Online Bayesian inference for multiple changepoints and risk assessment

Olivier Sorba, C Geissler

Advestis Working Paper
Online inference for multiple changepoints and risk assessment

Olivier Sorba[1] and C. Geissler[2]

[1]RandomPulse, 75116 Paris, France, olivier.sorba@randompulse.net
[2]Advestis, 69 Boulevard Haussmann, 75008 Paris, France, cgeissler@advestis.com

April 2021

[1,2]Work supported by Advestis

Abstract

The aim of the present study is to detect abrupt trend changes in the mean of a multidimensional sequential signal. Directly inspired by papers of Fernhead and Liu ([4] and [5]), this work describes the signal in a hierarchical manner : the change dates of a time segmentation process trigger the renewal of a piece-wise constant emission law. Bayesian posterior information on the change dates and emission parameters is obtained. These estimations can be revised online, i.e. as new data arrive. This paper proposes explicit formulations corresponding to various emission laws, as well as a generalization to the case where only partially observed data are available. Practical applications include the returns of partially observed multi-asset investment strategies, when only scant prior knowledge of the movers of the returns is at hand, limited to some statistical assumptions. This situation is different from the study of trend changes in the returns of individual assets, where fundamental exogenous information (news, earnings announcements, controversies, etc.) can be used.

Keywords: Regime changes, Bayesian inference, Assets returns

1 Introduction

This document outlines a method for the risk assessment of investment rules with specific attention to regime changes over time. It is based on the inference method proposed by P.Fearnhead et al. in [4] and [5]. The intention of the authors is to have a robust method for detecting changes in a vector emission law assumed to be piecewise stationary, such as the daily performance of a set of assets. The analytical framework proposed by Fernhead et al. ([4] and [5]) proves to be particularly well suited as it allows for online detection of changes in the vector emission law of asset returns under parsimonious assumptions on the parameters of the emission law. The prior knowledge is limited to the instantaneous probability of occurrence of a change, and to the parameters of a Gaussian distribution the emission law trend is drawn from. The purpose here is not to establish a causal or correlation link between observable variables from the 'real' world, and assets returns. Instead, the goal is to detect with a limited lag, the most likely distribution of changes in the underlying distribution.
We slightly generalize the framework of the original paper to the case where only some random components

of the performance vector can be observed. In the financial domain, an intended application in finance is to get an up-to-date estimate of an 'intermittent' investment strategy. Such a strategy is governed by decision rules that can be active or inactive at each time and for each invested stock. Estimating the law of returns for such strategies has therefore an additional layer of complexity, compared that of individual stocks. The introduction of random activation moves the model away from a pure econometric model. By this latter term, we mean a set of assumptions explicitly connecting companies' fundamental factors (such as size, profitability, industrial sector, etc.) and their returns. When such assumptions cannot anymore realistically be formulated, the Bayesian approach is very fruitful at providing online estimates for changes in the emission law.

1.1 Learning setup, definitions and notations

One observes a real possibly multi-dimensional random variable y_t at discrete dates $t \in [\![1, n]\!]$, for instance a series of asset performances. Let $y_{i:j}$ denote the observations from time i to time j included. We describe the process $y_{1:n}$ by the following multiple changepoint model, based on the assumption that the data before and after a changepoint are independent:

- There is an unknown random segmentation (a partition in contiguous segments) $\mathcal{S}$ of $[\![1, n]\!]$ so that $[\![1, n]\!] = \dot{\bigcup}_{\tau \in \mathcal{S}} \tau$.
- For any two integers $i < j$, we write $i \longmapsto j$ for the event $\{ \exists \tau \in \mathcal{S}, [\![i, j)\!] \subset \tau \}$, in other words $i, i+1, \cdots, j-1$ belong to the same segment.
- To each segment $\tau \in \mathcal{S}$ are associated a tuple β_τ of random parameters describing the emission law of $y_\tau := (y_t)_{t \in \tau}$, for instance $\beta_\tau = (\mu_\tau, v_\tau)$ where a μ_τ is a location parameter and v_τ is a size or dispersion parameter. We denote $\pi(\beta)\mathrm{d}\beta$, $\pi(\mu)\mathrm{d}\mu$, $\pi(v)\mathrm{d}v$ and so on the corresponding prior densities.
- Hierarchical structure:
  - conditional on the segmentation $\mathcal{S}$, the parameters sets and the observations of different segments are mutually independent,
  - the parameter sets follow the same distribution $\pi(\beta)\mathrm{d}\beta$,
  - the observations follow in each segment $\tau = [\![t, s)\!]$ the distribution $\mathbb{P}[\, y_{t:s-1} \mid t \longmapsto s, \beta_\tau \,]\, \mathrm{d}y_t \cdots \mathrm{d}y_{s-1}$.
- The starts of the segments of $\mathcal{S}$ (except $t = 1$) arise from a homogeneous point process on $\mathbb{Z}$ observed on the interval $[\![2, n]\!]$. This point process is defined by the probability mass function $g(t)$ of the distance between two consecutive changepoints.

Objective of this paper For any time $t \in [\![1, n]\!]$, denote β_t the unknown random parameter associated with the unique segment containing t. One particular point of interest is the last one β_n, and we would like to estimate some characteristics of its distribution conditional to the observations $y_{1:n}$. We start by restating the method of Fearnhead and Liu in [5], with some slight changes in notation and conventions. In particular, in our convention a changepoint is the leftmost point of a segment.

