Multiple Change Point Detection by Sparse Parameter Estimation

Jiří Neubauer and Vítězslav Veselý

Department of Econometrics
Fac. of Economics and Management
University of Defence Brno,
Czech Republic

Dept. of Appl. Math. and Comp. Sci.
Fac. of Economics and Administration
Masaryk University Brno,
Czech Republic

COMPSTAT 2010, August 22 – 27
Paris, FRANCE

Research supported GAČR P402/10/P209 and MSM0021622418

Outline

- Introduction
- Heaviside Dictionary for Change Point Detection
- Change Point Detection by Basis Pursuit
- Multiple Change Point Detection by Basis Pursuit
- References

Introduction

- Chen, S. S. et al. (1998) proposed a new methodology based on basis pursuit for spectral representation of signals (vectors). Instead of just representing signals as superpositions of sinusoids (the traditional Fourier representation) they suggested alternate dictionaries – collections of parametrized waveforms – of which the wavelet dictionary is only the best known.
- A recent review paper by Bruckstein et al. (2009) demonstrates a remarkable progress in the field of sparse modeling since that time. Theoretical background for such systems (also called frames) can be found for example in Christensen, O. (2003).
- In traditional Fourier expansion a presence of jumps in the signal slows down the convergence rate preventing sparsity. The Heaviside dictionary (see Chen et al. (1998)) merged with the Fourier or wavelet dictionary can solve the problem quite satisfactorily.

- A lot of other useful applications in a variety of problems can be found in Veselý and Tonner (2005), Veselý et al. (2009) and Zelinka et al. (2004).
- In Zelinka et al. (2004) kernel dictionaries showed to be an effective alternative to traditional kernel smoothing techniques.

In this paper we are using Heaviside dictionary in the same manner to denoise signal jumps (discontinuities) in the mean.
Consequently, the basis pursuit approach can be proposed as an alternative to conventional statistical techniques of change point detection (see Neubauer and Veselý (2009)). The mentioned paper is focused on using the basis pursuit algorithm with the Heaviside dictionary for one change point detection.
This paper presents results of an introductory empirical study for the simplest case of detecting two change points buried in additive gaussian white noise.

Introduction – Linear expansions in a separable Hilbert space

In what follows we use terminology and notation common in the *theory of frames* (cf. Christensen, 2003)

Dictionary (frame) and atomic decomposition

- H closed separable subspace of a bigger Hilbert space $X(\langle \cdot, \cdot \cdot \rangle)$ over $\mathbb{R}$ or $\mathbb{C}$ with induced norm $\|\cdot\| := \sqrt{\langle \cdot, \cdot \rangle}$,
- $G := \{G_j\}_{j \in J} \subset H$, $\|G_j\| = 1$ (or bounded), $\mathrm{card}\, J \leq \aleph_0$ complete in H such that any $x \in H$ can be expanded via a sequence of **spectral coefficients** $\xi = \{\xi_j\}_{j \in J} \in \ell^2(J)$

$$x = \sum_{j \in J} \xi_j G_j =: \underbrace{[G_1, G_2, \ldots]}_{=:T} \xi \qquad (1)$$

  where summation is unconditional (oder-independent) with respect to the norm $\|\cdot\|$ and clearly defines a surjective linear operator $T : \ell^2(J) \to H$.
- G is called **dictionary** or **frame** in H, expansion (1) **atomic decomposition** and G_j its **atoms**.

