A few key issues in finance that machine learning is helping solve

Pierre Brugière and Gabriel Turinici

CEREMADE, University Paris Dauphine-PSL

brugiere@ceremade.dauphine.fr, turinici@ceremade.dauphine.fr

November 29, 2022

Overview

1. Introduction
2. Overfitting and Outlayers
3. Sensitivity and Data mining
4. Model Risk and Simulations

Introduction

- Statistical learning and machine learning provide theoretical answers to several fundamental data analysis questions (on top of providing implementable solutions).
- A few examples of these key data analysis or modelling issues that machine learning addresses are presented here.
- For the part related to Generative Models I thanks my colleague Gabriel Turinici for his help and for providing useful insights and material.

Introduction

The topics discussed here concern:

- **Overfitting** i.e the problem of choosing a model which calibrates well but with poor prediction performance.
- **Outlayers** i.e when some observations, legitimate or illegitimate, jeopardise the estimation process.
- **Sensitivity** i.e when the ouput is very sensitive to the input parameters, as in Markowitz portfolio optimisation.
- "**Data mining** or "**False Discovery**" i.e when so many assets are analysed that one emerges by pure luck and not by merit.
- **Model risk and Simulation** i.e the risk of using the wrong model or the wrong scenarios to analyse a situation.

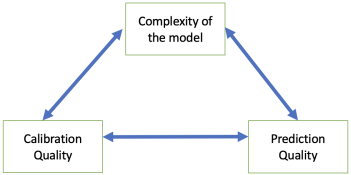
Overfitting and Outlayers

Overfitting and Outlayers

- Calibrating is not always easy, but predicting is even harder....
- For a long time there have been simple rules to guide the beginner practitioner (as saying that for a linear model there should be at least 30 observations per explanatory variable).
- The notion of **complexity of a model** from Vapnik and Chervonenkis (VC) links : the quality of the calibration, the complexity of the model and the expected quality of the prediction in a very universal and elegant way.
- For a classification problem the relationship between the three is as follows:

Overfitting and Outlayers

$$P\Big(E(\text{error on prediction}) > \text{error on calibration} + \phi_\eta(\frac{VC}{n})\Big) < 1 - \eta$$



Vapnik-Chervonenkis relationship

This gave birth to Support Vector Machines (SVM), which address the problem of complexity and robustness control as well as outlayers management.

Overfitting and Outlayers

We illustrate these characteristics when estimating the Probability of Default (PD) of a company, by showing some key differences between:

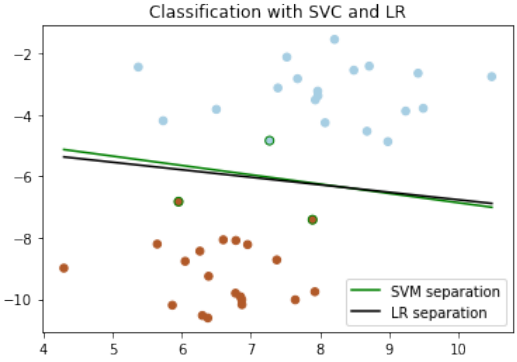
- SVMs (in their simplest formulation) and
- Logistic Regressions (which are accepted by the regulator).

For both models the inputs are some key financial ratios and the aim is to predict the likelihood of a default at a 12 month horizon.
The calibration is done on a large sample.

For Logistic Regression:

- The PD estimate is calculated as a sigmoid of a linear combination of the ratios.
- The linear coefficients are estimated by maximising the likelihood of the sample.

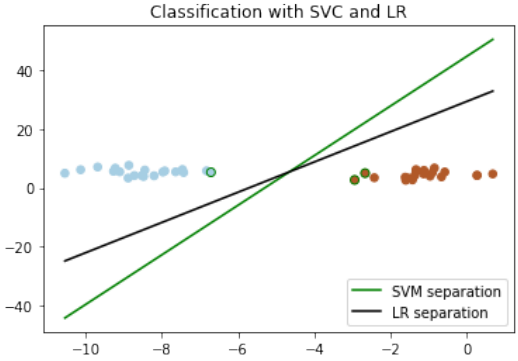
Therefore, the PD for a company is estimated by the distance to an hyperplane on which the PD is estimated at 0.5.