1.1.1 Prior probability distribution of the segmentations

Denote $G(\cdot)$ the distribution function of the distance between two successive changepoints. then

$$G(t) = \sum_{s=1}^{t} g(s)$$

Denote $g_0(\cdot)$ the mass distribution function of the distance $d = \tau_1 - 1$ the first changepoint τ_1 after $t = 1$ to the origin, which is the residual time in the language of renewal processes. We know classically that the mass distribution function $g_0(\cdot)$ of this distance has the following expression, quantifying the survival bias:

$$g_0(d) = \frac{\sum_{s=d}^{\infty} s^{-1} s g(s)}{\sum_{s=1}^{\infty} s g(s)} = \frac{1 - G(d-1)}{\sum_{s=1}^{\infty} 1 - G(s-1)}.$$

The probability of the segmentation $\mathcal{S} = \{ [1, \tau_1 - 1], [\tau_1, \tau 2 - 1], \cdots, [\tau_m, n] \}$ is expressed as:

$$g_0(\tau_1 - 1) \prod_{i=1}^{m-1} g(\tau_{i+1} - \tau_i) \left(1 - G(n - \tau_m)\right).$$

The authors of [5] suggest a negative binomial distribution. With parameters p and n, then this is the distribution of the number of independent Bernoulli trials before reaching n successes. Then classically:

$$g(t) = \binom{t-1}{n-1} p^n (1-p)^{t-n},$$
$$g_0(t) = \sum_{i=1}^{n} \binom{t-1}{i-1} p^i (1-p)^{t-i} / n.$$

Proof. Consider a Markov process of on $\mathbb{Z}_n$ with $x \to x$ with probability $1-p$ and $x \to x+1$ with probability p. Then $g()$ is the probability distribution of the return time to $x = 0$ (via $x = n - 1$) starting from 0, and $g(t) = p\,\mathbb{P}[x_{t-1} = n-1] = p\,\mathbb{P}[\mathcal{B}(t-1, p) = n-1]$. Additionally, $g_0()$ is the same starting in the process' stationary distribution, which is uniform by cyclicity. □

For $n = 1$ the survival law is geometric and the point process is Markov. Higher values of n can reduce the number of very short segments.

2 Filtering Recursions

Let us first define quantities that will appear frequently in subsequent calculations.

2.1 Segment likelihood

First define for $t \leq s$ the *ex ante* distribution on segment observations:

$$P(t, s) = \mathbb{P}[y_{t:s} \mid t \longmapsto s+1], \tag{2.1}$$
$$= \int \mathbb{P}[y_{t:s} \mid t \longmapsto s+1, \beta]\, \pi(\beta) \mathrm{d}\beta \tag{2.2}$$

assuming that $P(t, s)$ can be calculated analytically or numerically for all $[t, s] \subset [1, n]$. As the authors of [5] point, this requires either conjugate priors on β or numerical integration. We postpone to section 4 the description of concrete tractable examples.

2.2 Predecessor changepoint process

Define C_t as the location of the changepoint immediately preceding t, or 1 if there is none. C_{n+1} refers to the last breakpoint. In other words C_t is the start of the segment containing $t - 1$. The only manner for

C_{t+1} to differ from C_t is for t to be a changepoint. It follows from the model described in introduction that C_t is a Markov process with for $i > 1$:

$$\mathbb{P}\left[C_{t+1} = j \mid\ C_t = i\right] = \begin{cases} \frac{1-G(t-i)}{1-G(t-i-1)} & \text{if } j = i, \\ \frac{G(t-i)-G(t-i-1)}{1-G(t-i-1)} & \text{if } j = t, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The case of $i = 1$ is obtained by substituting $G_0(\cdot)$ to $G(\cdot)$ in the relation above.
As usual with hidden Markov models and Viterbi algorithms, the process C_t will follow a reverse Markov chain when conditioned to the observations [3, chap. 3 p. 51]. In this perspective define for $1 \leq j \leq t < n+1$ the conditional probabilities

$$p_t^{(j)} = \mathbb{P}\left[C_{t+1} = j \mid\ y_{1:t}\right]. \tag{2.4}$$

Assume the posterior distributions $p_t^{(\cdot)}$ are known. This allows to sample the position of the last changepoint from its exact posterior distribution. Finally the full segmentation can be sampled from its exact posterior distribution by iterating backwards until the origin of the observations is reached. In the same manner, maximum a posteriori (MAP) estimates are obtained by recursively taking the most probable previous changepoint. A single backward pass yields the marginal posterior changepoint probabilities thanks to the recursion relation:

$$\mathbb{P}\left[i \,\text{changepoint} \mid\ y_{1:n}\right] = p_n^{(i)} + \sum_{j=i+1}^{n} p_{j-1}^{(i)}\, \mathbb{P}\left[j \,\text{changepoint} \mid\ y_{1:n}\right]. \tag{2.5}$$

2.3 Forward recursion on the predecessor changepoint $p_t^{(\cdot)}$ distributions

When $y_1, y_2, \cdots, y_t$ is known and y_{t+1} becomes available, the additional information provided by y_{t+1} is contained in the following likelihood ratios, that Fearnhead and Liu propose to use as update weights for the posterior probability $p_t^{(\cdot)}$:

$$w_{t+1}^{(j)} := \mathbb{P}\left[y_{t+1} \mid\ C_{t+2} = j, y_{1:t}\right], \tag{2.6}$$

$$= \mathbb{P}\left[y_{t+1} \mid\ j \longmapsto t+1, y_{j:t}\right], \tag{2.7}$$

$$= \begin{cases} \frac{P(j,t+1)}{P(j,t)} & \text{if } j < t+1, \\ P(t+1,t+1) & \text{if } j = t+1. \end{cases} \tag{2.8}$$

In order to do so, the following recursion relation is available:

$$p_t^{(j)} \propto \begin{cases} w_t^{(j)} \frac{1-G(t-j)}{1-G(t-j-1)} p_{t-1}^{(j)} & \text{if } j < t, \\ w_t^{(t)} \sum_{i=1}^{t-1} \frac{G(t-i)-G(t-i-1)}{1-G(t-i-1)} p_{t-1}^{(i)} & \text{if } j = t. \end{cases} \tag{2.9}$$

We recall the demonstration for completeness ([5]) in Section 6.1.
These weights usually rely on summary statistics that can be incrementally updated, limiting the processing cost.