Introduction – Sparse Estimators

- IDEAL SPARSE ESTIMATOR may be formulated as a solution of the NP-hard combinatorial problem:

$$\xi^* = \operatorname{argmin}_{T\xi \in \mathcal{O}_\varepsilon(\widehat{x})} \|\xi\|_0^0 \text{ where } \|\xi\|_0^0 = \operatorname{card}\{j \in J \mid \xi_j \neq 0\} < \infty.$$

- ℓ_p-OPTIMAL SPARSE ESTIMATOR $(0 < p \leq 1)$ can reduce the computational complexity by solving simpler programming problem which is either nonlinear nonconvex (with $0 < p < 1$) or linear convex (with $p = 1$) approximation of the above in view of $\|\xi\|_p^p \to \|\xi\|_0^0$ for $p \to 0$:

$$\xi^* = \operatorname{argmin}_{T\xi \in \mathcal{O}_\varepsilon(\widehat{x})} \|\xi\|_{p,\mathbf{w}}^p \text{ where } \|\xi\|_{p,\mathbf{w}}^p = \sum_{j \in J} w_j |\xi_j|^p < \infty.$$

  ℓ_1-optimal sparse estimator=**Basis Pursuit Algorithm (BPA)** by [Chen & Donoho & Saunders, 1998].

The weights $\mathbf{w} = \{w_j\}_{j \in J}$, $w_j > 0$, have to be chosen appropriately to balance contribution of individual parameters to the model on $\mathcal{O}_\varepsilon(\widehat{x})$. (If $\|G_j\| = 1$ for all j, then $w_j = 1$.)

In this section we propose the method based on basis pursuit algorithm (BPA) for the detection of the change point in the sample path $\{y_t\}$ in one dimensional stochastic process $\{Y_t\}$. We assume a deterministic functional model on a bounded interval $\mathcal{I}$ described by the dictionary $G = \{G_j\}_{j \in J}$ with atoms $G_j \in L^2(\mathcal{I})$ and with additive white noise e on a suitable finite discrete mesh $\mathcal{T} \subset \mathcal{I}$:

$$Y_t = x_t + e_t, \ t \in \mathcal{T},$$

where $x \in \mathrm{sp}(\{G_j\}_{j \in J})$, $\{e_t\}_{t \in \mathcal{T}} \sim WN(0, \sigma^2)$, $\sigma > 0$, and J is a big finite indexing set.

Smoothed function $\hat{x} = \sum_{j \in J} \hat{\xi}_j G_j =: \mathbf{G}\hat{\xi}$ minimizes on $\mathcal{T}$ ℓ^1-penalized optimality measure $\frac{1}{2}\|\mathbf{y} - \mathbf{G}\xi\|^2$ as follows:

$$\hat{\xi} = \operatorname{argmin}_{\xi \in \ell^2(J)} \frac{1}{2}\|\mathbf{y} - \mathbf{G}\xi\|^2 + \lambda\|\xi\|_1, \;\; \|\xi\|_1 := \sum_{j \in J} \|G_j\|_2 \xi_j,$$

where $\lambda = \sigma\sqrt{2 \ln(\mathrm{card}\, J)}$ is a smoothing parameter chosen according to the soft-thresholding rule commonly used in wavelet theory. This choice is natural because one can prove that with any orthonormal basis $G = \{G_j\}_{j \in J}$ the shrinkage via soft-thresholding produces the same smoothing result $\hat{x}$. (see Bruckstein et al. (2009)). Such approaches are also known as basis pursuit denoising (BPDN).

Solution of this minimization problem with λ close to zero may not be sparse enough: we are searching small $F \subset J$ such that $\hat{x} \approx \sum_{j \in F} \hat{\xi}_j G_j$ is a good approximation. That is why we apply the following four-step procedure described in Zelinka et al. (2004) in more detail and implemented in Veselý (2001–2008).

Heaviside Dictionary for Change Point Detection

(A0) Choice of a raw initial estimate $\xi^{(0)}$, typically $\xi^{(0)} = \mathbf{G}^{+}\mathbf{y}$.

(A1) We improve $\xi^{(0)}$ iteratively by stopping at $\xi^{(1)}$ which satisfies optimality criterion BPDN. The solution $\xi^{(1)}$ is optimal but not sufficiently sparse in general (for small values of λ).

(A2) Starting with $\xi^{(1)}$ we are looking for $\xi^{(2)}$ by BPA which tends to be nearly sparse and is optimal.

