Segmentation of time-series with various types of dependence

S. Chakar, E. Lebarbier, C. Lévy-Leduc and S. Robin

AgroParisTech / INRA MIA 518, Paris, France

Statistics month in Marseille, Processes 15-19 February, 2016

Introduction

Segmentation/breakpoint detection problems arise in many fields.

General aim: Given a series of observations, find abrupt change-points in their distribution.

Introduction

Segmentation/breakpoint detection problems arise in many fields.

General aim: Given a series of observations, find abrupt change-points in their distribution.

Examples:	Genomics	Meteorology	Geodesy
Time t	Position along the genome	Time	Time
Signal Y_t	Micro-array	Temperature	GPS location
Breakpoints $\{t_k\}$	Endpoints of altered regions	Change of instrument	Earth's crust shifts

Issues and outline

1. Segmentation: statistical and algorithmic issues
    - Statistics: choice of the number of segments.
    - Algorithmic: optimal repartition of the breakpoints (efficient exploration of the segmentation space).

Issues and outline

1. Segmentation: statistical and algorithmic issues
    - Statistics: choice of the number of segments.
    - Algorithmic: optimal repartition of the breakpoints (efficient exploration of the segmentation space).
2. Segmentation (breakpoint detection in the mean) of an independent Gaussian process (model, inference procedure & algorithm)

Issues and outline

1. Segmentation: statistical and algorithmic issues
   - Statistics: choice of the number of segments.
   - Algorithmic: optimal repartition of the breakpoints (efficient exploration of the segmentation space).
2. Segmentation (breakpoint detection in the mean) of an independent Gaussian process (model, inference procedure & algorithm)
3. Take into account for two types of dependence:

   (i) When dealing with time-series, it is likely that time-dependence exists.

   (ii) When dealing with multiple series, it is likely that some dependencies between series exist.

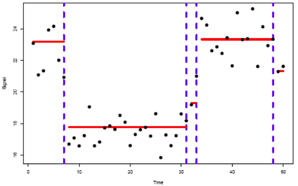
For each: model (how we model the dependence) / inference / simulation

Breakpoint detection in the mean of an independent Gaussian process

Model. The signal $Y = (Y_1, \ldots, Y_n)$ is such that:

$$Y_t = \mu_k + \eta_t \quad \text{if } t \text{ is in } I_k = [t_{k-1}+1, t_k],$$

where $\{\eta_t\}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, the convention $t_0 = 0$ and $t_K = n$, and $k = 1, \ldots, K$.



Parameters. The breakpoints $t = (t_1, ..., t_{K-1})$, the means $\mu = (\mu_1, \ldots, \mu_K)$, the variance σ^2 and the number of segments K.

Inference.

- K being fixed, estimation of t, μ and σ^2 by maximum likelihood.
- Choice of K. Penalized Log-Likelihood.

Estimation of t, μ and σ^2

The optimization problem consists in

$$\max_{t \in \mathcal{M}_{n,K}} \max_{\sigma>0} \max_{\mu \in \mathbb{R}^K} -n \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \mu_k)^2,$$

where $\mathcal{M}_{n,K} = \{(t_1, \ldots, t_{K-1}) \in \mathbb{N}^{K-1}, 0 = t_0 < t_1 < \ldots, t_{K-1} < t_K = n\}$.
Maximized in σ, it becomes

$$\min_{t \in \mathcal{M}_{n,K}} \min_{\mu \in \mathbb{R}^K} \text{crit}(t, \mu) = \min_{t \in \mathcal{M}_{n,K}} \min_{\mu \in \mathbb{R}^K} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \mu_k)^2.$$

Estimation of μ and t:

- The estimation of μ is straightforward $\widehat{\mu}_k = \frac{1}{n_k} \sum_{t \in I_k} Y_t$
- How to find the breakpoints t? We have to minimize

$$\text{crit}(t, \hat{\mu}) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (Y_t - \widehat{\mu}_k)^2.$$

Algorithmics

- There are $\binom{n-1}{K-1}$ possible ways to divide a series with length n into K segments ($n = 1000,\ K = 10 \rightarrow 10^{21}$ possibilities).

$\rightarrow$ A naive search is hopeless.

Algorithmics

- There are $\binom{n-1}{K-1}$ possible ways to divide a series with length n into K segments ($n = 1000,\ K = 10 \to 10^{21}$ possibilities).

  $\rightarrow$ A naive search is hopeless.

- Dynamic programming (DP) allows to recover the optimal segmentation with $\mathcal{O}(Kn^2)$ complexity, provided that the contrast to be optimized is additive.

Algorithmics

- There are $\binom{n-1}{K-1}$ possible ways to divide a series with length n into K segments ($n = 1000$, $K = 10 \to 10^{21}$ possibilities).

  $\rightarrow$ A naive search is hopeless.

- Dynamic programming (DP) allows to recover the optimal segmentation with $\mathcal{O}(Kn^2)$ complexity, provided that the contrast to be optimized is additive.