2.3.1 Approximate inference

In Equation 2.9 above, the set of indices $[\,1, t-1\,]$ may be seen as a particle swarm of candidate changepoints knowing the signal up to date t. The authors of [5] show how to speedup calculations by limiting this particle swarm to the most likely changepoints, based on an efficient re-sampling scheme. As we deal with time series of moderate length, there is no immediate need for optimization and we only restate the recursion method.

3 Risk evaluation

After recalling the method proposed by the authors of [5], we would like to apply the same principles to evaluate the posterior probability of an event Ω related to the parameters of the last segment, for instance $\Omega_\theta = \{\, \mu_n < \theta \,\}$ for some real θ assuming that we know how to express the corresponding probability within a connected segment:

$$P_\Omega(i,j) := \mathbb{P}\left[\, \Omega, y_{i:j} \mid\ i \longmapsto j+1 \,\right]. \tag{3.1}$$

For the frequent case $\Omega = \{\, \mu_j \leq \theta \,\}$, we denote

$$P_\theta(i,j) := \mathbb{P}\left[\, \mu_j \leq \theta, y_{i:j} \mid\ i \longmapsto j+1 \,\right]. \tag{3.2}$$

Proposition 1. *Consider an event Ω depending only on the last segment's parameter set (e.g. $\beta_n = (\,\mu_n, v_n\,)$, Then the following relation holds:*

$$\begin{aligned}\mathbb{P}\left[\, \Omega \mid\ y_{1:n} \,\right] &= \sum_{j=1}^{n} p_n^{(j)}\, \mathbb{P}\left[\, \Omega \mid\ y_{j:n}, C_{n+1} = j \,\right], \\ &= \sum_{j=1}^{n} p_n^{(j)}\, \frac{P_\Omega(j,n)}{P(j,n)}\end{aligned}$$

The proof is given in Section 6.2

4 Examples with known models

In this section we provide worked out examples of models suitable to calculate both the segment likelihood function $P(s, s+k-1) := \mathbb{P}\left[\, y_{s:s+k} \mid\ s \longmapsto s+k \,\right]$ and the distribution $\mathbb{P}\left[\, \mu \mid\ y_{s:s+k-1}, s \longmapsto s+k \,\right]$ of the segment parameter μ conditioned to the observation of $y_{s:s+k-1}$. Following [5] we favor models where the location parameters and the noise parameters are governed by a common scale parameter σ^2.
As this section deals with a fixed segment $[\, s, s+k-1 \,]$ of length k, so may assume that $s = 1$ by time invariance and also simply write y for $y_{s:s+k-1}$.

4.1 Gaussian multilinear regression with fixed variance parameter

Assume y_t is d-dimensional for each $1 \leq t \leq n$. To simplify notation we write y as a flat vector of dimension kd

$$y := \bigoplus_{t=s}^{s+k-1} y_t \tag{4.1}$$

On a segment of length k, consider the linear regression model:

$$y = H\mu + \epsilon, \tag{4.2}$$

$$\epsilon \sim \mathcal{N}_{kd}(0_{kd}, \sigma^2 \Sigma), \tag{4.3}$$

$$\mu \sim \mathcal{N}_q(0_q, \sigma^2 D) \tag{4.4}$$

where

- σ^2 is a variance parameter that we assume fixed for the present section,

- H is a $kd \times q$ matrix of q regression vectors with $q \leq d$,
- Σ is a $kd \times kd$ covariance matrix of rank kd,
- and $D = Diag(\delta_1^2, \cdots, \delta_q^2)$ is a fixed $q \times q$ positive diagonal matrix, representing the *a priori* size ratios between the explanatory variables terms and the observation noise.

Often one assumes that the noise ϵ has no time autocorrelation and is identically distributed over time, so that the noise covariance matrix writes as a block diagonal matrix:

$$\sigma^2\Sigma = \mathbb{I}_k \bigotimes \mathrm{Cov}\left(\epsilon_s\right).$$

Contrary to [5], as our observations are multidimensional, we do not assume the observation noise i.i.d across the signal's dimension at fixed date, hence the presence of the Σ covariance matrix.
In the same manner, if the regression vectors have no time evolution, they write as k repetitions of a bloc of size d:

$$H_i = \bigoplus_{t=s}^{s+k-1} h_i,$$

where for instance h_i can be a principal component of the d-dimensional process y_t.
In this setup, classically:

Proposition 2 (Multilinear setup with fixed noise parameter)**.**
the following relations hold:

$$P(s, s+k-1|\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \left(\frac{|M|}{|D|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\|y\|_P^2\right], \tag{4.5}$$

$$\mu \,|\, y, \sigma^2 \sim \mathcal{N}_q(\hat{\mu}, \sigma^2 M), \tag{4.6}$$

$$H\mu \,|\, y, \sigma^2 \sim \mathcal{N}_{kd}(\hat{y}, \sigma^2 HMH^T), \tag{4.7}$$

$$v^T H\mu \,|\, y, \sigma^2 \sim \mathcal{N}(v^T\hat{y}, \sigma^2 v^T HMH^T v). \tag{4.8}$$

where v is any kd-dimensional real vector and

$$\begin{aligned}
M &= \left[H^T\Sigma^{-1}H + D^{-1}\right]^{-1}, \\
P &= \Sigma^{-1} - \Sigma^{-1}HMH^T\Sigma^{-1}, \\
\|y\|_P^2 &= y^T P y, \\
\hat{\mu} &= MH^T\Sigma^{-1}y, \\
\hat{y} &= H\hat{\mu} = HMH^T\Sigma^{-1}y.
\end{aligned}$$

This set of results is close to [6, 3.2 p54], although with non i.i.d noise. The proof is given for completeness in Section 6.3

4.2 Gaussian multilinear model with Inverse-Gamma prior on variance

Still following [5], assume that in the multilinear model of Section 6.3 above the variance parameter σ^2 is not fixed but follows an inverse Gamma law of parameters $\nu/2$ and $\gamma/2$ (see [2]):

$$\pi(\sigma^2)\mathrm{d}\sigma^2 = \frac{1}{\Gamma(\frac{\nu}{2})}\left(\frac{\gamma}{2\sigma^2}\right)^{\frac{\nu}{2}} \exp\left(-\frac{\gamma}{2\sigma^2}\right)\sigma^{-2}\mathrm{d}\sigma^2. \tag{4.9}$$

Then:

Proposition 3 (Multilinear model with inverse Gamma prior on variance)**.**
The following relation holds:

$$P(s, s+k-1) = (\pi)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \left(\frac{|M|}{|D|} \right)^{\frac{1}{2}} \frac{\Gamma(\frac{dk+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\gamma)^{\frac{\nu}{2}}}{\left(\gamma + \|y\|_P^2\right)^{\frac{dk+\nu}{2}}}. \tag{4.10}$$

Conditioned to y, the $\sigma^2|y$ random parameter follows an inverse Gamma law of parameter $(\frac{\nu+kd}{2}, \frac{\gamma+\|y\|_P^2}{2})$. Notably,

$$\mathbb{E}\left[\sigma^2 \mid y\right] = \frac{\gamma + \|y\|_P^2}{\nu + kd - 2} \text{ for } \nu + kd > 2,$$
$$\mathbb{V}\left[\sigma^2 \mid y\right] = \frac{2}{\nu + kd - 4} \mathbb{E}\left[\sigma^2 \mid y\right]^2 \text{ for } \nu + kd > 4.$$

For any $v \in \mathbb{R}^q$ The random variable $v^T\mu|y$ is distributed like $v^T\hat{\mu} + \left(\frac{\gamma v^T M v}{\nu}\right)^{\frac{1}{2}} \mathcal{T}_\nu$ where $\mathcal{T}_\nu$ is a Student's T variable with fractional ν degrees of freedom. Notably

$$\mathbb{E}\left[v^T\mu \,\middle|\, y\right] = v^T\hat{\mu}, \tag{4.11}$$
$$= v^T M \Sigma^{-1} y, \tag{4.12}$$
$$\mathbb{V}\left[v^T\mu \,\middle|\, y\right] = \frac{\gamma}{\nu - 2} v^T M v \text{ if } \nu > 2. \tag{4.13}$$

Last the law of linear transform $A^T(\mu - \hat{\mu})$ of higher rank is a multivariate t-distribution (see [1]) of covariance $\frac{\gamma}{\nu-2} A^T M A$, assuming this last matrix is inversible.

The classical Equation 4.10 is given without proof in [5] in the case $\Sigma = \mathbb{I}_{kd}$. This results are also similar to [6, 3.2 p.54], still with $\Sigma = \mathbb{I}_{kd}$.

4.3 Partially observed multilinear model

In this section we keep the setup of Section 4.2 above, but assume the observations are only partial. More precisely, we assume there is a random projector P such that only Py is observed. For instance, the case:

$$\Pi = \bigoplus_{t=s}^{s+k-1} \sum_{i=1}^{d} R_{i,t} e_i e_i^T,$$

where $(e_1, \cdots, e_d)$ is a basis of $\mathbb{R}^d$ and R_i, t a random activation rule. There is no need to specify independence rules between Π and other variables since we are only interested in what happens in the image of Π, in other words in the components of y where the rules are activated. Then all estimates can be conditioned to Π.
In this modified setup, the model of Section 4.2 is maintained but considered latent, the observations being modeled by:

$$\begin{aligned}
\Pi y &= \Pi H \mu + \Pi \epsilon, \\
\Pi \epsilon &\sim \mathcal{N}_{\mathrm{Tr}\,\Pi}(0, \sigma^2 \Pi \Sigma \Pi^T), \\
\mu &\sim \mathcal{N}_q(0, \sigma^2 D), \\
\frac{1}{\sigma^2} &\sim \Gamma(\frac{\gamma}{2}, \frac{\nu}{2}).
\end{aligned}$$

This is exactly the setup of Section 4.2 when y, H and Σ are replaced by Py, PH and $P\Sigma P^T$. So the same considerations lead to:

Proposition 4. *conditioned to* Πy *and* Π*, the following relations hold:*

$$P(s, s+k-1|\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{\operatorname{Tr}\Pi}{2}} |\Sigma_\Pi|_\Pi^{-\frac{1}{2}} \left(\frac{|M_\Pi|}{|D|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\|y\|_{P_\Pi}^2\right], \tag{4.14}$$

$$P(s, s+k-1) = (\pi)^{-\frac{\operatorname{Tr}\Pi}{2}} |\Sigma_\Pi|_\Pi^{-\frac{1}{2}} \left(\frac{|M_\Pi|}{|D|}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{\operatorname{Tr}\Pi+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{(\gamma)^{\frac{\nu}{2}}}{\left(\gamma + \|y\|_{P_\Pi}^2\right)^{\frac{\operatorname{Tr}\Pi+\nu}{2}}}. \tag{4.15}$$

as well as the following relations in law:

$$\mu \,|\, \sigma^2 \sim \mathcal{N}_q(\hat{\mu}, \sigma^2 M_\Pi), \tag{4.16}$$

$$\sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu + \operatorname{Tr}\Pi}{2}, \frac{\gamma + \|y\|_{P_\Pi}^2}{2}\right), \tag{4.17}$$

$$v^T(\mu - \hat{\mu}) \sim \left(\frac{\gamma v^T M v}{\nu}\right)^{\frac{1}{2}} \text{Student's-t}(\nu). \tag{4.18}$$

where $v \in \mathbb{R}^d$ *and*

$$\begin{aligned}
&\Sigma_\Pi = \Pi\Sigma\Pi^T,\\
&M_\Pi = \left[H^T\Sigma_\Pi^+ H + D^{-1}\right]^{-1},\\
&P_\Pi = \Sigma_\Pi^+ - \Sigma_\Pi^+ H M_\Pi H^T \Sigma_\Pi^+,\\
&\|y\|_{P_\Pi}^2 = y^T P_\Pi y,\\
&\hat{\mu} = M_\Pi H^T \Sigma_\Pi^+ y,\\
&\hat{y} = H M_\Pi H^T \Sigma_\Pi^+ y.
\end{aligned}$$

To keep all the matrices of same dimension, we introduced the pseudo inverse Σ_Π^+ obtained with a few bloc matrices manipulations as

$$\Pi\left[(\mathbb{I}_{kd} - \Pi) + \Pi\Sigma\Pi\right]^{-1}\Pi$$

and the determinant $|\Sigma_\Pi|_\Pi$ of its restriction to $Im(\Pi)$ obtained as

$$\left|(\mathbb{I}_{kd} - \Pi) + \Pi\Sigma\Pi\right|.$$

Note that with a null P, which means no observation, we recover the prior distributions of the segment parameters μ and σ^2.

5 Examples

In this section we detail the calculation of the segment likelihood function is some common cases.

5.1 Regression by step functions : time invariant noise and covariates

A simple case occurs when one assumes the signal's conditional expectation is constant over each segment and one wants to explain the observed signal by step functions. then H may be expressed as

$$H = \bigoplus_{t=s}^{s+k-1} H_0, \tag{5.1}$$

where H_0 is a $d \times q$ matrix of q covariate vectors of dimension d. In addition, let us assume that the distribution of the noise ϵ is time invariant and presents no cross-correlation over different dates. Then its correlation matrix may be written as a bloc diagonal matrix:

$$\Sigma = \bigoplus_{t=s}^{s+k-1} \Sigma_0, \tag{5.2}$$

where Σ_0 is the noise correlation structure at any date.
Last, we assume the partial information situation is described by the activations $r_{i,t}$, so:

$$\begin{aligned} \Pi &= \bigoplus_{t=s}^{s+k-1} \Pi_t, \\ &= \sum_{t=s}^{s+k-1} \sum_{i=1}^{d} r_{i,t} e_{i,t} e_{i,t}^T, \end{aligned}$$

with the same bloc structure than the noise correlation.
In this case, some bloc matrix algebra leads to:

$$\Sigma_\Pi^+ = \bigoplus_{t=s}^{s+k-1} \Pi_t \left[\left(\mathbb{I}_d - \Pi_t \right) + \Pi_t \Sigma_0 \Pi_t \right]^{-1} \Pi_t, \tag{5.3}$$

and

$$|\Sigma_\Pi| = \prod_{t=s}^{s+k-1} \left| \left(\mathbb{I}_d - \Pi_t \right) + \Pi_t \Sigma_0 \Pi_t \right|. \tag{5.4}$$

where Σ_0 is the noise correlation structure at any date and Π_t the projector on the component space observable at date t. Finally, all the relevant quantities can be expressed as functions of running sums of matrices or vectors:

$$M_\Pi = \left[D^{-1} + \sum_{t=s}^{s+k-1} H_0^T \left(\Pi_t \Sigma_0 \Pi_t \right)^+ H_0 \right]^{-1}, \tag{5.5}$$

$$\hat{\mu} = M_\Pi \sum_{t=s}^{s+k-1} H_0^T \left(\Pi_t \Sigma_0 \Pi_t \right)^+ y_t, \tag{5.6}$$

$$\hat{y} = \bigoplus_{t=s}^{s+k-1} \hat{y}_0, \tag{5.7}$$

with

$$\hat{y}_0 = H_0 M_\Pi \sum_{t=s}^{s+k-1} H_0^T \left(\Pi_t \Sigma_0 \Pi_t \right)^+ y_t, \tag{5.8}$$

$$y^T P_\Pi y = \sum_{t=s}^{s+k-1} y_t^T \left(\Pi_t \Sigma_0 \Pi_t \right)^+ y_t - \hat{\mu}^T M^{-1} \hat{\mu}, \tag{5.9}$$

$$= \sum_{t=s}^{s+k-1} y_t^T \left(\Pi_t \Sigma_0 \Pi_t \right)^+ y_t - \sum_{t=s}^{s+k-1} \hat{y}_0^T \left(\Pi_t \Sigma_0 \Pi_t \right)^+ \hat{y}_0 - \hat{\mu}^T D^{-1} \hat{\mu}. \tag{5.10}$$