(A3) We construct a sparse and optimal solution ξ^{*} by removing negligible parameters and corresponding atoms from the model, namely those satisfying $|\xi_j^{(2)}| < \alpha \|\xi^{(2)}\|_1$ where $0 < \alpha << 1$ is a suitable sparsity level, a typical choice being $\alpha = 0.05$ following an analogy with the statistical significance level.

(A4) We repeat the step (A1) with the dictionary reduced according to the step (A3) and with a new initial estimate $\xi^{(0)} = \xi^{*}$. We expect to obtain a possibly improved sparse estimate ξ^{*}.

Hereafter we refer to this four-step algorithm as to BPA4. The steps (A1), (A2) and (A4) use Primal-Dual Barrier Method designed by M. Saunders (see Saunders (1997–2001)). This up-to-date sophisticated algorithm allows one to solve fairly general optimization problems minimizing convex objective subject to linear constraints. A lot of controls provide a flexible tool for adjusting the iteration process.

Heaviside Dictionary for Change Point Detection

We build our dictionary from heaviside-shaped atoms on $L^2(\mathbb{R})$ derived from a fixed 'mother function' via shifting and scaling following the analogy with the construction of wavelet bases. We construct an oversized shift-scale dictionary $G = \{G_{a,b}\}_{a \in \mathcal{A}, b \in \mathcal{B}}$ derived from the 'mother function' by varying the shift parameter a and the scale (width) parameter b between values from big finite sets $\mathcal{A} \subset \mathbb{R}$ and $\mathcal{B} \subset \mathbb{R}^+$, respectively ($J = \mathcal{A} \times \mathcal{B}$), on a bounded interval $\mathcal{I} \subset \mathbb{R}$ spanning the space $H = \mathrm{sp}(\{G_{a,b}\})_{a \in \mathcal{A}, b \in \mathcal{B}}$, where

$$G_{a,b}(t) = \begin{cases} 1 & \text{for} \quad t - a > b/2, \\ 2(t-a)/b & \qquad\;\; |t-a| \le b/2, b > 0, \\ 0 & \qquad\;\; t = a, b = 0, \\ -1 & \qquad\;\; \text{otherwise.} \end{cases}$$

Heaviside Dictionary for Change Point Detection

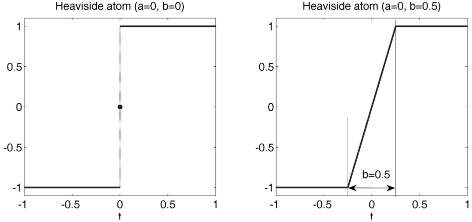


Figure: Heaviside atoms with parameters $a = 0, b = 0$ and $a = 0, b = 0.5$

In the simulations below $\mathcal{I} = [0, 1]$, $\mathcal{T} = \{t/T\}$ (typically with mesh size $T = 100$), $\mathcal{A} = \{t/T\}_{t=t_0}^{T-t_0}$ (t_0 is a boundary trimming, $t_0 = 4$ was used in the simulations) and scale b fixed to zero ($\mathcal{B} = \{0\}$). Clearly the atoms of such Heaviside dictionary are normalized on $\mathcal{I}$, i.e. $\|G_{a,0}\|_2 = 1$.

Neubauer and Veselý (2009) proposed the method of change point detection if there is just one change point in a one-dimensional stochastic process (or in its sample path). We briefly describe a given method. We would like to find a change point in a stochastic process