- In the independent case, we have to minimize

$$\sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (Y_t - \widehat{\mu}_k)^2$$

Note that on very large signals: 'Pruned' DP with a complexity $\mathcal{O}(Kn \log n)$ [Rigaill, 2010]

Choice of K

Classical penalized criterion: AIC and BIC.

Numerous criteria have been proposed

- Penalized likelihood [Lavielle, 2005; Lebarbier, 2005] ($\hat{K} = \underset{K}{\mathrm{Argmax}}\ \hat{\mathcal{L}}_K - \ pen(K)$ with $pen(K) = \beta K$ and $pen(K) = c_2 K + c_1 \log(\mathrm{C}_n^K)$ respectively).
- ICL [Rigaill et al., 2011];
- mBIC [Zhang and Siegmund, 2007]

  ...

$$mBIC_K(Y) \quad = \quad -\left(\frac{n-K}{2}+1\right)\log SS_{\text{wg}} + \log\Gamma\left(\frac{n-K}{2}+1\right) - \frac{1}{2}\sum_{k=1}^{K}\log\hat{n}_k - (K-1)\log n$$

where $\log SS_{\text{wg}} = \min_t \min_\mu \sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}(y_t-\mu_k)^2$ and $\hat{n}_k = \hat{t}_k - \hat{t}_{k-1}$.

$mBIC$ achieves the best performances (see [Picard et al., 2011b]).

Summary and proposed works

Breakpoint detection in the mean of an independent Gaussian process.

- Algorithmics: DP
- Choice of the number of segments : mBIC

Take into account for two types of dependence:

- When dealing with one time-series, it is likely that time-dependence exists: we model it with an AR(1) (PhD thesis of Souhil Chakar).
- When dealing with multiple series, it is likely that some dependencies between series exist: we propose a joint segmentation model and assume that the signals at each position are correlated from one series to another.

Summary and proposed works

Breakpoint detection in the mean of an independent Gaussian process.

- Algorithmics: DP
- Choice of the number of segments : mBIC

Take into account for two types of dependence:

- When dealing with one time-series, it is likely that time-dependence exists: we model it with an AR(1) (PhD thesis of Souhil Chakar).
- When dealing with multiple series, it is likely that some dependencies between series exist: we propose a joint segmentation model and assume that the signals at each position are correlated from one series to another.

Main problem: inference of t since the additivity condition is not satisfied and DP can not be used.

Summary and proposed works

Breakpoint detection in the mean of an independent Gaussian process.

- Algorithmics: DP
- Choice of the number of segments : mBIC

Take into account for two types of dependence:

- When dealing with one time-series, it is likely that time-dependence exists: we model it with an AR(1) (PhD thesis of Souhil Chakar).
- When dealing with multiple series, it is likely that some dependencies between series exist: we propose a joint segmentation model and assume that the signals at each position are correlated from one series to another.

Main problem: inference of t since the additivity condition is not satisfied and DP can not be used.

Our strategy consists in removing the dependency so that this algorithm can be applied.

Breakpoint detection in the mean of a Gaussien AR(1) process

Model. The signal $Y = (Y_1, \ldots, Y_n)$ is such that:

$$Y_t = \mu_k + \eta_t \quad \text{if } t \text{ is in } I_k = [t_{k-1}+1, t_k],$$

where $\{\eta_t\}_{t \in \mathbb{Z}}$ is a zero-mean stationary AR(1) Gaussian process defined as

$$\eta_t = \rho \eta_{t-1} + \epsilon_t$$

where $|\rho| < 1$ and $\{\epsilon_t\}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Parameters.
the breakpoints $t = (t_1, ..., t_{K-1})$, $t_k = [n\tau_k]$
the variance σ^2
the means $\mu = (\mu_1, \ldots, \mu_K)$, $\delta_k = \mu_k(1-\rho)$,
the auto-correlation ρ
the number of segments K

Inference, algorithmics and strategy

Inference. We have to maximize the $\log$-likelihood (conditionally to y_0), yet optimized with respect to σ, e.g. minimize

$$
\begin{aligned}
\text{crit}(\rho, t, \delta) \quad = \quad & \sum_{k=1}^{K} \sum_{t=t_{k-1}+2}^{t_k} (y_t - \rho\ y_{t-1} - \delta_k)^2 \\
& + (y_1 - \rho\ y_0 - \delta_0)^2 \\
& + \sum_{k=2}^{K} \left(\left(y_{t_{k-1}+1} - \frac{\delta_k}{1-\rho} \right) - \rho \left(y_{t_{k-1}} - \frac{\delta_{k-1}}{1-\rho} \right) \right)^2
\end{aligned}
$$

Inference, algorithmics and strategy

Inference.

$$\begin{aligned}
\text{crit}(\rho, t, \delta) &= \sum_{k=1}^{K} \sum_{t=t_{k-1}+2}^{t_k} (y_t - \rho\, y_{t-1} - \delta_k)^2 \\
&\quad + (y_1 - \rho\, y_0 - \delta_0)^2 \\
&\quad + \sum_{k=2}^{K} \left(\left(y_{t_{k-1}+1} - \frac{\delta_k}{1-\rho} \right) - \rho \left(y_{t_{k-1}} - \frac{\delta_{k-1}}{1-\rho} \right) \right)^2
\end{aligned}$$

$\Rightarrow$ Due to the term that involves both δ_{k-1} and δ_k, this criterion cannot be efficiently minimized.

