HYBRID BAYESIAN VARIATIONAL SCHEME TO HANDLE PARAMETER SELECTION IN TOTAL VARIATION SIGNAL DENOISING

Jordan Frecon[(1)]*, Nelly Pustelnik*[(1)]*, Nicolas Dobigeon*[(2)]*, Herwig Wendt*[(2)] *and Patrice Abry*[(1)]

[(1)] Physics Dept., CNRS UMR 5672, ENS Lyon, France, `firstname.lastname@ens-lyon.fr`
[(2)] IRIT, CNRS UMR 5505, INP-ENSEEIHT Toulouse, France, `firstname.lastname@irit.fr`

ABSTRACT

Change-point detection problems can be solved either by variational approaches based on total variation or by Bayesian procedures. The former class leads to small computational time but requires the choice of a regularization parameter that significantly impacts the achieved solution and whose automated selection remains a challenging problem. Bayesian strategies avoid this regularization parameter selection, at the price of high computational costs. In this contribution, we propose a hybrid Bayesian variational procedure that relies on the use of a hierarchical Bayesian model while preserving the computational efficiency of total variation optimization procedures. Behavior and performance of the proposed method compare favorably against those of a fully Bayesian approach, both in terms of accuracy and of computational time. Additionally, estimation performance are compared to the Stein unbiased risk estimate, for which the knowledge of the noise variance is needed.

***Index Terms*—** Parameter selection, total variation, convex optimization, hierarchical Bayesian model.

1. INTRODUCTION

Change-point detection problems are of considerable potential interest in many different applications ranging e.g, from econometrics to signal processing (see [1, 2], for an overview). Formally, a change-point detection problem consists in estimating a piecewise constant signal $\mathbf{x} \in \mathbb{R}^N$ from noisy observations $\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ denotes an additive degradation.

To solve this problem, variational methods based on total variation have received considerable interest and research efforts over the past years (see, e.g., [3] for genomic data processing). They aim at providing an estimate for $\mathbf{x}$ by minimizing the following non-smooth convex criterion by iterative strategies [4–7] or straightforward computations [8, 9]:

$$\mathbf{x}_\lambda^* = \arg\min_{\mathbf{u}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{u}-\mathbf{y}\|_2^2 + \lambda \sum_{i=1}^{N-1} |u_{i+1}-u_i| \qquad (1)$$

where $\mathbf{u} = \big(u_i\big)_{1\le i\le N}$ and λ is a (positive) regularization parameter that controls the trade-off between the fit to the observations $\mathbf{y}$ and the amount of regularization. The selection of λ is critical for the performance of (1) because the solution $\mathbf{x}_\lambda^*$ strongly depends on its value. Indeed, for $\lambda \to 0$, the first term in (1) dominates and $\mathbf{x}_\lambda^*$ consists of many constant segments, resulting in large variance and overestimation of the number of segments in $\mathbf{x}$. To the contrary, when $\lambda \to +\infty$, the regularization term is dominant and (1) yields a solution with very few constant segments, small variance, and large bias. Currently existing methods for selecting a value for λ rely on the *Stein unbiased risk estimate* (SURE) [10, 11], which minimizes the unbiased estimator of the mean squared error between $\mathbf{x}$ and $\mathbf{x}_\lambda^*$. While this approach is effective, it requires the knowledge of the variance of $\boldsymbol{\epsilon}$, which is often unavailable.

As a second important class of methods for the estimation of $\mathbf{x}$, Bayesian approaches were intensively studied and successfully applied (cf., e.g, [12, 13]). They rely on the choice of appropriate prior distributions for the unknown parameters of $\mathbf{x}$ (e.g., the number of change points, the value of the signal on each segment,...) and of a model for the noise $\boldsymbol{\epsilon}$. Estimation of the unknown parameters is then performed on the basis of their posterior distribution. By adopting a hierarchical strategy with additional hyperparameters, no tuning of the parameters of the prior and noise distributions is needed, see, e.g., [12]. However, Bayesian methods suffer from a high computational cost which stems from the fact that the posterior distribution cannot, in most cases, be expressed analytically and needs to be approximated numerically by sampling methods such as *Markov chain Monte Carlo* (MCMC) to construct estimators for the unknown parameters.