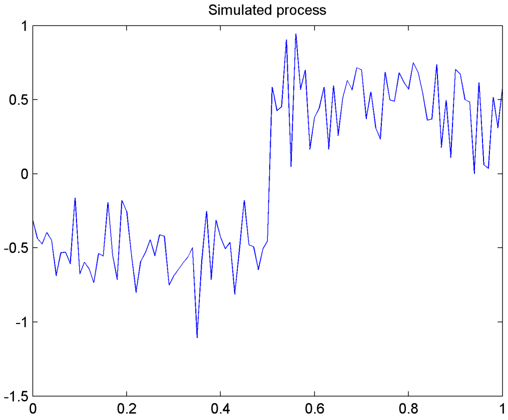
$$Y_t = \begin{cases} \mu + \epsilon_t & t = 1, 2, \ldots, c \\ \mu + \delta + \epsilon_t & t = c+1, \ldots, T, \end{cases} \tag{2}$$

where $\mu, \delta \neq 0, t_0 \leq c < T - t_0$ are unknown parameters and ϵ_t are independent identically distributed random variables with zero mean and variance σ^2. The parameter c indicates the change point in the process.

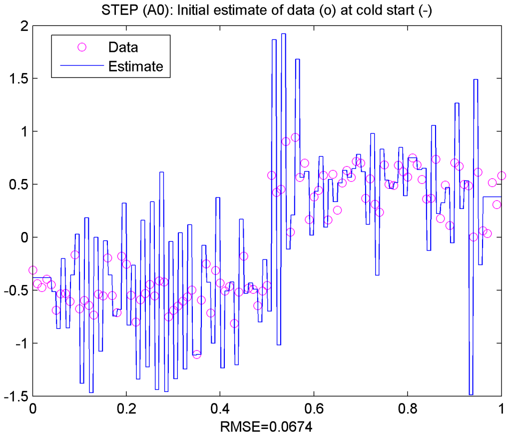
Change Point Detection by Basis Pursuit

Using the basis pursuit algorithm we obtain some significant atoms, we calculate correlation between significant atoms and analyzed process. The shift parameter of the atom with the highest correlation is taken as an estimator of the change point c.