The Probability of Default is estimated at 0.5 on the line

For SVMs:

- There is also an hyperplane, on which the PD is estimated at 0.5, and for other points the PD is linked to the distance to this hyperplane.
- This time the hyperplane is defined with the principle of separating the points on each side with a maximum margin.
- This maximum margin technique enables to control the complexity of the model and therefore the quality of its prediction and its robusteness.
- It is also less sensitive to outlayers than logistic regression as most remote points are not taken into account when defining the maximum margin hyperplane.



The Probability of Default is estimated at 0.5 on the line

Overfitting and Outlayers

- SVMs became well known in 1998 for winning the competition of the MNIST database of handwritten digits recognition.
- It is a natural alternative approach to logistic regression or to any method using scores.
- It addresses overfitting and robustness questions in an elegant way.
- It does not systematically outperform logistic regression models.
- It is more complex to use than logistic regression because of the flexibility to choose the Kernel and the hyper-parameters.
- The current view of the European Banking Authority on the use of this model or other advanced ML techniques is as follow:

Overfitting and Outlayers

European Banking Authority (EBA) stance on ML technics [5]:

- The logistic regression model is currently accepted as an internal ratings-based (IRB) model for PD evaluation.
- The direct use of ML models (other than linear and logistic regression) is currently prohibited to calculate regulatory capital requirements under the IRB approach.
- A consultation on the topic has been launched by the EBA on February 11th 2022. The situation may evolve in the future and particularly for the evaluation of the PD.
- According to a review of the EBA in 2021 ML techniques are applied to loan origination, credit monitoring and restructuring.

Sensitivity and Data mining

Sensitivity and Data mining

We illustrate the issues with two examples.

- The first one concerns mean-variance **portfolio optimisation**.
- When examining a large number of stocks some very high correlations are likely to appear only by chance.
- This leads to the integration of large pair trading strategies, without any merit, inside the portfolio.
- It also induces variance-covariance matrices with very small eigenvalues producing unstable allocations.
- The second one concerns **fund selection**.
- Some funds may exhibit high performances on a given period only by pure luck and at the same time overshadow the real skilled ones.
- This is a problem both of False Discovery and inflated estimated skills.

Some ML Bayesian technics help solving these problem as well as some ad-hoc approaches.

Exemple 1: Markowitz's portfolio.
We solve here:

$$\arg\max_{\pi} E(R(\pi)) - \lambda V(R(\pi))$$

where $R(\pi)$ is the excess return of the portfolio of risky allocation π.
Fundamental issues arising are:

- The estimation $\hat{\Sigma}$ of the matrix of variance-covariance.
- The invertibility of $\hat{\Sigma}$.
- The implementability and stability of the allocation found.

These issues get dramatic when the number of assets considered is large.

To estimate $\hat{\Sigma}$ for the S&P500 with the basic rule of 30 independent observations by covariance estimate $30 \times \frac{d+1}{2}$ time points would be needed i.e approx 29 years of daily historic!

Sensitivity and Data mining

Also, with large covariance matrices many very high correlations are likely to be found between some stocks or group of stocks by pure luck (data mining) leading to a distortion of the optimisation process (and with less than 501 data points $\hat{\Sigma}$ would not be invertible).

To remedy this problem various solutions can be considered :

- Factor models (Fama, French & All) which by assuming the independence, after factor decomposition, of the residuals noises produce de facto (amongst other things) invertible $\hat{\Sigma}$ matrices.
- Bayesian methods for which "significant proof" must be given by the data before the prior parameters of the model (with adequate covariance matrices) are changed significantly.
- Some simple ad-hoc methods which can be linked to Bayesian methods.

Sensitivity and Data mining

A Bayesian method with conjugate priors on μ and Σ would be :

- $R|(\mu, \Sigma) \sim \mathcal{N}(\mu, \Sigma)$ (R the vector of returns).
- $\mu|\Sigma \sim \mathcal{N}(\mu_0, \frac{1}{\tau_0}\Sigma)$ (τ_0 precision parameter)
- $\Sigma \sim IW(V_0, \nu_0)$ (inverted Wishart with ν_0 degrees of freedom)

and from this we get :

$$\Sigma^{-1} \sim W(V_0^{-1}, \nu_0) \text{ and } E(\Sigma^{-1}) = \nu_0 V_0^{-1}$$

After observing the returns of the N risky assets over T time steps, the updated parameters become :

$$\mu_1 = \frac{\tau_0}{T + \tau_0}\mu_0 + \frac{T}{T + \tau_0}\hat{\mu} \text{ and } \tau_1 = \tau_0 + T$$

$$V_1 = V_0 + T\hat{\Sigma} + \frac{T\tau_0}{T + \tau_0}(\mu_0 - \hat{\mu})(\mu_0 - \hat{\mu})' \text{ and } \nu_1 = \nu_0 + T$$