In the present contribution, we propose a *hybrid Bayesian variational* (HBV) method that combines the advantages of both worlds. It relies on the straightforward computation of the solution of (1), from which we can extract a parametric vector depending on the regularization parameter λ. The optimal parameter vector and its associated estimate of $\mathbf{x}$ are then determined automatically based on the maximization of the posterior distribution of a Bayesian hierarchical model for which knowledge of the noise variance is not needed. The performance of the proposed hybrid procedure compare fa-

This work was supported by the CNRS GdR ISIS under the junior research project GALILEO.

vorably against those of a fully Bayesian approach, both in terms of estimation accuracy and computational cost. It is also compared against the benchmark SURE that assumes the a priori knowledge of the noise variance.

The remainder of this contribution is organized as follows. In Section 2, we describe the proposed HBV denoising algorithm. Numerical experiments validating and illustrating the proposed method are conducted in Section 3. Conclusions are drawn in Section 4.

2. HYBRID BAYESIAN VARIATIONAL DENOISING

2.1. Parametric model

To formalize the change point detection problem and the automated regularization parameter selection, we explicitly model the piecewise constant signal $\mathbf{x} \in \mathbb{R}^N$ as a signal constituted of K segments, with corresponding constant values μ_k, $k = 1, \ldots, K$, referred to as the vector $\boldsymbol{\mu} = (\mu_k)_{1\leq k\leq K}$. Use is also made of the change-point indicator vector $\mathbf{r} = \left(r_i\right)_{1\leq i\leq N}$,

$$r_i = \begin{cases} 1, & \text{if there is a change-point at time instant } i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Convention $r_N = 1$ ensures that the number K of segments is equal to the number of change-points, i.e., $K = \sum_{i=1}^{N} r_i$. By definition, $r_i = 1$ indicates that x_i is the last sample belonging to the current segment, and thus that x_{i+1} belongs to next segment. Equivalently, we can deduced from $\mathbf{r}$ the set of time indices $\mathcal{I}_k \subset \{1, \ldots, N\}$ corresponding to the k-th segment for $k = 1, \ldots, K$, such that $\mathcal{I}_k \cap \mathcal{I}_{k'} = \{\emptyset\}$ for $k \neq k'$ and $\cup_{k=1}^{K} \mathcal{I}_k = \{1, \cdots, N\}$. Following [14], we also introduce the pixel-wise change occurrence probability p. In addition, the noise $\boldsymbol{\epsilon}$ is assumed to be zero-mean and with constant variance σ^2. Therefore, observations $\mathbf{y}$ depend upon the vector parameter $\boldsymbol{\Theta} = \{\mathbf{r}, \boldsymbol{\mu}, \sigma^2, p\}$.

2.2. Regularization parameter selection

The strategy of the proposed HBV denoising algorithm can now be summarized in the following steps. First, a TV denoising algorithm is used to solve (1) for a large number of candidate values $\lambda \in \Lambda$. Then, from each recovered solution $\mathbf{x}^*_\lambda$, we derive an estimate $\widehat{\boldsymbol{\Theta}}_\lambda$ of the parameter vector $\boldsymbol{\Theta}$. Finally, the optimal value λ_{opt} of λ is chosen as the value λ for which $\widehat{\boldsymbol{\Theta}}_\lambda$ maximizes the posterior, i.e.,

$$\lambda_{\text{opt}} \in \underset{\lambda\in\Lambda}{\text{Argmax}}\, f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y}) \tag{3}$$

where the notation $\in$ Argmax indicates that the solution is not necessarily unique. To model the posterior distribution $f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$, we have recourse to a hierarchical Bayesian model, detailed in next section.