Inference, algorithmics and strategy

Inference.

$$\text{crit}(\rho, t, \delta) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \rho \, y_{t-1} - \delta_k)^2$$

Inference, algorithmics and strategy

Inference.

$$\text{crit}(\rho, t, \delta) \quad = \quad \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \left(y_t - \rho \; y_{t_1} - \delta_k\right)^2$$

$\Rightarrow$ Due to the global parameter ρ, this criterion cannot be efficiently minimized.

Inference, algorithmics and strategy

Inference.

$$\text{crit}(\rho, t, \delta) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \rho \; y_{t_1} - \delta_k)^2$$

$\Rightarrow$ Due to the global parameter ρ, this criterion cannot be efficiently minimized.

Proposed strategy.

- Propose an estimator of the autocorrelation parameter ρ, denoted $\widetilde{\rho}_n$,
- Minimize crit$(\widetilde{\rho}_n, t, \delta)$.

$$\Leftrightarrow$$

Apply the independent strategy on the decorrelated series $y_t - \widetilde{\rho}_n y_{t-1}$.

A robust estimator of ρ

- A robust (to the presence of outliers) scale estimator (Rousseeuw and Croux (1993)):

$$\begin{aligned} S_n &\propto \text{med}_i \text{med}_j \left| x_i - x_j \right| \ , \\ Q_n &\propto \left(|x_i - x_j| \right)_{(\lfloor n/4 \rfloor)} \ . \end{aligned}$$

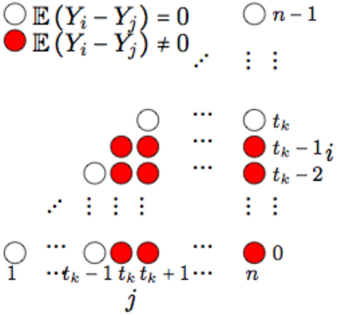
- A robust estimator of the autocorrelation function of a stationnary time series based on Q_n (Ma and Genton (2000)):

$$\widetilde{\rho}_{MG} = \frac{Q_n^2(y^+) - Q_n^2(y^-)}{Q_n^2(y^+) + Q_n^2(y^-)},$$

where $y^+ = (y_{i+1} + y_i)_{1 \le i \le n-1}$ and $y^- = (y_{i+1} - y_i)_{1 \le i \le n-1}$.

A robust estimator of ρ

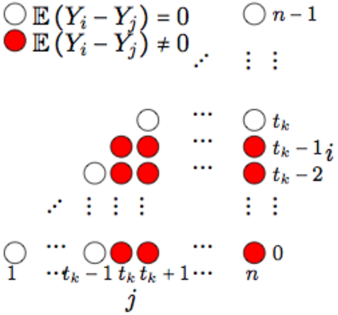
Breakpoint detection: a more robust estimator of the autocorrelation parameter ρ.



$\text{med}_i \text{med}_j |Y_i - Y_j|$

A robust estimator of ρ

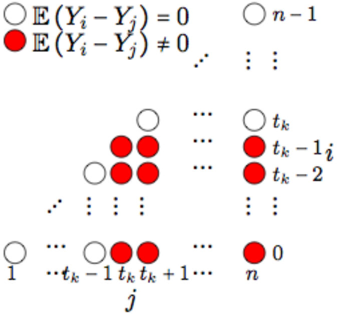
Breakpoint detection: a more robust estimator of the autocorrelation parameter ρ.



$\mathrm{med}_i |Y_i - Y_{i+1}|$

A robust estimator of ρ

Breakpoint detection: a more robust estimator of the autocorrelation parameter ρ.



$$\widetilde{\rho}_n = \left(\frac{\text{med}_i \, |Y_{i+2} - Y_i|}{\text{med}_i \, |Y_{i+1} - Y_i|} \right)^2 - 1$$

Convergence of the estimators

Proposition (Convergence of $\widetilde{\rho}_n$)

$\widetilde{\rho}_n$ *satisfies the following Central Limit Theorem*

$$\sqrt{n}(\widetilde{\rho}_n - \rho) \xrightarrow{d} \mathcal{N}(0, \tilde{\sigma}^2)\ ,\ \textit{as } n \to \infty$$