5.2 Regression by step functions : time invariant covariates and white noise

In addition to the preceding section, let us assume first hand that the noise ϵ is i.i.d so that $\Sigma_0 = \mathbb{I}_d$, and second that the covariates at any date are the natural basis of $\mathbb{R}^d$, so that $q = d$ and $H_0 = \mathbb{I}_d$.
Finally, some algebraic transformations lead to

$$\begin{aligned}
\Sigma_\Pi &= \sum_{t=s}^{s+k-1} \sum_{i=1}^{d} e_{i,t} r_{i,t} e_{i,t}^T, \\
\left| \Sigma_\Pi \text{ restricted to } \Pi\left(\mathbb{R}^{kd}\right) \right| &= 1 \\
M_\Pi &= \sum_{i=1}^{q} e_i \frac{\delta_i^2}{1+n_i\delta_i^2} e_i^T, \\
|M_\Pi| &= \prod_{i=1}^{q} \frac{\delta_i^2}{1+n_i\delta_i^2}, \\
P_\Pi &= \sum_{i,t} e_{i,t} r_{i,t} e_{i,t}^T - \sum_{i,t,t'} e_{i,t} r_{i,t} \frac{\delta_i^2}{1+n_i\delta_i^2} r_{i,t'} e_{i,t'}^T, \\
\|y\|_{P_\Pi}^2 &= \sum_{i,t} r_{i,t} y_{i,t}^2 - \sum_i \frac{\delta_i^2}{1+n_i\delta_i^2} \left(\sum_t r_{i,t} y_{i,t} \right)^2, \\
&= \sum_i n_i \left(\bar{y_i^2} - \frac{n_i\delta_i^2}{1+n_i\delta_i^2} \bar{y_i}^2 \right), \\
&= \sum_i n_i \left(\bar{y_i^2} - \bar{y_i}^2 + \frac{1}{1+n_i\delta_i^2} \bar{y_i}^2 \right), \\
\hat{\mu} &= \sum_{i=1}^{d} \left(1 - \frac{1}{1+n_i\delta_i^2} \right) \bar{y_i} e_i, \\
\hat{y}_t &= \sum_{i=1}^{d} \left(1 - \frac{1}{1+n_i\delta_i^2} \right) \bar{y_i} e_i.
\end{aligned}$$

where for $1 \leq i \leq q$

$$\begin{aligned}
n_i &= \sum_{t=s}^{s+k-1} r_{i,t}, \\
\bar{y_i} &= \frac{1}{n_i} \sum_{t=s}^{s+k-1} r_{i,t} y_{i,t}, \\
\bar{y_i^2} &= \frac{1}{n_i} \sum_{t=s}^{s+k-1} r_{i,t} y_{i,t}^2.
\end{aligned}$$

As expected, both the likelihood and the linear estimates appear as mixtures between their counterparts arising one hand from the prior Bayesian model and on the other hand from the purely linear regression model. This example also shows how naturally the formulation above deals with missing values or dates in the time series.

6 Proofs

6.1 Proof of Equation 2.9

Proof. The first step is to recall that the prior "previous changepoint" process C_t is a Markov chain with the transition probabilities of Equation 2.3.
Then by Bayes relation the following filtering recursions hold:

$$\mathbb{P}\left[C_{t+1} = j \mid y_{1:t}\right] \propto \mathbb{P}\left[y_t \mid C_{t+1} = j, y_{1:t-1}\right] \mathbb{P}\left[C_{t+1} = j \mid y_{1:t-1}\right]$$

and

$$\mathbb{P}\left[C_{t+1} = j \mid y_{1:t-1}\right] = \sum_{i=1}^{t} \mathbb{P}\left[C_{t+1} = j \mid C_t = i\right] \mathbb{P}\left[C_t = i \mid y_{1:t-1}\right]$$

so that

$$\begin{aligned}
\mathbb{P}\left[C_{t+1} = j \mid y_{1:t}\right] \propto & \mathbb{P}\left[y_t \mid C_{t+1} = j, y_{1:t-1}\right] \\
& \sum_{i=1}^{t-1} \mathbb{P}\left[y_t \mid C_{t+1} = j, y_{1:t-1}\right] \mathbb{P}\left[C_{t+1} = j \mid C_t = i\right] \mathbb{P}\left[C_t = i \mid y_{1:t-1}\right], \\
\propto & w_t^{(j)} \sum_{i=1}^{t-1} \mathbb{P}\left[C_{t+1} = j \mid C_t = i\right] \mathbb{P}\left[C_t = i \mid y_{1:t-1}\right].
\end{aligned}$$

