ». *Mathematical Programming* 45.1 (1989), p. 503-528 (cf. p. 105).

[Liu, 2022] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, Björn W. Schuller. « Audio self-supervised learning : A survey ». *Patterns* 3.12 (2022), p. 100616 (cf. p. 13, 30, 100).

[Lognonné, 2020] Philippe Lognonné, William Bruce Banerdt, WT Pike, Domenico Giardini, U Christensen, Raphaël F Garcia et al. « Constraints on the shallow elastic and anelastic structure of Mars from InSight seismic data ». *Nature Geoscience* 13.3 (2020), p. 213-220 (cf. p. 101).

[Lombardo, 2014] F. Lombardo, E. Volpi, D. Koutsoyiannis, S. M. Papalexiou. « Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology ». *Hydrology and Earth System Sciences* 18.1 (2014), p. 243-255 (cf. p. 67).

[Lorenz, 1963] Edward N Lorenz. « Deterministic nonperiodic flow ». *Journal of atmospheric sciences* 20.2 (1963), p. 130-141 (cf. p. 10, 27).

[Lorenz, 2021] Ralph D. Lorenz, Aymeric Spiga, Philippe Lognonné, Matthieu Plasman, Claire E. Newman, Constantinos Charalambous. « The whirlwinds of Elysium : A catalog and meteorological characteristics of "dust devil" vortices observed by InSight on Mars ». *Icarus* 355 (2021), p. 114119 (cf. p. 101).

[Lustig, 1998] Rolf Lustig. « Microcanonical Monte Carlo simulation of thermodynamic properties ». *The Journal of Chemical Physics* 109.20 (1998), p. 8816-8828. eprint : https://pubs.aip.org/aip/jcp/article-pdf/109/20/8816/10794132/8816_1_online.pdf (cf. p. 2, 19).

[Mallat, 1999] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999 (cf. p. 4, 21, 40, 71, 91).

[Mallat, 2012] Stéphane Mallat. « Group invariant scattering ». *Communications on Pure and Applied Mathematics* 65.10 (2012), p. 1331-1398 (cf. p. 5, 23, 49, 64, 70, 71, 134).

[Mallat, 2020] Stéphane Mallat, Sixin Zhang, Gaspar Rochette. « Phase harmonic correlations and convolutional neural networks ». *Information and Inference : A Journal of the IMA* 9.3 (2020), p. 721-747 (cf. p. 7, 37, 43, 44, 46, 64, 70).

[Mandelbrot, 1963] Benoit B Mandelbrot. « The variation of certain speculative prices ». *The Journal of Business* 36.4 (1963), p. 394-419 (cf. p. 55).

[Mandelbrot, 1982] Benoit B Mandelbrot. *The fractal geometry of nature*. W. H. Freeman et Co., 1982 (cf. p. 37).

[Mandelbrot,] Benoit B Mandelbrot. « Multifractals and 1/f Noise Wild Self-Affinity in Physics (1963–1976) » () (cf. p. 37).

[Mandelbrot, 1997] Benoit B Mandelbrot, Adlai J Fisher, Laurent E Calvet. « A multifractal model of asset returns » (1997) (cf. p. 5, 8, 22, 25, 38, 39, 54, 146).

[Mandelbrot, 1968] Benoit B Mandelbrot, John W Van Ness. « Fractional Brownian motions, fractional noises and applications ». *SIAM review* 10.4 (1968), p. 422-437 (cf. p. 37, 48, 51, 54).

[Marchand, 2022] Tanguy Marchand, Misaki Ozawa, Giulio Biroli, Stéphane Mallat. « Wavelet Conditional Renormalization Group ». *arXiv e-prints*, arXiv :2207.04941 (2022), arXiv :2207.04941. arXiv : 2207.04941 [cond-mat.stat-mech] (cf. p. 72, 82, 147).

[Markowitz, 1952] Harry Markowitz. « Portfolio Selection ». *The Journal of Finance* 7.1 (1952), p. 77-91 (cf. p. 85).

[Masry, 1993] E. Masry. « The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion ». *IEEE Transactions on Information Theory* 39.1 (1993), p. 260-264 (cf. p. 40).

[Matilla, 2016] José Manuel Zorrilla Matilla, Zoltán Haiman, Daniel Hsu, Arushi Gupta, Andrea Petri. « Do dark matter halos explain lensing peaks ? » *Physical Review D* 94.8 (2016), p. 083506 (cf. p. 75).