Proposition (Estimators of t and δ and convergence)

$$(\hat{\delta}_n, \hat{t}_n) \quad = \quad \arg \min_{(\delta,t)\in\mathbb{R}^K \times \mathcal{A}_{n,K}} crit(\widetilde{\rho}_n, t, \delta)\ , \hat{\tau}_n = \hat{t}_n/n,$$

where $\mathcal{A}_{n,K} = \{t; 0 = t_0 < t_1 < \ldots, < t_{K-1} < t_K = n, \forall k, t_k - t_{k-1} \geq \Delta_n\}$ *with a real sequence* (Δ_n) *such that* $n^{-\alpha}\Delta_n \longrightarrow \infty$ *when* $n \to \infty$ *and* $\alpha > 0$. *Then*

$$(\| \hat{\tau}_n - \tau \|) = O_P(n^{-1}) \ \textit{ and } \ \left(\| \hat{\delta}_n - \delta \|\right) = O_P(n^{-1/2})$$

Notice that we have the same asymptotic properties than in the independent case ([Lavielle & Moulines, 2000] ; [Bai & Perron, 1998])

Model selection

For the independent case, we select K as follows:

$$\hat{K} = \arg\max_K mBIC_K(Y)$$

where $mBIC_K$ is the modified BIC criterion [Zhang & Siegmund, 2007].

Model selection

For the independent case, we select K as follows:

$$\hat{K} = \arg\max_K mBIC_K(Y)$$

where $mBIC_K$ is the modified BIC criterion [Zhang & Siegmund, 2007].

For the Gaussian AR(1) process,

Proposition

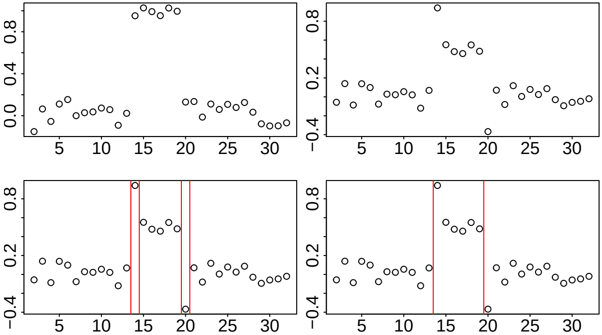
Let denote $X = (Y_t - \rho Y_{t-1})_t$ *and* $\tilde{X} = (Y_t - \widetilde{\rho}_n Y_{t-1})_t$, *then*

$$mBIC_m(\tilde{X}) = mBIC_m(X) + O_P(1)$$

We propose

$$\hat{K} = \arg\max_K mBIC_K(\tilde{X})$$

Post-processing (PP)



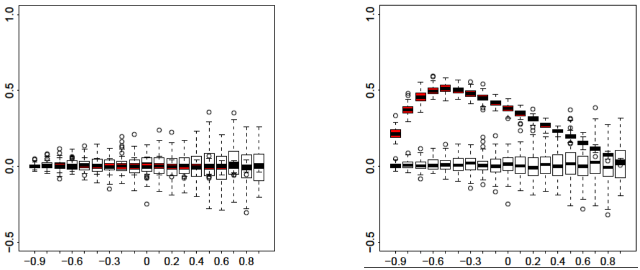
The decorrelation procedure introduces spurious breakpoints in the series, at distance 1 of the true breakpoints. The PP consists in removing segments of size 1:

$$PP = \left\{\hat{t}_{n,k} \in \hat{t}_n\right\} \setminus \left\{\hat{t}_{n,k} \;\text{ such that }\; \hat{t}_{n,i} = \hat{t}_{n,i-1} + 1 \;\text{ and }\; \hat{t}_{n,i+1} \neq \hat{t}_{n,i} + 1\right\}$$

Simulation study

Performance of $\widetilde{\rho}_n$. Comparison of $\widetilde{\rho}_n - \rho$ (in white) with $\widetilde{\rho}_{MG} - \rho$ (in red) with $\sigma = 0.6$ and $n = 1600$.

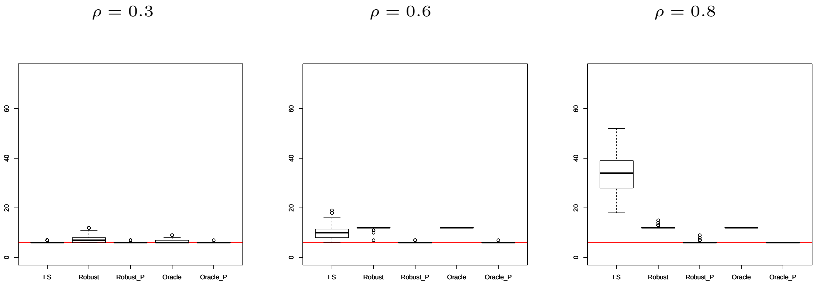
without breakpoints with breakpoints



Results. Without breakpoints, $\widetilde{\rho}_{MG}$ performs better. This is the contrary in the presence of breakpoints.

Simulation study

Selection of K. Different methods: LS ($\widetilde{\rho}_n = 0$), Robust ($\widetilde{\rho}_n = \widetilde{\rho}_n$) without PP and with PP, Oracle ($\widetilde{\rho}_n = \rho$) without PP and with PP. $\sigma = 0.1$ and $n = 1600$.