2.3. Hierarchical Bayesian model

We assume that the ϵ_i, for $i = \{1, \ldots, N\}$, are i.i.d. zero mean Gaussian variables with constant (but unknown) variance σ^2: $\mathcal{N}\left(0, \sigma^2\right)$. The joint likelihood function of the observations $\mathbf{y}$, depending on the piecewise constant model and noise parameters $\{\mathbf{r}, \boldsymbol{\mu}, \sigma^2\}$, then reads:

$$f\left(\mathbf{y}|\mathbf{r}, \boldsymbol{\mu}, \sigma^2\right)= \prod_{k=1}^{K}\prod_{i\in\mathcal{I}_k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mu_k\right)^2}{2\sigma^2}\right). \tag{4}$$

To model the posterior distribution, further assumptions are needed. Following [12], it is first assumed that $\mathbf{r}$ are a priori independent and distributed according to a Bernoulli distribution with parameter p, quantifying the prior probability of having a change-point at a given location:

$$f(\mathbf{r}|p) = \prod_{i=1}^{N} p^{r_i}(1-p)^{1-r_i}. \tag{5}$$

Furthermore, a Beta(α_0, α_1) distribution is assigned to p:

$$f\left(p|\alpha_0, \alpha_1\right) = \frac{\Gamma\left(\alpha_0, \alpha_1\right)}{\Gamma\left(\alpha_0\right)\Gamma\left(\alpha_1\right)} p^{\alpha_1 - 1}(1-p)^{\alpha_0 - 1}. \tag{6}$$

Segment values μ_k, for $k = \{1, \ldots, K\}$, are also assumed to be i.i.d., distributed according to a common Gaussian distribution $\mathcal{N}(\mu_0, \sigma_0^2)$:

$$f\left(\boldsymbol{\mu}|\mu_0, \sigma_0^2\right) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu_k - \mu_0)^2}{2\sigma_0^2}\right). \tag{7}$$

Finally, a non-informative Jeffreys prior is assigned to the noise variance σ^2:

$$f\left(\sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{8}$$

With these assumptions, the posterior distribution reads:

$$f(\boldsymbol{\Theta}|\mathbf{y}) = \frac{1}{f(\mathbf{y})} f\left(\mathbf{y}|\mathbf{r}, \boldsymbol{\mu}, \sigma^2\right) f\left(\mathbf{r}|p\right) f\left(p|\alpha_0, \alpha_1\right) \times f\left(\boldsymbol{\mu}|\mu_0, \sigma_0^2\right) f\left(\sigma^2\right) \tag{9}$$

where $f\left(\mathbf{y}|\mathbf{r}, \boldsymbol{\mu}, \sigma^2\right)$, $f\left(\mathbf{r}|p\right)$, $f\left(p|\alpha_0, \alpha_1\right)$, $f\left(\boldsymbol{\mu}|\mu_0, \sigma_0^2\right)$, and $f\left(\sigma^2\right)$ have been defined in (4), (5), (6), (7), and (8), respectively.

2.4. Proposed algorithm

The proposed HBV algorithm consists in repeating for a large number of candidates $\lambda \in \Lambda$ the following three-step procedure.

Step 1 consists in computing $\mathbf{x}^*_\lambda \in \mathbb{R}^N$, the solution of (1), using a variational strategy, based on our own implementation of Condat 1D-TV algorithm [8].

Step 2 relies on the fact that, by nature, the solution $\mathbf{x}^*_\lambda$ is piecewise constant with $\widehat{K}_\lambda$ segments. Therefore, from each solution $\mathbf{x}^*_\lambda$, an estimate, denoted $\widehat{\boldsymbol{\Theta}}_\lambda = \{\widehat{\mathbf{r}}_\lambda, \widehat{\boldsymbol{\mu}}_\lambda, \widehat{\sigma}^2_\lambda, \widehat{p}_\lambda\}$, of the parameter vector $\boldsymbol{\Theta}$ involved in the hierarchical Bayesian model, can be obtained: Estimates $\widehat{\mathbf{r}}_\lambda$ of the change-point locations (or, equivalently, of the sets $(\widehat{\mathcal{I}}_{\lambda,k})_{1\le k\le \widehat{K}_\lambda}$) can be easily computed and empirical estimates of the remaining parameters can be derived as follows:

$$\widehat{\boldsymbol{\mu}}_\lambda = (\widehat{\mu}_{\lambda,k})_{1\le k\le \widehat{K}_\lambda} \text{ where } \widehat{\mu}_{\lambda,k} = \frac{1}{|\widehat{\mathcal{I}}_{\lambda,k}|}\sum_{i\in\widehat{\mathcal{I}}_{\lambda,k}} y_i, \quad (10)$$

$$\widehat{\sigma}^2_\lambda = \frac{1}{N}\sum_{k=1}^{\widehat{K}_\lambda}\sum_{i\in\widehat{\mathcal{I}}_{\lambda,k}} (y_i - \widehat{\mu}_{\lambda,k})^2, \quad (11)$$

$$\widehat{p}_\lambda = \widehat{K}_\lambda / N. \quad (12)$$

We denote $\widehat{\mathbf{x}}_\lambda = (\widehat{x}_{\lambda,i})_{1\le i\le N}$, the estimate of $\mathbf{x}$ such that:

$$(\forall k\in\{1,\dots \widehat{K}_\lambda\})(\forall i\in\widehat{\mathcal{I}}_{\lambda,k}), \quad \widehat{x}_{\lambda,i} = \widehat{\mu}_{\lambda,k}. \quad (13)$$

Note that $\widehat{\mathbf{x}}_\lambda$ is different from $\mathbf{x}^*_\lambda$ but both estimates share the same change-point locations.

Step 3 computes the posterior distribution $f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$ based on the hierarchical Bayesian model (9).

Finally, the optimal regularization parameter λ_{opt} is selected as the one that maximizes the posterior distribution (9) according to (3). Consequently we denote $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$ the estimate $\widehat{\mathbf{x}}_\lambda$ when $\lambda = \lambda_{\text{opt}}$. The HBV procedure is sketched in Algorithm 1.

Algorithm 1 HBV algorithm

```
Input: Observed signal $\mathbf{y}\in\mathbb{R}^N$.
       Predefined set of regularization parameters $\Lambda$.
       Prior parameters $\boldsymbol{\Phi} = \{\alpha_0, \alpha_1, \mu_0, \sigma_0^2\}$.
Iterations:
 1: for $\lambda\in\Lambda$ do
 2:     Estimate $\mathbf{x}^*_\lambda$ with Condat 1D-TV algorithm.
 3:     Estimate $\widehat{\boldsymbol{\Theta}}_\lambda = \{\widehat{\mathbf{r}}_\lambda, \widehat{\boldsymbol{\mu}}_\lambda, \widehat{\sigma}^2_\lambda, \widehat{p}_\lambda\}$ from $\mathbf{x}^*_\lambda$.
 4:     Compute $f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$ with (9) and compute $\widehat{\mathbf{x}}_\lambda$.
 5: end for
Output: $\lambda_{\text{opt}}\in\text{Argmax}_{\lambda\in\Lambda} f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$;
        Solution $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$.
```

3. PERFORMANCE ASSESSMENT

3.1. Experimental settings

Non informative prior parameters are used for the Bayesian hierarchical model ($\alpha_0 = \alpha_1 = 1$, from which it follows that p has an a priori uniform distribution on $[0,1]$; $\mu_0 = \frac{1}{N}\sum_{i=1}^N y_i$ and $\sigma_0^2 = \frac{1}{10}\widehat{\text{Var}}(\mathbf{y})$, where $\widehat{\text{Var}}$ stands for the variance).

First example – To illustrate and quantify the performance of the proposed HBV denoising procedure, we consider the piecewise constant signal $\mathbf{x}\in\mathbb{R}^N$ with $N = 120$ samples plotted in black in Fig.1 (top) together with the resulting data $\mathbf{y}$ (gray) for one realization of an additive Gaussian noise with signal-to-noise ratio (SNR) of 1.8 dB.