Sensitivity and Data mining

So, before the observations, the matrix to use to calculate Markowitz portfolios is:

$$\Sigma_0^{-1} = \nu_0 V_0^{-1}$$

and after observations (if we omit the adjustment terms for μ):

$$\Sigma_1^{-1} = (\nu_0 + T)(V_0 + T\hat{\Sigma})^{-1} = \Big(\frac{T}{\nu_0 + T}\hat{\Sigma} + \frac{\nu_0}{\nu_0 + T}\Sigma_0\Big)^{-1}$$

So, the Bayesian approach has the benefit of controlling the invertibility and stability of the variance-covariance matrix to use. It is also possible to choose the parameters of the priors so that the prior optimal portfolio is equally weighted or match the benchmark. New data evidence will then lead to adjustments in the assets held.

Sensitivity and Data mining

A simple way to reproduce the "Bayesian effects" is to integrate some model uncertainty such as:

$$R_{future} = \sqrt{1-\delta^2} R_{past} + \delta\epsilon$$

$$\text{with } \epsilon \sim \mathcal{N}(0, \Sigma_0) \text{ independent from } R_{past}$$

this leads to a variance-covariance matrix for R_{future} equal to :

$$\Sigma_f = (1-\delta^2)\hat{\Sigma} + \delta^2\Sigma_0$$

with deflated correlation terms compared to the ones deriving from $\hat{\Sigma}$ (and guaranteed invertibility).

Sensitivity and Data mining

For example:
If $\epsilon \sim \mathcal{N}(0, diag(\sigma_i^2))$ we get $\forall i \neq j, \rho^*_{i,j} = (1-\delta^2)\rho_{i,j}$
If $\epsilon \sim \mathcal{N}(0, [\rho\sigma_i\sigma_j])$ we get $\forall i \neq j, \rho^*_{i,j} = (1-\delta^2)\rho_{i,j} + \delta^2\rho$

We illustrate below this stabilisation effect for an allocation on the DAX.

- The expected returns are taken from an average of analyst target prices.
- The correlation matrix $(\rho_{i,j})$ is estimated historically.
- The volatilities σ_i are estimated historically (even if for real investments it would be better to use implied volatilities).
- The variance-covariance matrices used for portfolio constructions are the $(\rho_{i,j}\sigma_i\sigma_j)$ and $(\rho^*_{i,j}\sigma_i\sigma_j)$.

Sensitivity and Data mining

	Sharpe	alloc delta =0	alloc delta =0.5	alloc delta =1
ALV.DE	0.42	0.23	-0.06	0.03
HNR1.DE	-0.05	-0.58	-0.38	-0.00
BEI.DE	0.21	-0.13	-0.13	0.02
ADS.DE	0.12	-0.74	-0.32	0.01
DBK.DE	0.48	-0.06	-0.01	0.02
SIE.DE	0.41	-0.17	-0.12	0.02
FRE.DE	1.36	0.87	0.58	0.09
BAS.DE	0.27	-0.21	-0.14	0.02
DPW.DE	0.95	0.30	0.24	0.05
FME.DE	0.53	-0.27	-0.08	0.03

DAX 30 - Allocation of the portfolio with the maximum Sharpe Ratio (first 10 components)

- The max Sharpe ratio for the 30 stocks is 1.4
- The Sharpe ratios obtained for the portfolios are: 3.37 (for $\delta = 0$), 2.82 for ($\delta = 0.5$) and 3.64 for ($\delta = 1$).

Sensitivity and Data mining

Example 2: Fund selection.
Some important literature exists concerning the detection of funds with:

- The best alpha.
- The best Sharpe ratio (or excess Sharpe ratio to a Benchmark).

The issues are:

- To differentiate skills from luck (problem of data mining).
- To calculate the expected performance parameter (alpha, Sharpe ratio or other) of the selected fund in the future.

Sensitivity and Data mining

We consider $d+1$ funds

- d with no skill, for which the excess (log) performances R_t^i to the benchmark satisfies $R_t^i \sim N(0, \sigma\sqrt{t})$.
- One skilled, for which the the excess performance R_t^* satisfies $R^* \sim N(mt, \sigma\sqrt{t})$. (with $m > 0$)

The pitfalls to avoid are:

- To pick a fund which is not the skilled one (false discovery) and
- to have inflated expectations of what future excess returns will be.