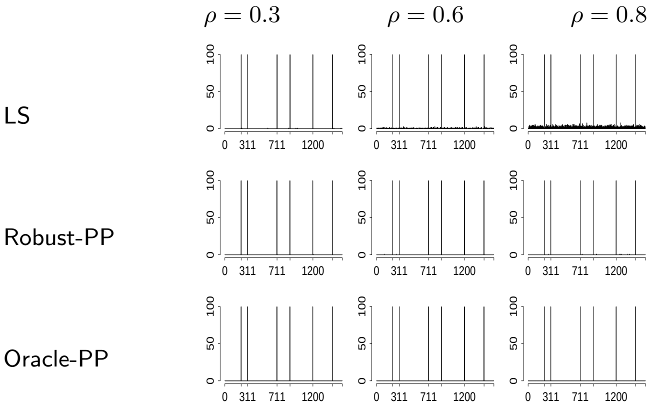


Results. Same result for small σ and ρ. Otherwise, LS tends to strongly overestimate K; Robust and Oracle tend to select twice the true number of segments; PP corrects it.

Simulation study

Breakpoint positionning. Frequencies of all possible breakpoint estimator with $n = 1600$.

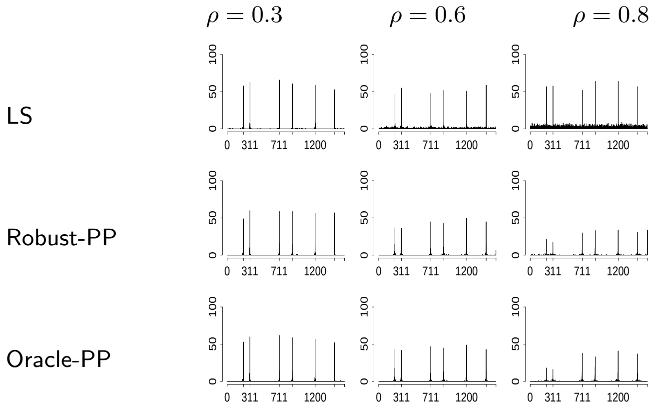
$\sigma = 0.1$



Simulation study

Breakpoint positionning. Frequencies of all possible breakpoint estimator with $n = 1600$.

$\sigma = 0.5$



Towards multiple sample analysis

Multiple series. One often deals with multiple series, observed

- at different locations or
- for different patients.

Series can be analyzed either independently or jointly.

Joint analysis allows to account for dependencies between series.

Model (joint segmentation) Denote Y_{mt} the signal at position t ($t = 1, \dots, n$) for series m ($m = 1 \dots, M$),

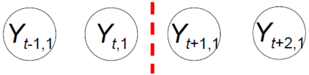
$$\forall t \in I_k^m, \qquad Y_{mt} = \mu_{mk} + F_{mt},$$

where $\{\mathbf{F}_t\}_t$ i.i.d. $\mathcal{N}_M(\mathbf{0}, \mathbf{\Sigma})$ (joint segmentation refers to the case where the breakpoints are specific to each series).

Graphical model and computational issue

Graphical model.

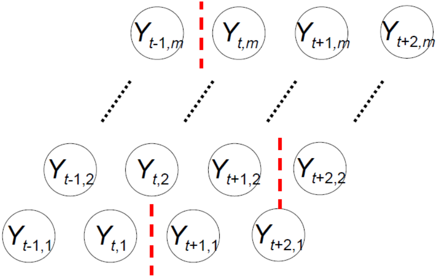
One series: OK with classical DP.



Graphical model and computational issue

Graphical model.

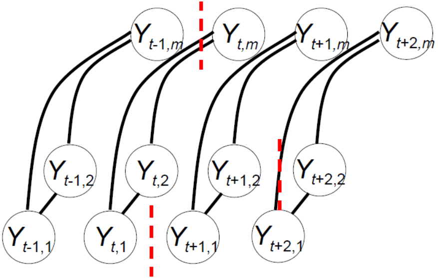
Independent series: OK with 2-stage DP ($\mathbf{\Sigma} = \sigma^2 \mathbf{I}$).



Graphical model and computational issue

Graphical model.

Joint segmentation: Non-additivity of the likelihood
$\rightarrow$ DP cannot apply as such.



$$-2 \log \mathcal{L}(\mathbf{Y}; \phi) \quad = \quad N \log(2\pi) + n \log(|\mathbf{\Sigma}|) + \sum_{t=1}^{n} \|\mathbf{Y}_t - \boldsymbol{\mu}_t\|^2_{\mathbf{\Sigma}^{-1}}$$

Factor model

Factor model. Factor model refers to the rewriting of the covariance matrix: if $\mathbf{\Sigma}$ can be written as

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Psi}, \qquad \text{with } \mathbf{B} = [b_{qm}] : M \times Q, \qquad Q < M,$$

then the model can be rewritten as

$$Y_{tm} = \mu_{km} + \sum_{q=1}^{Q} Z_{tq} b_{qm} + E_{tm}, \qquad \forall t \in I_k^m$$

where $\mathbf{Z}_t$ i.i.d. $\sim \mathcal{N}_Q(\mathbf{0}, \mathbf{I})$, $\mathbf{E}_t$ i.i.d. $\sim \mathcal{N}_M(\mathbf{0}, \sigma^2\mathbf{I})$, $(\{\mathbf{Z}_t\}, \{\mathbf{E}_t\})$ indep, $\mathbf{\Psi} = \sigma^2\mathbf{I}$.