[McCoy, 1996] EJ McCoy, AT Walden. « Wavelet analysis and synthesis of stationary long-memory processes ». *Journal of computational and Graphical statistics* 5.1 (1996), p. 26-56 (cf. p. 40).

[Mordant, 2002] N Mordant, J Delour, E Léveque, A Arnéodo, J-F Pinton. « Long time correlations in Lagrangian dynamics : a key to intermittency in turbulence ». *Physical review letters* 89.25 (2002), p. 254502 (cf. p. 54).

[Morel, 2023a] Rudy Morel, Stéphane Mallat, Jean-Philippe Bouchaud. « A maximum entropy factor model of financial stocks ». *in preparation* (2023) (cf. p. 17, 34).

[Morel, 2023b] Rudy Morel, Stéphane Mallat, Jean-Philippe Bouchaud. « Path Shadowing Monte-Carlo ». *arXiv preprint arXiv :2308.01486* (2023) (cf. p. 17, 34).

[Morel, 2022] Rudy Morel, Gaspar Rochette, Roberto Leonarduzzi, Jean-Philippe Bouchaud, Stéphane Mallat. « Scale Dependencies and Self-Similar Models with Wavelet Scattering Spectra ». *arXiv preprint arXiv :2204.10177* (2022) (cf. p. 17, 34).

[Muzy, 1994] Jean-François Muzy, Emmanuel Bacry, Alain Arneodo. « The multifractal formalism revisited with wavelets ». *International Journal of Bifurcation and Chaos* 4.02 (1994), p. 245-302 (cf. p. 5, 22, 37).

[Nadaraya, 1964] Elizbar A Nadaraya. « On estimating regression ». *Theory of Probability & Its Applications* 9.1 (1964), p. 141-142 (cf. p. 15, 32, 125).

[Nandi, 1996] A.K. Nandi, V. Zarzoso. « Fourth-order cumulant based blind source separation ». *IEEE Signal Processing Letters* 3.12 (1996), p. 312-314 (cf. p. 13, 30, 100).

[Neri, 2021] Julian Neri, Roland Badeau, Philippe Depalle. « Unsupervised Blind Source Separation with Variational Auto-Encoders ». *2021 29th European Signal Processing Conference (EUSIPCO)*. 2021, p. 311-315 (cf. p. 13, 30, 100).

[Olah, 2017] Chris Olah, Alexander Mordvintsev, Ludwig Schubert. « Feature Visualization ». *Distill* (2017) (cf. p. 79).

[Olshausen, 1996] Bruno A. Olshausen, David J. Field. « Emergence of simple-cell receptive field properties by learning a sparse code for natural images ». *Nature* 381.6583 (1996), p. 607-609 (cf. p. 71, 92).

[Oord, 2016] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves et al. « Wavenet : A generative model for raw audio ». *arXiv preprint arXiv :1609.03499* (2016) (cf. p. 56).

[Panning, 2023] M. P. Panning, W. B. Banerdt, C. Beghein, S. Carrasco, S. Ceylan, J. F. Clinton et al. « Locating the Largest Event Observed on Mars With Multi-Orbit Surface Waves ». *Geophysical Research Letters* 50.1 (2023), e2022GL101270. eprint : `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL101270` (cf. p. 101).

[Parisi, 1985] Giorgio Parisi, Uriel Frisch et al. « A multifractal model of intermittency ». *Turbulence and predictability in geophysical fluid dynamics and climate dynamics* (1985), p. 84-88 (cf. p. 10, 27).

[Parra, 2003] Lucas Parra, Paul Sajda. « Blind Source Separation via Generalized Eigenvalue Decomposition ». *J. Mach. Learn. Res.* 4.null (2003), p. 1261-1269 (cf. p. 13, 30, 100).

[Pedersen, 2008] Michael Syskind Pedersen, Jan Larsen, Ulrik Kjems, Lucas C. Parra. « Convolutive Blind Source Separation Methods ». *Springer Handbook of Speech Processing*. Sous la dir. de Jacob Benesty, M. Mohan Sondhi, Yiteng Arden Huang. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 1065-1094 (cf. p. 100).

[Perlman, 2007] Eric Perlman, Randal Burns, Yi Li, Charles Meneveau. « Data Exploration of Turbulence Simulations Using a Database Cluster ». *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing*. SC '07. Reno, Nevada : Association for Computing Machinery, 2007 (cf. p. 75).