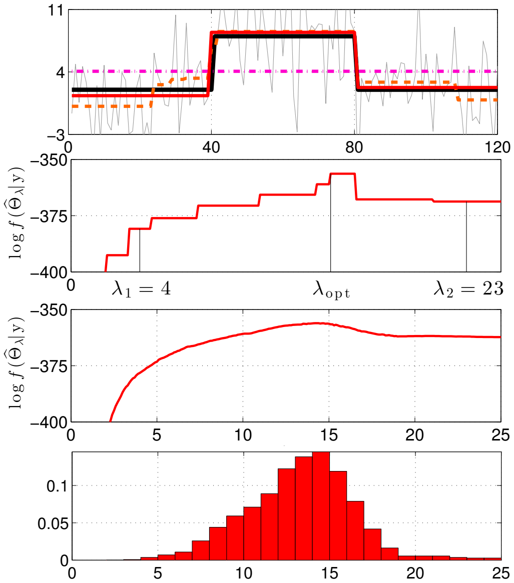

Fig. 1: **Illustration of HBV procedure.** Top row: true signal $\mathbf{x}$ (solid black), single realization of observations $\mathbf{y}$ (gray, SNR = 1.8 dB), solution $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$ (solid red), non-optimal solutions $\widehat{\mathbf{x}}_\lambda$, with $\lambda_1 = 4$ (dashed orange) and $\lambda_2 = 23$ (mixed magenta); $\log f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$ evaluated for the set Λ (second row). Ensemble average of $\log f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$ (third row) and empirical distribution of λ_{opt} (bottom row).

Second example – Complementary results arise from the study of another example, where $\mathbf{x}$ is made of $N = 240$ samples whose segments are 40 samples long with different amplitudes μ_k, and whose interest will be shown in the following.

3.2. Illustration of regularization parameter selection

In Fig. 1 (top), the selected optimal solution $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$ (solid red) is compared against two non-optimal solutions $\widehat{\mathbf{x}}_\lambda$ obtained with $\lambda_1 = 4$ (dashed orange) and $\lambda_2 = 23$ (mixed magenta), respectively, for one single realization. While the non-optimal solutions clearly fail to reproduce the true signal $\mathbf{x}$, the solution $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$ obtained with the proposed procedure provides a visually good estimate for $\mathbf{x}$. In Fig.1 (second row), the corresponding log posterior distribution $\log f(\widehat{\boldsymbol{\Theta}}_\lambda|\mathbf{y})$ is plotted as a function of the regularization parameter λ. The piecewise behavior of the posterior stems from the fact that the solutions $\mathbf{x}^*_\lambda$ have the same discontinuities and are identical for ranges of λ. We denote by Λ_{opt} the range of values of λ for which the posterior is maximal and define, by convention, λ_{opt} as the smallest value in Λ_{opt}. We note that the log posterior distribution for a single realization provides a relevant approximation for the ensemble average of the posterior distribution, plotted in Fig. 1 (third row) for 100 realizations. The empiri-