Factor model

Factor model. Factor model refers to the rewriting of the covariance matrix: if $\mathbf{\Sigma}$ can be written as

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Psi}, \qquad \text{with } \mathbf{B} = [b_{qm}] : M \times Q, \qquad Q < M,$$

then the model can be rewritten as

$$Y_{tm} = \mu_{km} + \sum_{q=1}^{Q} Z_{tq} b_{qm} + E_{tm}, \qquad \forall t \in I_k^m$$

where $\mathbf{Z}_t$ i.i.d. $\sim \mathcal{N}_Q(\mathbf{0}, \mathbf{I})$, $\mathbf{E}_t$ i.i.d. $\sim \mathcal{N}_M(\mathbf{0}, \sigma^2\mathbf{I})$, $(\{\mathbf{Z}_t\}, \{\mathbf{E}_t\})$ indep, $\mathbf{\Psi} = \sigma^2\mathbf{I}$.

Matrix form. The model can be written as

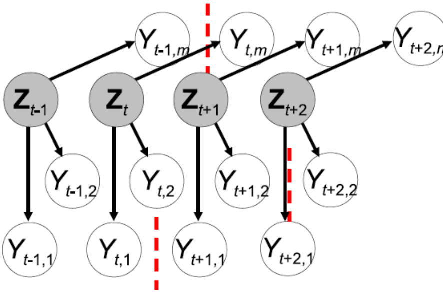
$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{Z}\mathbf{B}' + \mathbf{E}$$

where $\mathbf{T}$ corresponds to the breakpoint positions and $\boldsymbol{\mu}$ contains the segment means.
Parameters to estimate. $\phi = (\mathbf{T}, \boldsymbol{\mu}, \mathbf{B}, \sigma^2)$.

Factor model

Graphical model.

The dependency is modeled thought Z.

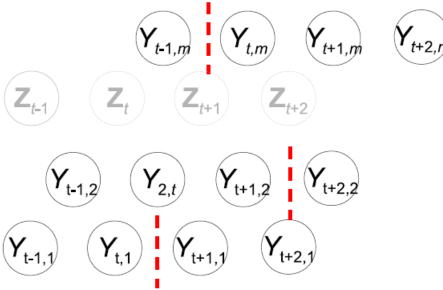


Factor model

Graphical model.

The dependency is modeled thought Z.

Conditional likelihood: because $\mathbf{\Psi}$ is diagonal, $\log \mathcal{L}(\mathbf{Y}|\mathbf{Z})$ is additive w.r.t. the segments.



$$-2\sum_t \log \mathcal{L}(\mathbf{Y}_t|\mathbf{Z}_t;\phi) \quad = \quad N\log(2\pi) + n\log(|\mathbf{\Psi}|) + \sum_{t=1}^{n} \|\mathbf{Y}_t - \boldsymbol{\mu}_t - \mathbf{Z}_t\mathbf{B}'\|^2_{\mathbf{\Psi}^{-1}}$$

and

$$\sum_{t=1}^{n} \|\mathbf{Y}_t - \mu_t - \mathbf{Z}_t\mathbf{B}'\|^2_{\mathbf{\Psi}^{-1}} \quad = \quad \sum_{m=1}^{M}\sum_{k=1}^{K_m}\sum_{t=t^m_{k-1}+1}^{t^m_k} (Y_{tm} - \mu_{km} - \sum_q Z_{tq}b_{qm})^2/\psi_m$$

EM algorithm

E-M strategy. In presence of latent variable $\mathbf{Z}$, it consists in maximizing the conditional expectation of the complete-data likelihood cond. to $\mathbf{Y}$,

$$\mathbb{E}[\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \phi)|\mathbf{Y}] = \mathbb{E}[\log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \phi)|\mathbf{Y}] + \mathbb{E}[\log \mathcal{L}(\mathbf{Z})|\mathbf{Y}]$$

EM algorithm

E-M strategy. In presence of latent variable $\mathbf{Z}$, it consists in maximizing the conditional expectation of the complete-data likelihood cond. to $\mathbf{Y}$,

$$\mathbb{E}[\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \phi)|\mathbf{Y}] = \mathbb{E}[\log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \phi)|\mathbf{Y}] + \mathbb{E}[\log \mathcal{L}(\mathbf{Z})|\mathbf{Y}]$$

E-step: Calculate the conditional distribution $p(\mathbf{Z}|\mathbf{Y}; \phi^h)$ e.g. the moments

$$\mathbf{Z}_t^{(h)} = \mathbb{E}_{\phi^h}(\mathbf{Z}_t|\mathbf{Y}) \text{ and } \mathbb{V}_{\phi^h}(\mathbf{Z}_t|\mathbf{Y}).$$