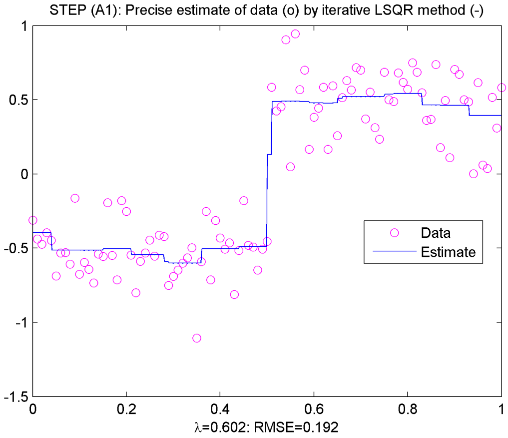
BP algorithm - example



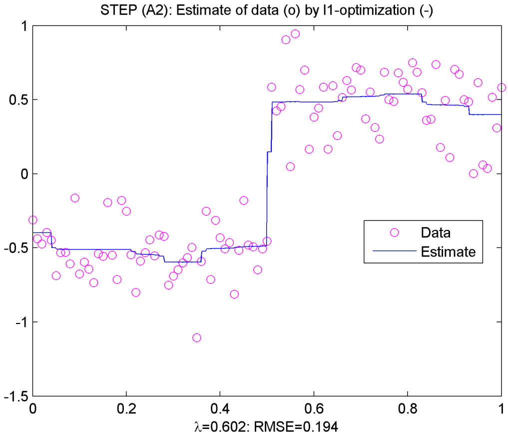
BP algorithm - example



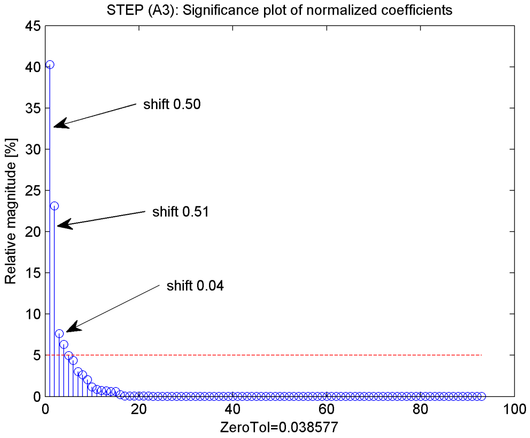
BP algorithm - example



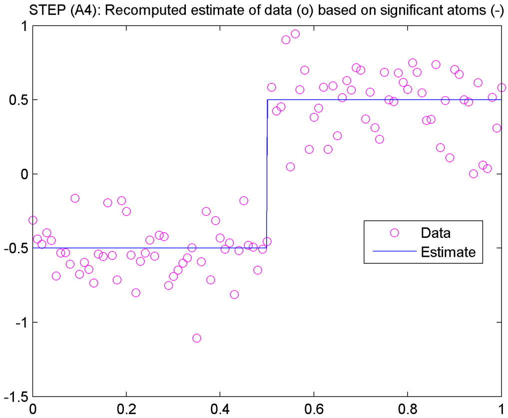
BP algorithm - example



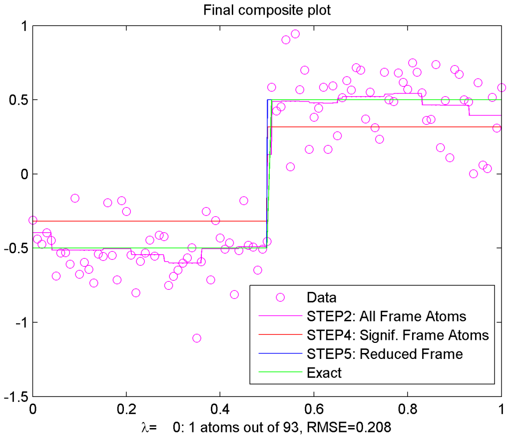
BP algorithm - example



BP algorithm - example



BP algorithm - example



Now let us assume the model with two change points

$$Y_t = \begin{cases} \mu + \epsilon_t & t = 1, 2, \ldots, c_1 \\ \mu + \delta_1 + \epsilon_t & t = c_1 + 1, \ldots, c_2, \\ \mu + \delta_2 + \epsilon_t & t = c_2 + 1, \ldots, T, \end{cases} \tag{3}$$

where $\mu, \delta_1, \delta_2 \neq 0$, $t_0 \leq c_1 < c_2 < T - t_0$ are unknown parameters and ϵ_t are independent identically distributed random variables with zero mean and variance σ^2.

We use the method of change point estimation described above for detection two change points c_1 and c_2 in the model (3). Instead of finding only one significant atom with the highest correlation with the process Y_t we can identify two significant atoms with the highest correlation. The shift parameters of these atoms determine estimators for the change points c_1 and c_2.

Another possibility is to apply the procedure of one change point detection two times in sequence. In the first step we identity one change point in the process Y_t, then we subtract given significant atom from the process (by linear regression)

$$Y_t = \beta G_{0,\hat{c}_1} + e_t$$
$$Y_t' = Y_t - \hat{\beta} G_{0,\hat{c}_1}$$

and finally we apply the method to the new process Y_t'. The shift parameters of selected atoms are again identifiers of the change points c_1 and c_2. Observe that this can be seen as two steps of orthogonal matching pursuit (OMP) combined with BPA.

We demonstrate the method of multiple change point detection by BPA4 on simulations of the process (3) with the change points $c_1 = 30$ and $c_2 = 70$, $T = 100$, $\mu = -1, \delta_1 = 1, \delta_2 = 2$ and $\sigma = 0.5$ (for BPA4 we transform it to the interval [0,1]).