In fine,

$$\mathbb{P}\left[C_{t+1} = j \mid y_{1:t}\right] \propto \begin{cases} w_t^{(j)} \frac{1-G(t-i)}{1-G(t-i-1)} \mathbb{P}\left[C_t = j \mid y_{1:t-1}\right] & \text{if } j < t, \\ w_t^{(j)} \sum_{i=1}^{t-1} \frac{G(t-i)-G(t-i-1)}{1-G(t-i-1)} \mathbb{P}\left[C_t = i \mid y_{1:t-1}\right] & \text{if } j = t, \end{cases}$$

or equivalently, Equation 2.9. □

6.2 Proof of proposition 1

Proof. Again, we proceed by conditioning on the position of the last changepoint. Then as the event Ω only depends on the last segment parameters,

$$\begin{aligned}
\mathbb{P}\left[\Omega \mid y_{1:n}\right] &= \sum_{j=1}^{n} \mathbb{P}\left[C_{n+1} = j \mid y_{1:n}\right] \mathbb{P}\left[\Omega \mid y_{1:n}, C_{n+1} = j\right], \\
&= \sum_{j=1}^{n} p_n^{(j)} \mathbb{P}\left[\Omega \mid y_{j:n}, C_{n+1} = j\right]. \\
&= \sum_{j=1}^{n} p_n^{(j)} \frac{\mathbb{P}\left[\Omega, y_{j:n} \mid C_{n+1} = j\right]}{\mathbb{P}\left[y_{j:n} \mid C_{n+1} = j\right]}, \\
&= \sum_{j=1}^{n} p_n^{(j)} \frac{P_\Omega(j,n)}{P(j,n)}.
\end{aligned}$$

□

6.3 Proof of Proposition 2

Proof. Recall that the segment parameter set β is reduced here to the location parameter μ, and that by definition,

$$
\begin{aligned}
&M = \left[H^T \Sigma^{-1} H + D^{-1} \right]^{-1}, \\
&P = \Sigma^{-1} \left(\Sigma - HMH^T \right) \Sigma^{-1}, \\
&\|y\|_P^2 = y^T P y, \\
&\hat{\mu} = MH^T \Sigma^{-1} y, \\
&\hat{y} = H\hat{\mu} = HMH^T \Sigma^{-1} y.
\end{aligned}
$$

The model definition and a few algebraic transformations based on the definitions above lead to:

$$
\begin{aligned}
\mathbb{P}\left[y \mid \ \mu \right] \pi(\mu) &= \left(2\pi\sigma^2 \right)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(y - H\mu \right)^T \Sigma^{-1} \left(y - H\mu \right) \right] \\
&\quad \left(2\pi\sigma^2 \right)^{-\frac{q}{2}} |D|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \mu^T D^{-1} \mu \right], \qquad (6.1) \\
&= \left(2\pi\sigma^2 \right)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(y^T \Sigma^{-1} y - 2 y^T \Sigma^{-1} H\mu \right) \right] \\
&\quad \left(2\pi\sigma^2 \right)^{-\frac{q}{2}} |D|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \mu^T M^{-1} \mu \right], \qquad (6.2) \\
&= \left(2\pi\sigma^2 \right)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(y^T \Sigma^{-1} y - y^T \Sigma^{-1} HMH^T \Sigma^{-1} y \right) \right] \\
&\quad \left(2\pi\sigma^2 \right)^{-\frac{q}{2}} |D|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(\mu - MH^T \Sigma^{-1} y \right)^T M^{-1} \left(\mu - MH^T \Sigma^{-1} y \right) \right], \qquad (6.3) \\
&= \left(2\pi\sigma^2 \right)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \|y\|_P^2 \right] \\
&\quad \left(2\pi\sigma^2 \right)^{-\frac{q}{2}} |D|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(\mu - \hat{\mu} \right)^T M^{-1} \left(\mu - \hat{\mu} \right) \right], \qquad (6.4)
\end{aligned}
$$

The first and second results in the proposition follow by the relation

$$
P(s, s+k) = \int \mathbb{P}\left[y \mid \ \mu \right] \pi(\mu) \mathrm{d}\mu.
$$

and by the Bayesian relation

$$
\mathbb{P}\left[\mu \mid \ y \right] \propto \mathbb{P}\left[y \mid \ \mu \right] \pi(\mu).
$$

□

The third relation follows from the relation $\hat{y} = H\hat{\mu}$. The last relation follows by a simple expectation and variance calculation

6.4 Proof of Proposition 3

Proof. Starting from the results of the fixed σ^2 model leads to:

$$\begin{aligned}
P(s, s+k-1) &= \iint \mathbb{P}\left[\, y \,|\, \mu, \sigma^2 \,\right] \pi(\mu)\pi(\sigma^2) \mathrm{d}\mu \mathrm{d}\sigma^2, \\
&= \int \frac{1}{\Gamma(\frac{\nu}{2})} \left(\frac{\gamma}{2\sigma^2}\right)^{\frac{\nu}{2}} \exp\left(-\frac{\gamma}{2\sigma^2}\right) \sigma^{-2} P(s, s+k-1|\sigma^2) \mathrm{d}\sigma^2, \\
&= \int \frac{1}{\Gamma(\frac{\nu}{2})} \left(\frac{\gamma}{2\sigma^2}\right)^{\frac{\nu}{2}} \exp\left(-\frac{\gamma}{2\sigma^2}\right) \left(2\pi\sigma^2\right)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \left(\frac{|M|}{|D|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \|y\|_P^2\right] \sigma^{-2} \mathrm{d}\sigma^2, \\
&= (\pi)^{-\frac{dk}{2}} |\Sigma|^{-\frac{1}{2}} \left(\frac{|M|}{|D|}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{dk+\nu)}{2}}{\Gamma(\frac{\nu}{2})} \frac{(\gamma)^{\frac{\nu}{2}}}{\left(\gamma + \|y\|_P^2\right)^{\frac{dk+\nu}{2}}}.
\end{aligned}$$