EM algorithm

E-M strategy. In presence of latent variable $\mathbf{Z}$, it consists in maximizing the conditional expectation of the complete-data likelihood cond. to $\mathbf{Y}$,

$$\mathbb{E}[\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \phi)|\mathbf{Y}] = \mathbb{E}[\log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \phi)|\mathbf{Y}] + \mathbb{E}[\log \mathcal{L}(\mathbf{Z})|\mathbf{Y}]$$

E-step: Calculate the conditional distribution $p(\mathbf{Z}|\mathbf{Y}; \phi^h)$ e.g. the moments

$$\mathbf{Z}_t^{(h)} = \mathbb{E}_{\phi^h}(\mathbf{Z}_t|\mathbf{Y}) \text{ and } \mathbb{V}_{\phi^h}(\mathbf{Z}_t|\mathbf{Y}).$$

M-step: Update the parameter value as $\phi^{h+1} = \arg\max_\phi \mathbb{E}_{\phi^h}[\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \phi)|\mathbf{Y}]$.
Each step focus on one parameter:

Estimation of σ^2

Estimation of $\mathbf{B}$

Estimation of segmentation parameters $\mathbf{T}\boldsymbol{\mu}$.

$$\left\{\mathbf{T}^{(h+1)}, \boldsymbol{\mu}^{(h+1)}\right\} = \arg\min_{\mathbf{T},\boldsymbol{\mu}} \sum_{t=1}^{n} \|\mathbf{Y}_t - \mathbf{T}\boldsymbol{\mu}_t - \mathbf{Z}_t^{(h)}\mathbf{B}^{(h+1)})\|^2$$

$\rightarrow$ (2 stages) DP

Model selection: a heuristic

Q: dimension of the latent vectors $\{\mathbf{Z}_t\}$.

$K = \sum_m K_m$: total number of segment.

Model selection: a heuristic

Q: dimension of the latent vectors $\{\mathbf{Z}_t\}$.

$K = \sum_m K_m$: total number of segment.

Factor model. For a fixed given K:

$$BIC_K(Q) = -2 \log \mathcal{L}(\mathbf{Y}; \widehat{\mathbf{T}\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}) + [1 + Q(2M - Q + 1)/2] \log n.$$

Model selection: a heuristic

Q: dimension of the latent vectors $\{\mathbf{Z}_t\}$.

$K = \sum_m K_m$: total number of segment.

Factor model. For a fixed given K:

$$BIC_K(Q) = -2\log\mathcal{L}(\mathbf{Y}; \widehat{\mathbf{T}\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}) + [1 + Q(2M - Q + 1)/2]\log n.$$

Segmentation. [Zhang & Siegmund, 2007] proposed a modified BIC ($mBIC$) for one series. We use an extension similar to [Picard et al., 2011b]: for a given Q,

$$mBIC_Q(K) = f(K, \widehat{\boldsymbol{\Sigma}}_Q, \{n_k^m\}).$$

Model selection: a heuristic

Q: dimension of the latent vectors $\{\mathbf{Z}_t\}$.

$K = \sum_m K_m$: total number of segment.

Factor model. For a fixed given K:

$$BIC_K(Q) = -2 \log \mathcal{L}(\mathbf{Y}; \widehat{\mathbf{T}\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}) + [1 + Q(2M - Q + 1)/2] \log n.$$

Segmentation. [Zhang & Siegmund, 2007] proposed a modified BIC ($mBIC$) for one series. We use an extension similar to [Picard et al., 2011b]: for a given Q,

$$mBIC_Q(K) = f(K, \widehat{\boldsymbol{\Sigma}}_Q, \{n_k^m\}).$$

Heuristic.

$$\widehat{Q}_K = \arg\max_Q BIC_K(Q), \qquad \widehat{K} = \arg\max_K mBIC_{\widehat{Q}_K}(K), \qquad \widehat{Q} = \widehat{Q}_{\widehat{K}}.$$

Simulation study

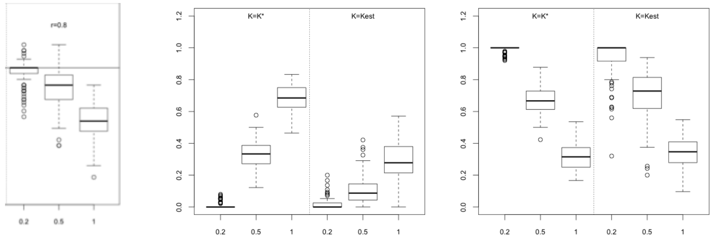
Parameters: $M = 10, n = 100$, $Q = Q^\star = M - 1$, increasing σ^2.

Simulation study

Parameters: $M = 10, n = 100$, $Q = Q^{\star} = M - 1$, increasing σ^2.