[Pipiras, 2017] Vladas Pipiras, Murad S Taqqu. *Long-range dependence and self-similarity*. T. 45. Cambridge university press, 2017 (cf. p. 4, 21, 38-40).

[Pochart, 2002] Benoit Pochart, Jean-Philippe Bouchaud. « The skewed multifractal random walk with applications to option smiles ». *Quantitative finance* 2.4 (2002), p. 303 (cf. p. 54).

[Podesta, 2009] J. J. Podesta. « Dependence of Solar-Wind Power Spectra on the Direction of the Local Mean Magnetic Field ». 698.2 (2009), p. 986-999. arXiv : `0901.4940 [astro-ph.EP]` (cf. p. 63).

[Portilla, 2000] Javier Portilla, Eero P Simoncelli. « A parametric texture model based on joint statistics of complex wavelet coefficients ». *International journal of computer vision* 40.1 (2000), p. 49-70 (cf. p. 6, 23, 37, 43, 70, 134).

[Potters, 2005] Marc Potters, Jean-Philippe Bouchaud, Laurent Laloux. « Financial applications of random matrix theory : Old laces and new pieces ». *arXiv preprint physics/0507111* (2005) (cf. p. 12, 29, 85).

[Potters, 2001] Marc Potters, Jean-Philippe Bouchaud, Dragan Sestovic. « Hedged Monte-Carlo : low variance derivative pricing with objective probabilities ». *Physica A : Statistical Mechanics and its Applications* 289.3 (2001), p. 517-525 (cf. p. 121, 129, 130).

[Raina, 2007] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, Andrew Y Ng. « Self-taught learning : transfer learning from unlabeled data ». *Proceedings of the 24th international conference on Machine learning*. 2007, p. 759-766 (cf. p. 92).

[Ranftl, 2021] René Ranftl, Alexey Bochkovskiy, Vladlen Koltun. « Vision transformers for dense prediction ». *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, p. 12179-12188 (cf. p. 136).

[Regaldo-Saint Blancard, 2021] Bruno Regaldo-Saint Blancard, Erwan Allys, François Boulanger, François Levrier, Niall Jeffrey. « A new approach for the statistical denoising of Planck interstellar dust polarization data ». *Astronomy & Astrophysics* 649 (2021), p. L18 (cf. p. 13, 14, 30, 31, 64, 82, 101, 134).

[Régaldo-Saint Blancard, 2023] Bruno Régaldo-Saint Blancard, Erwan Allys, Constant Auclair, François Boulanger, Michael Eickenberg, François Levrier, Léo Vacher, Sixin Zhang. « Generative Models of Multichannel Data from a Single Example—Application to Dust Emission ». *The Astrophysical Journal* 943.1 (2023), p. 9 (cf. p. 12, 29, 64, 82, 136).

[Reigneron, 2011] Pierre-Alain Reigneron, Romain Allez, Jean-Philippe Bouchaud. « Principal regression analysis and the index leverage effect ». *Physica A : Statistical Mechanics and its Applications* 390.17 (2011), p. 3026-3035 (cf. p. 12, 29, 85, 88, 89).

[Renaud, 2005] Olivier Renaud, J-L Starck, Fionn Murtagh. « Wavelet-based combined signal filtering and prediction ». *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 35.6 (2005), p. 1241-1251 (cf. p. 16, 33).

[Renaud, 2003] Olivier Renaud, Jean-Luc Starck, Fionn Murtagh. « Prediction based on a multiscale decomposition ». *International Journal of Wavelets, Multiresolution and Information Processing* 1.02 (2003), p. 217-232 (cf. p. 16, 33).

[Rodrıguez, 2021] Ángel Bueno Rodriguez, Randall Balestriero, Silvio De Angelis, M Carmen Benitez, Luciano Zuccarello, Richard Baraniuk, Jesus M Ibanez, Maarten de Hoop. « Recurrent Scattering Network detects metastable behavior in polyphonic seismo-volcanic signals for volcano eruption forecasting ». *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), p. 1-23 (cf. p. 5, 23).

[Saës, 2022] Guillaume Saës, Wejdene Ben Nasr, Stéphane Jaffard, Florent Palacin, Véronique Billat. « Multifractal analysis of physiological data from marathon runners ». *arXiv preprint arXiv :2206.08140* (2022) (cf. p. 5, 22).

[Salvi, 2021] Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, Terry Lyons. « Higher order kernel mean embeddings to capture filtrations of stochastic processes ». *Advances in Neural Information Processing Systems* 34 (2021), p. 16635-16647 (cf. p. 115).