Multiple Change Point Detection by Basis Pursuit

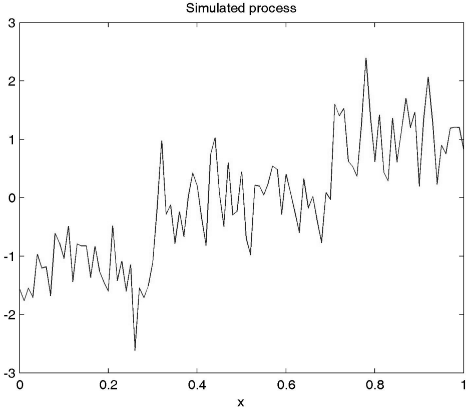


Figure: Simulated process

After first applying the method of one change point detection, we get estimate $\hat{c} = 30$. Using linear regression we can subtract this identified atom from the process and repeat the procedure. We obtain estimation of the second change point $\hat{c}_2 = 70$. We use linear regression model

$$y_t = \beta_1 G_{0.3,0} + \beta_2 G_{0.7,0} + u_t,$$

to get final fit on the simulated process. We obtain $\hat{\beta}_1 = 0.651$ with the confidence interval $[0.459, 0.664]$ and $\hat{\beta}_2 = 0.613$ with the confidence interval $[0.511, 0.716]$, see figure 3. Dashed lines denote 95% confidence and prediction intervals.

Multiple Change Point Detection by Basis Pursuit

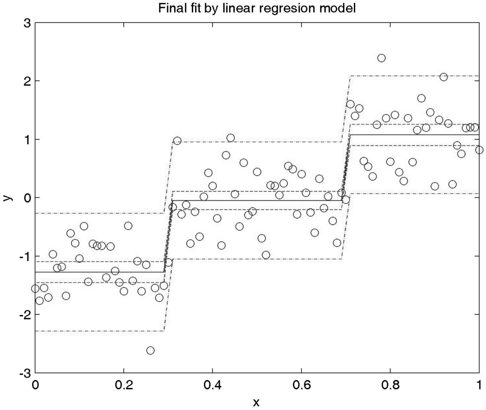


Figure: Final fit by linear regression model

For the purpose of introductory performance study of the proposed method of multiple change point detection we use simulations of the process (3). We put, analogously to the example, $\mu = -1, \delta_1 = 1, \delta_2 = 2$ and $T = 100$ where the error terms are independent normally distributed with zero mean and the standard deviations $\sigma = 0.2, 0.5, 1$ and 1.5, respectively. We calculate simulations of this model with change points $c_1 = 30$ and $c_2 = 70$ (500 simulations for each choice of standard deviation). We preferred the second method of multiple change point detection (an application the procedure of one change point detection two times in succession) which proved to be more suitable.