The first issue is linked to the fact that as d increases:

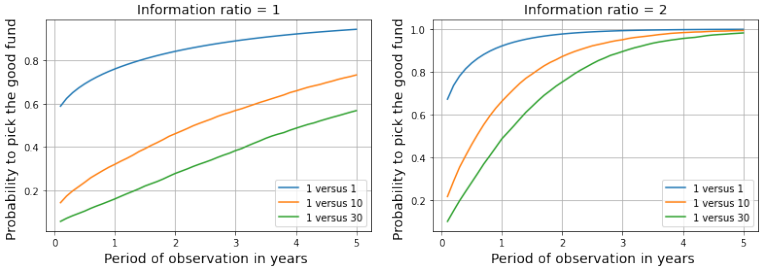
$$P(R_t^* < \max_{i \in [\![1,d]\!]} R_t^i) \text{ increases}$$

The second to the fact that as d increases:

$$E(\max_{i \in [\![1,d]\!]} R_t^i) - \max_{i \in [\![1,d]\!]} E(R_t^i) \text{ increases}$$

Sensitivity and Data mining

False Discovery:



Comments on the Discovery Problem

The "True Discovery" probability is determined by:

- The information ratio $\frac{m}{\sigma}$.
- The number of (independent) funds in the sample analysed.
- The length of the observation period.

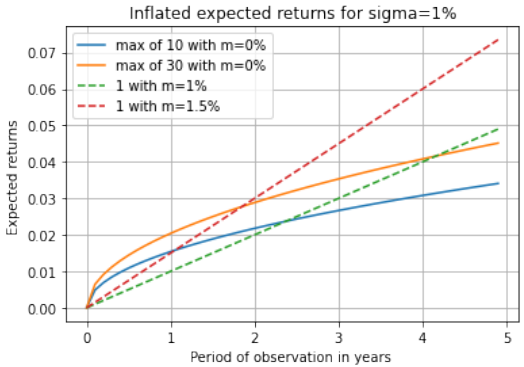
If the funds have a correlation ρ between them, the problem is mathematically equivalent to the problem with independent funds but with the true Information Ratio multiplied by the factor $\frac{1}{\sqrt{1-\rho^2}}$ as

$$P(mt + \sigma\sqrt{N} + \sqrt{1-\rho^2}\sigma Z > \max_{i \in [\![1,d]\!]} \sigma\sqrt{N} + \sqrt{1-\rho^2}\sigma Z_i)$$

$$P\Big(\frac{m}{\sigma\sqrt{1-\rho^2}}t + Z > \max_{i \in [\![1,d]\!]} Z_i\Big)$$

Sensitivity and Data mining

Inflated expected returns:



Comments on the inflated expected returns :

- For the unskilled funds $E(R_t^i) = 0$ but $E(\max_{i \in [\![1,d]\!]} R_t^i) > 0$
- In the model used here $E(\max_{i \in [\![1,d]\!]} R_t^i) > 0 = \lambda_d \sqrt{t}$ with
  - for 10 funds $\lambda_{10} = 1.54$
  - for 30 funds $\lambda_{30} = 2.04$
- Somebody picking the top performing fund out of 30, after 1 year of observation, is likely to pick a lucky unskilled fund and see the next year:
  - its excess performance (to the benchmark) going from 2% to an expected value of zero.
  - its ranking going from 1^{st} to an expected 15^{th}.
- When looking at correlated unskilled funds the inflation effect is shrunked by a factor $\sqrt{1-\rho^2}$.

Sensitivity and Data mining

So, how to find the best fund?

- In most cases the only possible thing to wish for is to get an estimate of the likelihood to get it right.
- When specifying the funds and particular performance indicators considered (alpha, Sharpe or Information Ratio) more sophisticated models can be built.
- Amongst them, some Bayesian models with some priors on the distribution of the skills and with a decision function (strategy) determined by minimising a Bayesian risk.

Model Risk and Simulations

Model Risk and Simulations

- Model risk is always an issue in finance.
- The default of AAA CDOs priced with Gaussian Copulas in 2008 illustrates this (as well as the LTCM debacle).
- There is an effort to devise models with as little assumptions as possible on the underlying.
- In this context some stochastic optimisation problems with strong model assumptions (such as in option pricing) see alternative approaches being developed with reinforcement learning.
- Most of these models need more data than available and require the creation of synthetic data.
- Here we look at this problem through the particular angle of *VaR* analysis.