[Saydjari, 2021] Andrew K Saydjari, Stephen KN Portillo, Zachary Slepian, Sule Kahraman, Blakesley Burkhart, Douglas P Finkbeiner. « Classification of magnetohydrodynamic simulations using wavelet scattering transforms ». *The Astrophysical Journal* 910.2 (2021), p. 122 (cf. p. 64).

[Schneider, 2006] Kai Schneider, Jörg Ziuber, Marie Farge, Alexandre Azzalini. « Coherent vortex extraction and simulation of 2D isotropic turbulence ». *Journal of Turbulence* 7 (2006), N44 (cf. p. 75).

[Scholz, 2020] John-Robert Scholz, Rudolf Widmer-Schnidrig, Paul Davis, Philippe Lognonné, Baptiste Pinot, Raphaël F. Garcia et al. « Detection, Analysis, and Removal of Glitches From InSight's Seismic Data From Mars ». *Earth and Space Science* 7.11 (2020), e2020EA001317 (cf. p. 100-102, 105, 108, 109, 153, 157).

[Seydoux, 2020] Léonard Seydoux, Randall Balestriero, Piero Poli, Maarten de Hoop, Michel Campillo, Richard Baraniuk. « Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning ». *Nature communications* 11.1 (2020), p. 3972 (cf. p. 5, 23).

[Shannon, 1948] Claude Elwood Shannon. « A mathematical theory of communication ». *The Bell system technical journal* 27.3 (1948), p. 379-423 (cf. p. 3, 20).

[Sherman, 2018] Michael Sherman. « Variance Estimation for Statistics Computed from Spatial Lattice Data ». *Journal of the Royal Statistical Society : Series B (Methodological)* 58.3 (2018), p. 509-523. eprint : `https://academic.oup.com/jrsssb/article-pdf/58/3/509/49100624/jrsssb\_58\_3\_509.pdf` (cf. p. 73).

[Siahkoohi, 2023a] Ali Siahkoohi, Rudy Morel, Randall Balestriero, Erwan Allys, Grégory Sainton, Taichi Kawamura, Maarten de Hoop. « Martian time-series unraveled : A multi-scale nested approach with factorial variational autoencoders ». *arXiv preprint arXiv :2305.16189* (2023) (cf. p. 17, 34).

[Siahkoohi, 2023b] Ali Siahkoohi, Rudy Morel, Maarten de Hoop, Erwan Allys, Grégory Sainton, Taichi Kawamura. « Unearthing InSights into Mars : unsupervised source separation with limited data ». *International Conference on Machine Learning*. PMLR. 2023, p. 31754-31772 (cf. p. 17, 34).

[Siggia, 1978] Eric D Siggia. « Model of intermittency in three-dimensional turbulence ». *Physical review A* 17.3 (1978), p. 1166 (cf. p. 10, 27).

[Sinai, 1972] Yakov G Sinai. « Gibbs measures in ergodic theory ». *Russian Mathematical Surveys* 27.4 (1972), p. 21 (cf. p. 125).

[Sklar, 1959] M Sklar. « Fonctions de répartition à n dimensions et leurs marges ». *Annales de l'ISUP*. T. 8. 3. 1959, p. 229-231 (cf. p. 95).

[Stähler, 2021] Simon C. Stähler, Amir Khan, W. Bruce Banerdt, Philippe Lognonné, Domenico Giardini, Savas Ceylan et al. « Seismic detection of the martian core ». *Science* 373.6553 (2021), p. 443-448. eprint : `https://www.science.org/doi/pdf/10.1126/science.abi7730` (cf. p. 110).

[Starck, 2004] Jean-Luc Starck, DL Donoho, Michael Elad. *Redundant multiscale transforms and their application for morphological component separation*. Rapp. tech. CM-P00052061, 2004 (cf. p. 13, 30, 100).

[Stein, 1956] Charles Stein. « Inadmissibility of the usual estimator for the mean of a multivariate normal distribution ». *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*. University of California Press, Berkeley-Los Angeles, Calif., 1956, p. 197-206 (cf. p. 73).

[Stott, 2022] A E Stott, R F Garcia, A Chédozeau, A Spiga, N Murdoch, B Pinot et al. « Machine learning and marsquakes : a tool to predict atmospheric-seismic noise for the NASA InSight mission ». *Geophysical Journal International* 233.2 (2022), p. 978-998. eprint : `https://academic.oup.com/gji/article-pdf/233/2/978/48755380/ggac464.pdf` (cf. p. 101).