Selection of K (when $Q = \hat{Q}$)

$\hat{K} - K$ | FPR for $K = K^*$ and $K = \hat{K}$ | TPR for $K = K^*$ and $K = \hat{K}$



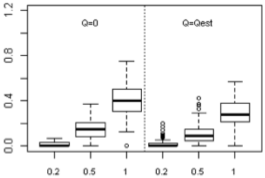
- The more difficult is the detection (large σ^2), the more under-estimated is the number of segments
- This leads to a better precision of the breakpoints (small FPR) (compared to the results with the true value of K).

Simulation study

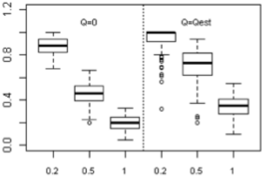
Accounting for dependency (using $\hat{K}$)

Whatever the difficulty of the detection problem, accounting for the dependency increases the performance of the segmentation (smaller FPR and larger TPR).

FPR for $Q = 0$ (segmentation only) and $Q = \hat{Q}$



TPR for $Q = 0$ (segmentation only) and $Q = \hat{Q}$



Simulation study

Selection of Q

	$(\widehat{Q}, \widehat{K})$			$(Q^*, \widehat{K})$		
σ	0.2	0.5	1	0.2	0.5	1
mean of Q	3.37	2.74	2.39	9		
$RMSE(\Sigma)$	0.005	0.032	0.119	0.0048	0.032	0.124
FPR	0.016	0.110	0.288	0.039	0.175	0.413
TPR	0.93	0.69	0.34	0.93	0.597	0.262

- The number of factors is strongly underestimated. This underestimation
  - does not alter much the estimation of Σ
  - increases the power of procedure in terms of breakpoint positioning
- $\hat{Q}$ decreases slightly with the increasing of σ^2 leading to a decreasing precision of the estimation of Σ

Summary, comments and future works

Summary. We consider separatly two types of dependence (within-series and between-series) in a segmentation model and propose two different strategies to infer the parameter such that DP can be applied.

Segmentation of multiple series.

- (Practical) identifiability. Although the model is theoretical identifiable, simultaneous breakpoints in a large fraction of series can be confounded with the 'random effect' $\mathbf{Z}_t$ (see [Picard et al., 2011b]).
- (Practical) Exploration of the (K, Q) space. The 2 BIC criteria cannot be combined into a single one, as they do not rely on the same contrast (likelihood vs within/between sum of squares).
  $\longrightarrow$ Time consuming exploration over a grid for (K, Q).
- (Theoretical) The proposed strategy is only heuristic. No criterion with theoretical guaranty exists at this time for such a dependency structure.

Summary, comments and future works

Time-dependence Gaussian process.

- Other estimators of ρ can be derived. It suffices that it has the good asymptotic property.
- For AR(1): Package R AR1seg.
- S. Chakar recently considered to model the dependence with an AR(p) and proposed the same decorrelation-type strategy.

Global model. Consider a joint segmentation model with these two dependencies.

References

- Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* 47-78.
- Bardet, J.M., Kengne, W. C. and Wintenberger, O. (2012). Detecting multiple change-points in general causal time series using penalized quasi-likelihood. *EJS* 1-50.
- Chakar, S., Lebarbier, E., Lévy-Leduc, C. and Robin, S. (2015). A robust approach for estimating change-points in the mean of an AR(1) process. *to appear in Bernouilli*.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J.R. Statist. Soc. B.* 39 1-38.
- Friguet, C., Kroareg, M. and Causeur, D. (2009). A factor model approach for multiple testing under independence. *J. Amer. Statist. Assoc.* 104(488) 1406-15. doi:10.1198/jasa.2009.tm08332.
- Lavielle, M. (2005). Using a penalized contrasts for the change-point problem. *Signal Proc.* 85 1501-1510.
- Lavielle, M. and Moulines, E. (2000). Least-squares Estimation of an Unknown Number of Shifts in a Time Series. *Journal of time series analysis* 21(1) 33-59.
- Lebarbier, E. (2005). Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection. *Signal Proc.* 85 717-736.
- Ma, Yanyuan and Genton, Marc G. (2000). Highly Robust Estimation of the Autocovariance Function. *Journal of Time Series Analysis* 21(6) 1467-9892.
- Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B. and Robin, S. (2011a), Joint segmentation, calling and normalization of multiple CGH profiles. *Biostatistics*. 12(3) 413-428 .
- Picard, F., Lebarbier, E., Budinska, E. and Robin, S. (2011b), Joint segmentation of multivariate Gaussian processes using mixed linear models. *Comp. Stat. & Data Analysis*. 55(2) 1160-70.
- Rigaill, G. (2010), Pruned dynamic programming for optimal multiple change-point detection. Technical report arXiv. arXiv:1004.0887v1.
- Rousseeuw, Peter J. and Croux, Christophe (1993), Alternatives to the Median Absolute Deviation. *JASA*, 88(424) 1273-1283.
- Zhang, N. R. and Siegmund, D. O. (2007), A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*. 63(1), 22-32.