Next we know by Bayes' relation combined with the equations above that

$$\begin{aligned}
\mathbb{P}\left[\sigma^2 \,|\, y\right] &\propto \mathbb{P}\left[y \,|\, \sigma^2\right] \pi(\sigma^2), \\
&\propto \left(\sigma^2\right)^{-\frac{\nu+kd}{2}} \exp\left[-\frac{\gamma + \|y\|_P^2}{2\sigma^2}\right] \sigma^{-2} \mathrm{d}\sigma^2,
\end{aligned}$$

so that $\sigma^2|y$ follows an inverse Gamma law of parameter $(\frac{\nu+kd}{2}, \frac{\gamma+\|y\|_P^2}{2})$. Notably, for $\nu + kd > 2$,

$$\begin{aligned}
\mathbb{E}\left[\sigma^2 \,|\, y\right] &= \frac{\gamma + \|y\|_P^2}{2} \frac{\Gamma(\frac{\nu+kd}{2} - 1)}{\Gamma(\frac{\nu+kd}{2})}, \\
&= \frac{\gamma + \|y\|_P^2}{\nu + kd - 2},
\end{aligned}$$

and for $\nu + kd > 4$,

$$\mathbb{V}\left[\sigma^2 \,|\, y\right] = \frac{2}{\nu + kd - 4} \mathbb{E}\left[\sigma^2 \,|\, y\right]^2.$$

Last, we know from Proposition 2 that conditioned to y and σ^2, the random scalar $v^T H\mu$ is distributed like $\mathcal{N}(v^T H\hat{\mu}, \sigma^2 v^T HMH^T v)$ so

$$\begin{aligned}
\mathbb{P}\left[v^T H\mu = v^T H\hat{\mu} + z \,\middle|\, y\right] \mathrm{d}z &= \int \mathbb{P}\left[v^T \mu = v^T \hat{\mu} + z \,\middle|\, y, \sigma^2\right] \pi(\sigma^2) \mathrm{d}\sigma^2 \mathrm{d}z, \\
&\propto \int \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{z^2}{2\sigma^2 v^T HMH^T v}\right] \pi(\sigma^2) \mathrm{d}\sigma^2 \mathrm{d}z, \\
&\propto \int \left(\sigma^2\right)^{-\frac{\nu+1}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\gamma + \frac{z^2}{v v^T HMH^T v}\right)\right] \sigma^{-2} \mathrm{d}\sigma^2 \mathrm{d}z, \\
&\propto \left(1 + \frac{z^2}{\gamma v^T HMH^T v}\right)^{-\frac{\nu+1}{2}} \mathrm{d}z, \\
&\propto \left(1 + \frac{1}{\nu} \frac{z^2}{\nu^{-1}\gamma v^T HMH^T v}\right)^{-\frac{\nu+1}{2}} \mathrm{d}z,
\end{aligned}$$

so $z = v^T H(\mu - \hat{\mu})|y$ is distributed like $\left(\frac{\gamma v^T HMH^T v}{\nu}\right)^{\frac{1}{2}} \mathcal{T}_\nu$ where $\mathcal{T}_\nu$ is a Student's T variable with fractional ν degrees of freedom. Notably

$$\mathbb{E}[z \mid y] = 0,$$
$$\mathbb{V}[z \mid y] = \frac{\gamma}{\nu - 2} v^T HMH^T v \text{ if } \nu > 2.$$

By the same arguments, a linear transform $A^T H\mu$ of higher rank, will produce a multivariate t-distribution (see [1]) of covariance $\frac{\gamma}{\nu-2} A^T HMH^T A$, assuming this last matrix is invertible. □

7 Forthcoming applications

The algorithm described in this paper from the original presentation by Fernhead and Liu [5], uses particle filter based sampling, in order to limit the number of selected candidates in the determination of breakup instants. This maintains a linear complexity in the number of dates, instead of a quadratic complexity that would be a major obstacle to use for time series with several thousand points. However, this random sampling makes the results of the algorithm themselves random, notably the *maximum a posteriori* estimates. The uncertainty observed in the results increases as the maximum number of candidates is decreased. There is therefore a trade-off between computation time and accuracy.
A first study will consist in the evaluation of this trade-off, in other words the search for the minimal allowed number of candidates given a required minimum level of accuracy. The acceptable uncertainty on the *maximum a posteriori* localization of the regime changes can be considered as a business constraint, that will in turn influence the computational performance.
A second direction of work concerns the effectiveness of the detection method, compared to commonly used trend measurement methods. A usual context of application of this paper is the online estimation of the average performance of a system, financial or energetic for example. An operator monitors an emission law based on the arrival of new data, and updates the evaluation of the last regime change. Depending on a possible change in the emission law, for example a change in sign of the trend, the operator can modify the scaling factor allocated to this system. It is easy to a posteriori simulate the cumulative effect in time of the actions caused by the detection of regime changes. Different detection methods can thus be compared on an objective basis. This subject will be the subject of a future publication in the field of financial investment strategies.

Contents

References

[1] Multivariate t-distribution. `https://en.wikipedia.org/wiki/Multivariate_t-distribution`, 2020.

[2] Inverse gamma distribution. `https://en.wikipedia.org/wiki/Inverse-gamma_distribution`, 2020.

[3] Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005.

[4] P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006.

[5] P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

[6] Jean-Michel Marin and Christian P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2007.