[Tumminello, 2007] Michele Tumminello, Fabrizio Lillo, Rosario Nunzio Mantegna. « Shrinkage and spectral filtering of correlation matrices : a comparison via the Kullback-Leibler distance ». *arXiv preprint arXiv :0710.0576* (2007) (cf. p. 12, 29, 85).

[Ustyuzhaninov, 2017] Ivan Ustyuzhaninov, Wieland Brendel, Leon Gatys, Matthias Bethge. « What does it take to generate natural textures ? » *International Conference on Learning Representations*. 2017 (cf. p. 6, 23).

[Valogiannis, 2022] Georgios Valogiannis, Cora Dvorkin. « Towards an optimal estimation of cosmological parameters with the wavelet scattering transform ». *Physical Review D* 105.10 (2022), p. 103534 (cf. p. 64).

[Vargas, 2015] Vincent Vargas, Tung-Lam Dao, Jean-Philippe Bouchaud. « Skew and implied leverage effect : smile dynamics revisited ». *International Journal of Theoretical and Applied Finance* 18.04 (2015), p. 1550022 (cf. p. 117, 123).

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin. « Attention is all you need ». *Advances in neural information processing systems* 30 (2017) (cf. p. 115, 136).

[Vielva, 2004] P. Vielva, E. Martinez-González, R. B. Barreiro, J. L. Sanz, L. Cayón. « Detection of Non-Gaussianity in the Wilkinson Microwave Anisotropy Probe First-Year Data Using Spherical Wavelets ». 609.1 (2004), p. 22-34. arXiv : `astro-ph/0310273 [astro-ph]` (cf. p. 63).

[Villaescusa-Navarro, 2020] Francisco Villaescusa-Navarro, ChangHoon Hahn, Elena Massara, Arka Banerjee, Ana Maria Delgado, Doogesh Kodi Ramanah et al. « The Quijote Simulations ». 250.1, 2 (2020), p. 2. arXiv : `1909.05273 [astro-ph.CO]` (cf. p. 10, 26, 75).

[Wang, 2018] DeLiang Wang, Jitong Chen. « Supervised speech separation based on deep learning : An overview ». *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018), p. 1702-1726 (cf. p. 100).

[Watson, 1964] Geoffrey S Watson. « Smooth regression analysis ». *Sankhyā : The Indian Journal of Statistics, Series A* (1964), p. 359-372 (cf. p. 15, 32, 125).

[Wen, 2022] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, Liang Sun. « Transformers in time series : A survey ». *arXiv preprint arXiv :2202.07125* (2022) (cf. p. 115, 136).

[Wendt, 2009] Herwig Wendt, Stéphane G Roux, Stéphane Jaffard, Patrice Abry. « Wavelet leaders and bootstrap for multifractal analysis of images ». *Signal Processing* 89.6 (2009), p. 1100-1114 (cf. p. 5, 22, 37).

[, 2017] *Why Stock Markets Crash : Critical Events in Complex Financial Systems*. REV - Revised. Princeton University Press, 2017 (cf. p. 67).

[Wisdom, 2020] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, John Hershey. « Unsupervised Sound Separation Using Mixture Invariant Training ». *Advances in Neural Information Processing Systems*. T. 33. Curran Associates, Inc., 2020, p. 3846-3857 (cf. p. 13, 30, 100).

[Wornell, 1993] G.W. Wornell. « Wavelet-based representations for the 1/f family of fractal processes ». *Proceedings of the IEEE* 81.10 (1993), p. 1428-1450 (cf. p. 6-8, 24, 40, 41, 49).

[Wu, 2022] Peng Wu, Jean-François Muzy, Emmanuel Bacry. « From rough to multifractal volatility : The log S-fBM model ». *Physica A : Statistical Mechanics and its Applications* 604 (2022), p. 127919 (cf. p. 115).

[Xu, 2022] Wuchuan Xu, Qiwen Zhu, Li Zhao. « GlitchNet : A Glitch Detection and Removal System for SEIS Records Based on Deep Learning ». *Seismological Research Letters* 93.5 (2022), p. 2804-2817 (cf. p. 101).

[Y Meyer, 1992] Y. Meyer. *Wavelets and Operators*. Advanced mathematics. Cambridge university press, 1992 (cf. p. 4, 21, 144).