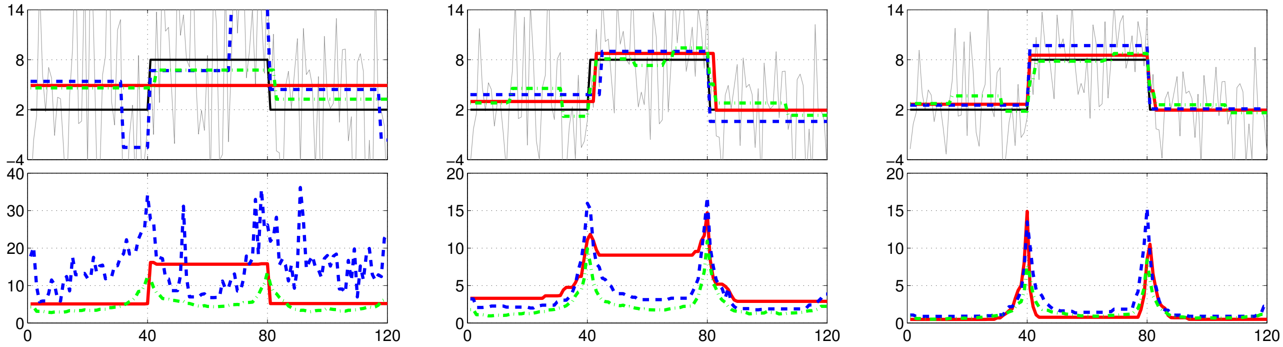

Fig. 2: **Estimation for single realizations (Top) and MSE over 100 realizations (Bottom) for different SNR.** From left to right SNR = -1.5dB, 0.4dB and 1.8dB. True signal $\mathbf{x}$ (black solid), observations $\mathbf{y}$ (gray solid), $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$ (red solid), $\widehat{\mathbf{x}}_{\text{FB}}$ (blue dashed) and $\mathrm{x}^*_{\lambda_{\text{SURE}}}$ (green mixed).

SNR (dB)	-1.5			0.4			1.8		
	GSURE	FB	HBV	GSURE	FB	HBV	GSURE	FB	HBV
MAD	**1.8**	2.5	2.7	**1.2**	1.4	2.0	0.9	0.9	**0.7**
MSE	**5.6**	13.5	8.7	**2.6**	3.8	5.6	1.5	2.0	**1.3**
time (s)	95.2	8.1	**0.045**	94.5	7.5	**0.048**	94.0	7.5	**0.050**

Table 1: **Estimation performance vs. SNR (first example).**

cal distribution of $\lambda \in \Lambda_{\text{opt}}$ is reported in Fig. 1 (bottom). Its mean (respectively median) equals 13.2 (respectively 13.4), which is consistent with the position of the maximum of the ensemble average of the log posterior distribution at $\lambda \simeq 13$.

3.3. Estimation performance

Comparisons with state-of-the-art methods – We proceed with comparing against a *fully Bayesian* (FB) procedure in which one may naturally compute the Bayesian estimators associated with the posterior distribution $f(\boldsymbol{\Theta}|\mathbf{y})$ in (9). Deriving the closed-form expression of the *maximum a posteriori* (MAP) or *minimum mean squared error* (MMSE) estimators associated with $f(\boldsymbol{\Theta}|\mathbf{y})$ is not straightforward. Alternatively, these estimators can be approximated by using MCMC procedures that essentially rely on a partially collapsed Gibbs sampler [15] similar to the algorithm derived in [14]. It consists in iteratively drawing samples (denoted $\cdot^{(t)}$) according to conditional posterior distributions that are associated with the joint posterior (9). The resulting procedure, detailed in Algorithm 2, provides a set of samples $\boldsymbol{\vartheta} = \left\{\mathbf{r}^{(t)}, \boldsymbol{\mu}^{(t)}, \sigma^{2(t)}, p^{(t)}\right\}_{t=1}^{T}$ that are asymptotically distributed according to (9). These samples can be used to approximate the MMSE estimators of the parameters of interest by empirical averaging. The corresponding optimal solution is referred to as $\widehat{\mathbf{x}}_{\text{FB}}$.

In addition, $\widehat{\mathbf{x}}_{\lambda_{\text{opt}}}$ is also compared against $\mathbf{x}^*_\lambda$ for $\lambda = \lambda_{\text{SURE}}$ that minimizes the *Stein unbiased risk estimate* (SURE). It is used as a benchmark solution since it requires the a priori knowledge of the noise variance σ^2. To do so, we have replaced steps 3 and 4 in Algorithm 1 with the computation of the Generalized SURE estimator proposed in [11], which requires to deal with an expectation that we approximate by an empirical mean computed over 10^5 realizations. Moreover, the output $\mathbf{x}^*_\lambda$, which minimizes the risk between $\mathbf{x}^*_\lambda$ and $\mathbf{x}$, is obtained with $\lambda = \lambda_{\text{SURE}} \in \text{Argmin}_{\lambda\in\Lambda}\text{GSURE}(\mathbf{y}, \mathbf{x}^*_\lambda, \sigma^2)$.