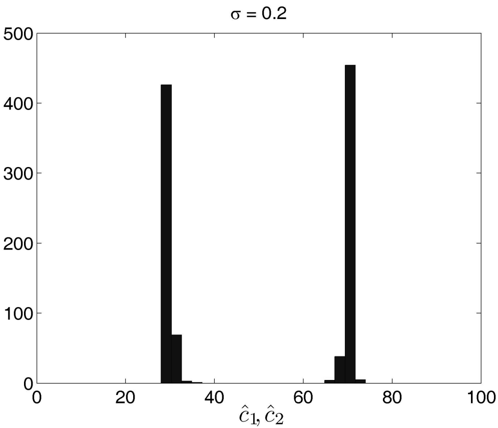


Figure: Histograms of estimated change points for $\sigma = 0.2$

Multiple Change Point Detection by Basis Pursuit

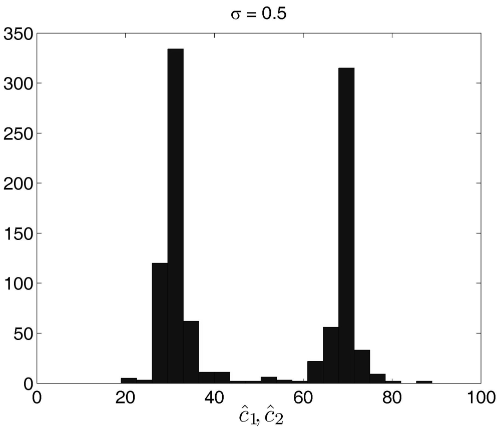


Figure: Histograms of estimated change points for $\sigma = 0.5$

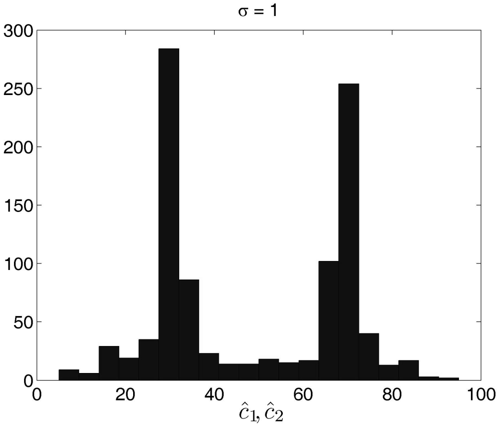


Figure: Histograms of estimated change points for $\sigma = 1$

Multiple Change Point Detection by Basis Pursuit

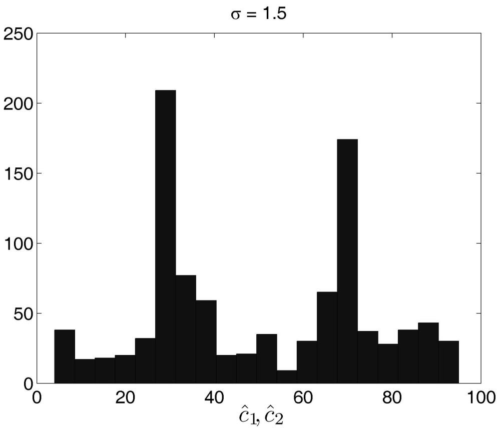


Figure: Histograms of estimated change points for $\sigma = 1.5$

Multiple Change Point Detection by Basis Pursuit

The model (3) can be easily extended to more than two change points. The number of the change points is in real situation unknown. Using the BP approach we assume that if there are any change points, we can detect significant atoms in BPA4 algorithm. In case there is not a significant atom, change point cannot be detected. In the first step we identify one change point, then we subtract given significant atom from the process by linear regression (according to the procedure of two change points detection mentioned above). If there are some significant atoms in the new process, we find the atom with highest correlation and subtract it from the new process etc. We stop when it is not possible to detect any significant atom.

Conclusion

According to the introductory simulation results the basis pursuit approach proposes a reasonable detection method of two change points in one-dimensional process. The outlined method can be used for detection of two or more change points, or another sort of change point with a dictionary G of different kind.
The change point detection techniques may be useful for instance in modeling of economical or environmental time series where jumps can occur.

References

- BRUCKSTEIN, A. M. et al. (2009): From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review 51 (1), 34–81.*
- CHEN, S. S. et al. (1998): Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput. 20 (1), 33-61* (2001 reprinted in *SIAM Review 43 (1), 129–159).*
- CHRISTENSEN, O. (2003): *An introduction to frames and Riesz bases*. Birkhuser, Boston-Basel-Berlin.
- NEUBAUER, J. and VESELÝ, V. (2009): Change Point Detection by Sparse Parameter Estimation. In: *The XIIIth International conference: Applied Stochastic Models and Data Analysis*. Vilnius, 158–162.
- SAUNDERS, M. A. (1997–2001): *pdsco.m: MATLAB code for minimizing convex separable objective functions subject to* $Ax = b, x \geq 0$.

References

- VESELÝ, V. (2001–2008): *framebox: MATLAB toolbox for overcomplete modeling and sparse parameter estimation.*
- VESELÝ, V. and TONNER, J. (2005): Sparse parameter estimation in overcomplete time series models. *Austrian Journal of Statistics, 35 (2&3), 371–378*.
- VESELÝ, V. et al. (2009): Analysis of PM10 air pollution in Brno based on generalized linear model with strongly rank-deficient design matrix. *Environmetrics, 20 (6), 676–698*.
- ZELINKA et al. (2004): Comparative study of two kernel smoothing techniques. In: Horová, I. (ed.) *Proceedings of the summer school DATASTAT'2003, Svratka.* Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica 15: Masaryk University, Brno, Czech Rep., 419–436.