[Zarka, 2019] John Zarka, Louis Thiry, Tomás Angles, Stéphane Mallat. « Deep network classification by scattering and homotopy dictionary learning ». *arXiv preprint arXiv :1910.03561* (2019) (cf. p. 92).

[Zhang, 2021] Sixin Zhang, Stéphane Mallat. « Maximum entropy models from phase harmonic covariances ». *Applied and Computational Harmonic Analysis* 53 (2021), p. 199-230 (cf. p. 11, 27, 43, 44, 46, 64, 70, 134).

[Zhu, 1997] Song Chun Zhu, Ying Nian Wu, David Mumford. « Minimax entropy principle and its application to texture modeling ». *Neural computation* 9.8 (1997), p. 1627-1660 (cf. p. 65).

[Zhu, 1998] Song Chun Zhu, Yingnian Wu, David Mumford. « Filters, random fields and maximum entropy (FRAME) : Towards a unified theory for texture modeling ». *International Journal of Computer Vision* 27.2 (1998), p. 107-126 (cf. p. 65).

[Zumbach, 2009] Gilles Zumbach. « Time reversal invariance in finance ». *Quantitative Finance* 9.5 (2009), p. 505-515 (cf. p. 54, 55, 59).

RÉSUMÉ

Les processus multi-échelles, qui présentent des variations sur une large gamme d'échelles, sont présents en physique, finance, biologie, médecine et de nombreux autres domaines. L'objectif principal de cette thèse est de construire des modèles probabilistes de tels processus observés à partir de peu de données et pouvant être échantillonnés numériquement. Ce sujet est crucial pour aborder plusieurs problèmes, notamment la génération, la prédiction et les problèmes inverses tels que la séparation de sources.

Dans cette thèse, nous introduisons les spectres en Scattering («Scattering Spectra»), qui sont basés sur une approximation diagonale de corrélations non linéaires de coefficients d'ondelettes. Ils peuvent être utilisés pour construire des modèles non-Gaussiens de processus multi-échelles, qu'il s'agisse de processus temporels, de processus temporels multi-canaux ou de processus d'image. Nous montrons qu'ils reproduisent des propriétés statistiques importantes de séries temporelles financières, de jets turbulents et de champs physiques.

Nous démontrons que cette représentation en «Scattering Spectra» peut être utilisée pour la séparation de sources à partir de peu de données. Appliquée aux données sismiques sur Mars, ils permettent de séparer avec succès les tremblements de Mars d'événements polluants transitoires appelés «Glitches».

La prédiction sur données limitées peut être abordée en utilisant un modèle précis du processus capable de capturer les dépendances à long terme. Nous introduisons le «Path-Shadowing Monte-Carlo» qui est une méthode à noyau non-locale qui propose de moyenner les quantités futures sur des chemins générés dont l'histoire passée est «proche» de l'histoire réelle (observée). Associée à un modèle basé sur les «Scattering Spectra», cette approche permet d'obtenir des résultats à l'état de l'art pour la prédiction de volatilité en finance et fournit des smiles d'option qui surpassent le marché dans un jeu de trading.

MOTS CLÉS

modélisation, multi-échelle, apprentissage non-supervisé, traitement du signal

ABSTRACT

Multi-scale processes, which have variations on a wide-range of scales, are encountered in Physics, Finance, Biology, Medicine and various other fields. The core purpose of this thesis is to construct probabilistic models of such processes observed from limited data and that can be sampled numerically. Such subject is crucial to tackle a number of problems among which generation, prediction and inverse problems such as source separation.

In this thesis we introduce wavelet Scattering Spectra which rely on a diagonal approximation of non-linear correlations of wavelet coefficients. They can be used to construct non-Gaussian models of multi-scale processes, including time-processes, multi-channel time-processes and image processes. Scattering Spectra are shown to capture important statistical properties of financial time-series, turbulent jet and physical fields.

We show that such Scattering Spectra representation can be used to perform source separation on limited data. Applied on Mars seismic data, we are able to successfully separate Marsquakes from transient polluting events called Glitches.

Prediction on limited data can be tackled by utilizing an accurate model of the process which captures long-range dependencies. We introduce Path-Shadowing Monte-Carlo, a non-local kernel method which proposes to predict future quantities by averaging over generated paths whose past history "shadows" the actual (observed) history. When combined with our Scattering Spectra model, this approach yields state-of-the-art volatility prediction in Finance and provides option smiles that outperform the market in a designed trading game.

KEYWORDS

modelling, multi-scale, unsupervised learning, signal processing