Model Risk and Simulations

- According to BIS rules, a bank must meet, on a daily basis, some capital requirements based on some *VaR* and stressed *VaR* measures.
- The *VaR* part (c_t) is defined as (MAR 30.15 [6]):

$$c_t = max(Var^{10}_{t-1,99\%}, \frac{m_t}{60}\sum_{i=1}^{60} Var^{1}_{t-i,99\%})$$

where m_t is at least 3 and at most 4 depending on the number of breaches of the daily *VaR*.

- The 10 days *VaR* is the "regulatory" *VaR* while the 1 day *VaR* is an "in-house" *VaR* adjusted based on its performances.

Model Risk and Simulations

- These quantities are calculated for whole portfolios (equity, currency, commodity..) and for positions which are linear or non linear (options).
- There are several approaches to calculate these *VaR* indicators.
- Very often banks focus on the calculation of the 1 day *VaR* and the 10 days *VaR* is derived from a model based formula by multiplication by a factor $\sqrt{10}$.
- Sometimes as well the 1 day *VaR* 99% is derived from a 1 day *VaR* 95% by multiplication by a factor $\frac{2.33}{1.65}$, assuming that for the data considered the ratio would be the same as for a Gaussian distribution.

We can distinguish three main approaches.

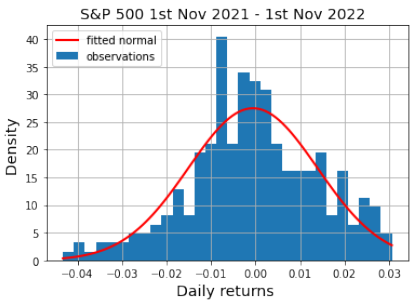
Model Risk and Simulations

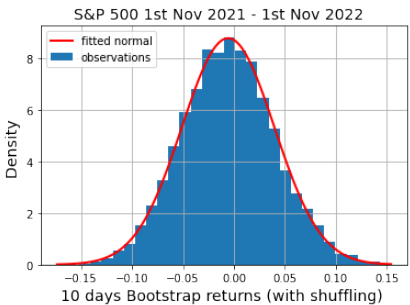
- **Model based:** For Gaussian models the tail risk may be underestimated and for (conditional) Gaussian models or Extreme Value models there is still a model risk. The Monte Carlo simulations based on these models suffer from the same model risk.
- **Historical based:** There is no model assumption but the number of scenarios, taken from historical time windows, may look small. The Bootstrap method can generate additional scenarios but: may disregards some phenomenon such as volatility clustering, cannot extrapolate or invent scenarios that have not already occurred, may still generate a number of scenarios which is too small.
- **Scenario based:** Very subjective and limited number of scenarios generated.

Model Risk and Simulations

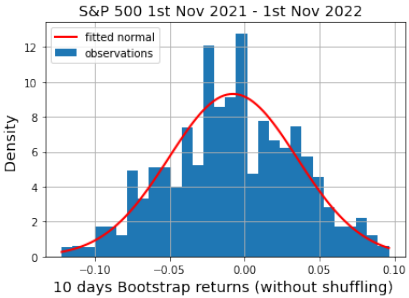
- For these reasons some alternative methods can be considered (VAE, GANs,..).
- But there is a "Zoo" of possible models, with their problems of implementation and convergence.
- We consider here the case of a single asset, the S&P 500.
- We illustrate the discrepancies between the standard methods.
- The generative model described here has not been devised specifically for *VaR* analysis and some details remain to fine tune but the logic behind it makes it interesting.
- This model has been developed by my colleague Gabriel Turinici for general purposes and has been adjusted here only slightly for predictive purposes.