Algorithm 2 FB Algorithm

```
Input: Observed signal $\mathbf{y} \in \mathbb{R}^N$.
       Prior parameters $\boldsymbol{\Phi} = \{\alpha_0, \alpha_1, \mu_0, \sigma_0^2\}$.
Iterations:
for $t = 1, \dots, T$ do
    for $i = 1, \dots, N-1$ do
        Draw $r_i^{(t)} \sim \mathrm{P}\left[r_i|\mathbf{y}, \mathbf{r}_{\backslash i}, p, \sigma^2, \mu_0, \sigma_0^2\right]$
    end for
    for $k = 1, \dots, \sum_{i=1}^N r_i^{(t)}$ do
        Draw $\mu_k^{(t)} \sim f\left(\mu_k|\mathbf{y}, \mathbf{r}, \sigma^2, \mu_0, \sigma_0^2\right)$
    end for
    Draw $\sigma^{2(t)} \sim f\left(\sigma^2|\mathbf{y}, \mathbf{r}, \boldsymbol{\mu}\right)$
    Draw $p^{(t)} \sim f\left(p|\mathbf{r}, \alpha_0, \alpha_1\right)$
end for
Output: $\boldsymbol{\vartheta} = \left\{\mathbf{r}^{(t)}, \boldsymbol{\mu}^{(t)}, \sigma^{2(t)}, p^{(t)}\right\}_{t=1}^{T}$;
```

The results for a single realization are plotted in Fig. 2 (top) for different SNR values. Corresponding average performance over 100 realizations are reported in Table 1: Mean absolute deviations $\text{MAD}(\widehat{\mathbf{x}}) = \widehat{\mathbb{E}}[\frac{1}{N}\|\widehat{\mathbf{x}} - \mathbf{x}\|_1]$ (where $\widehat{\mathbb{E}}$ stands for the empirical mean estimator), mean squared error $\text{MSE}(\widehat{\mathbf{x}}) = \widehat{\mathbb{E}}[\frac{1}{N}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2]$, and average execution times. Fig. 2 (bottom) provides a performance analysis in terms of MSE at each location $i \in \{1, \dots, N\}$.

Overall, the proposed HBV method yields estimates that

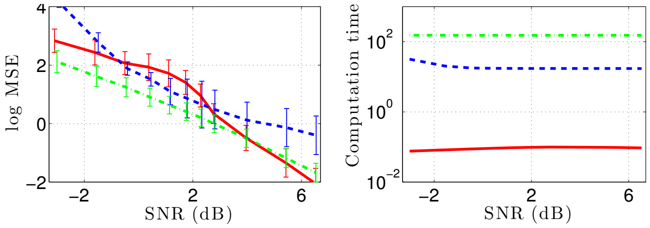

Fig. 3: **Estimation performance vs. SNR (second example).** MSE (left) and average computation time (right) associated to $\widehat{\mathbf{x}}_{\lambda_{\mathrm{opt}}}$ (red solid), $\widehat{\mathrm{x}}_{\mathrm{FB}}$ (blue dashed) and $\mathrm{x}^*_{\lambda_{\mathrm{SURE}}}$ (green mixed).

are comparable with FB results in terms of MSE and MAD, while its computation time is reduced by a factor of more than 150 with respect to the FB approach. For high SNR values, HBV clearly outperforms FB, the latter having wider MSE peaks at the change point locations, indicating less precision. Performance are consistent with those obtained with the second example and reported in Fig. 3 (left: MSE, right: average computation time) for different SNR values. In addition, the computational time associated with HBV is reduced by a larger factor as the sample size N grows.

The comparisons with GSURE illustrate that, at high SNR, the estimation performance of the proposed approach are close to those of the GSURE oracle (which requires the a priori knowledge of the noise variance) while the computation cost for the proposed approach is 3 orders of magnitude smaller. Note that $\widehat{\mathbf{x}}_{\lambda_{\mathrm{opt}}}$ is estimated a posteriori, cf., (13), while $\mathbf{x}^*_{\lambda_{\mathrm{SURE}}}$ is a direct solution of (1), resulting in a slightly lower MSE for high SNR values.