Model Risk and Simulations





Model Risk and Simulations

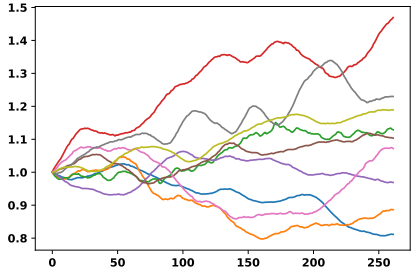


Methods	Values	Value x sqrt(10)
1d VaR_95% Histo x 2,33/1.65	0.037	0.118
1d VaR_99% Histo	0.038	0.121
1d VaR_99% Gaussian	0.035	0.109
XXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXX
10d VaR_99% Gaussian	XXXXXXXXXXXXXXXXXXX	0.114
XXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXX	XXXX Bootstrap XXXX
10d Var_99% with shuffling	XXXXXXXXXXXXXXXXXXX	0.156
10d Var_99% without shuffling	XXXXXXXXXXXXXXXXXXX	0.124

Deep neural network for VaR computation

• Goal: a deep generative neural network, of VAE (Variational Auto-Encoder) flavor, was trained to sample sequences of $252 + 10 = 262$ consecutive trading days.
• Training set: all blocks of 262 consecutive trading days available from 1/1/1928 to 1/11/2022, totaling 23561 (overlapping) sequences.
• The VAE is built within the class of 'energy' or Radon-Sobolev kernels, (cf. [8], see also GAN convergence issues at [7]).

Ten random evolution scenarios as constructed by the deep NN algorithm. We use 10'000 of them in practice.



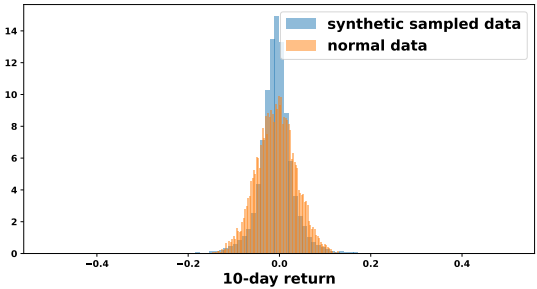
Deep neural network for VaR computation

• Once possible scenarios are sampled at random, a "proximity weight" is computed: the closer the first 252 days of the scenario are to the last 252 historical data, the more important the weight is; the last 10 days are NOT used.

Illustration of 3 close and 3 disparate scenarios from the 10'000 synthetic sampled scenarios; the weights are given in the legend. The weights appearing as zero are in fact smaller than 10^{-6}.

Deep neural network for VaR computation

The last 10 days of each scenario are renormalised, based on the mean and variance of the previous 252 days returns. These last 10 days trajectories are then used as 10 days samples to calculate the *VaR*.



• From these 10 days returns a *VaR* 99% is now computed. We find here a value 12.8% which is close to the 10 day bootstrap value without shuffling.

• More analysis and fine tuning of the model will be pursued but so far we find the results encouraging especially as we are able to simulate much more scenarios "model free" than with the bootstrap methods.

References I

[1] Doron Avramov and Guofu Zhou. "Bayesian portfolio analysis". In: *Annu. Rev. Financ. Econ.* 2.1 (2010), pp. 25–47.

[2] Pierre Brugière. *Quantitative portfolio management—with applications in Python*. Springer Texts in Business and Economics. Springer, Cham, [2020] ©2020, pp. xii+205. ISBN: 978-3-030-37740-3; 978-3-030-37739-7. DOI: 10.1007/978-3-030-37740-3. URL: https://doi.org/10.1007/978-3-030-37740-3.

[3] *DIS50 - Market risk*. en. 2019. URL: https://www.bis.org/basel_framework/index.htm (visited on 11/21/2022).

[4] *DP5/22 - Artificial Intelligence and Machine Learning*. en. URL: https://www.bankofengland.co.uk/prudential-regulation/publication/2022/october/artificial-intelligence (visited on 11/21/2022).

[5] *European Banking Authority : discussion paper on machine learning for IRB models*. en. Discussion Paper — EBA/DP/2021/04. (Visited on 11/2021).

[6] *MAR30 - Internal models approach*. en. 2019. URL: https://www.bis.org/basel_framework/index.htm (visited on 11/21/2022).

[7] Gabriel Turinici. "Convergence Dynamics of Generative Adversarial Networks: The Dual Metric Flows". en. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by Alberto Del Bimbo et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 619–634. ISBN: 978-3-030-68763-2. DOI: 10.1007/978-3-030-68763-2_47.

[8] Gabriel Turinici. "Radon–Sobolev Variational Auto-Encoders". In: *Neural Networks* 141 (Sept. 2021), pp. 294–305. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2021.04.018. URL: https://www.sciencedirect.com/science/article/pii/S0893608021001556.

[9] Lijia Wang, Xu Han, and Xin Tong. "Skilled Mutual Fund Selection: False Discovery Control under Dependence". In: *Journal of Business & Economic Statistics* (2022), pp. 1–15.

The End