3.4. Behavior comparisons

In addition, Fig. 2 shows that when increasing noise variance HBV tends to detect less change points and eventually no change point, while FB and GSURE yield a larger number of change point detection, with highly variable locations. Visually, no clear preference can be given to any of the methods, yet HBV has a lower MSE than FB at very low SNR. These pronouncedly different behaviors at low SNR are further illustrated in Fig. 4 where segment size empirical distributions are plotted as functions of SNR.

4. CONCLUSIONS

We developed a hybrid Bayesian variational method for the change point detection problem. The originality and advantage of the proposed procedure reside in combining the computational efficiency of variational methods with the statistical flexibility of a hierarchical Bayesian model, that thus permit to handle efficiently the automated regularization parameter selection. The proposed procedure compares favorably against a fully Bayesian approach in terms of estimation performances, while reducing computational cost by more than 2 orders of magnitude, and achieves, at large SNR, performance

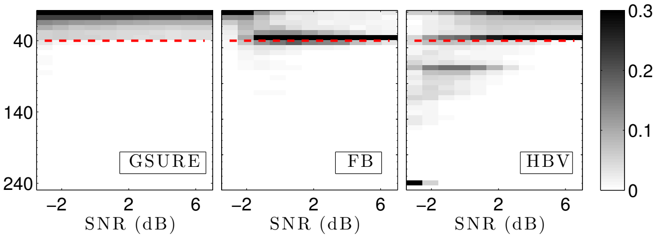

Fig. 4: **Size empirical distributions of estimated segment vs. SNR.** Second example where $\mathbf{x}$ consists of six segments which are 40 samples long (dashed red).

that are similar to those obtained with the GSURE "oracle" procedure, which requires the a priori knowledge of the noise variance. Extensions to image denoising and segmentation are under current investigation.

REFERENCES

[1] M. Basseville and I.V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[2] M.A. Little and N.S. Jones, "Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory," *Proc. R. Soc. A*, vol. 467, pp. 3088–3114, 2011.

[3] K. Bleakley and J.P. Vert, "The group fused lasso for multiple change-point detection," Tech. Rep., 2011.

[4] C.R. Vogel and M.E. Oman, "Iterative methods for total variation denoising," *J. Sci. Comput.*, vol. 17, pp. 227–238, 1996.

[5] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1-2, pp. 89–97, Jan. 2004.

[6] A. Barbero and S. Sra, "Fast Newton-type methods for total variation regularization," in *International Conference on Machine Learning*, Lise Getoor and Tobias Scheffer, Eds., New York, NY, USA, Jun. 2011, ICML '11, pp. 313–320, ACM.

[7] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[8] L. Condat, "A direct algorithm for 1D total variation denoising," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1054–1057, 2013.

[9] P.L. Davies and A. Kovac, "Local extremes, runs, strings and multiresolution," *Ann. Stat.*, vol. 29, no. 1, pp. 1–65, 2001.

[10] C.M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, 1981.

[11] C. Deledalle, S. Vaiter, G. Peyré, J.M. Fadili, and C. Dossal, "Unbiased risk estimation for sparse analysis regularization," in *Proc. Int. Conf. Image Process.*, Orlando, FL, USA, Sept. 30-Oct. 3 2012, pp. 3053 – 3056.

[12] M. Lavielle and E. Lebarbier, "An application of MCMC methods for the multiple change-points problem," *Signal Processing*, vol. 81, no. 1, pp. 39–53, Jan. 2004.

[13] E. Punskaya, C. Andrieu, A. Doucet, and W.J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Process.*, vol. 50, pp. 747–758, 2002.

[14] N. Dobigeon, J.-Y. Tourneret, and J. D. Scargle, "Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model," *IEEE Trans. Signal Process.*, vol. 55, no. 2, pp. 414–423, Feb. 2007.

[15] D.A. van Dyk and T. Park, "Partially collapsed Gibbs samplers: Theory and methods," *J. Am. Stat. Assoc.*, vol. 103, no. 482, pp. 790–796, June 2008.