BAYESIAN-DRIVEN CRITERION TO AUTOMATICALLY SELECT THE REGULARIZATION PARAMETER IN THE ℓ_1-POTTS MODEL

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

[1] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France firstname.lastname@ens-lyon.fr
[2] IRIT, CNRS, INP-ENSEEIHT, Toulouse, France firstname.lastname@irit.fr

ABSTRACT

This contribution focuses, within the ℓ_1-Potts model, on the automated estimation of the regularization parameter balancing the ℓ_1 data fidelity term and the TVℓ_0 penalization. Variational approaches based on total variation gained considerable interest to solve piecewise constant denoising problems thanks to their deterministic setting and low computational cost. However, the quality of the achieved solution strongly depends on the tuning of the regularization parameter. While recent works have tailored various hierarchical Bayesian procedures to additionally estimate the regularization parameter for Gaussian noise, less attention has been granted to Laplacian noise, of interested in numerous applications. This contribution promotes a fast and parameter-free denoising procedure for piecewise constant signals corrupted by Laplacian noise, that includes automated selection of the regularization parameter. It relies on the minimization of a Bayesian-driven criterion whose similarities with the ℓ_1-Potts model permit to derive a computationally efficient algorithm.

***Index Terms*—** Piecewise constant denoising, Laplacian noise, regularization parameter automated selection, Potts model, hierarchical Bayesian model.

1. INTRODUCTION

Denoising piecewise constant signals is of considerable interest in various applications [1, 2]. A large part of the literature relies on the assumption that noise is white and Gaussian. However, in numerous applications, it is advisable to assume a Laplace distribution for the noise rather than a Gaussian, in particular to account for large excursions of the noise, potentially corresponding to outliers [3, 4].

The problem of denoising piecewise constant signals corrupted by Laplace noise can be traced back to [5, 6]. Since then, it has rarely been envisaged in the Bayesian literature because, as opposed to Gaussian noise, posterior distributions cannot be, in general, represented in terms of simple and tractable functions of the observations. However, various methods have been developed to generalize the likelihood ratio test approach (see, e.g., [4] and references therein). In the variational setting, there exists a formulation based on the following ℓ_1-TVℓ_0 problem (usually referred to as ℓ_1-Potts model; here ℓ_1 indicates the norm of the data fidelity term) [7, 8]

$$\widehat{\boldsymbol{x}}_\lambda \in \underset{\boldsymbol{x}\in\mathbb{R}^N}{\operatorname{Argmin}} \|\boldsymbol{y}-\boldsymbol{x}\|_1 + \lambda \|L\boldsymbol{x}\|_0, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ denotes the linear superposition of data and noise, $L \in \mathbb{R}^{(N-1)\times N}$ is the first difference operator, i.e., $(L\boldsymbol{x})_k = x_{k+1} - x_k$, and $\lambda \geq 0$ is a regularization parameter aiming to balance the data fidelity and regularization terms. While (1) can be solved by means of, e.g., dynamic programming [7, 8], the estimation performance of $\widehat{\boldsymbol{x}}_\lambda$ strongly depends on the choice of λ, which is *a priori* unknown.

Related works. In the context of Gaussian noise and, thus, of the ℓ_2-TVℓ_q counterpart of (1) with $q = \{0, 1\}$, interesting ideas to select λ rely on a hierarchical Bayesian structure between $\boldsymbol{y}$, $\boldsymbol{x}$ and λ. For instance the selection of λ in the ℓ_2-TVℓ_1 problem can be solved by assuming a certain prior for λ and maximizing either the conditional distribution of $(\boldsymbol{x}, \lambda)$ given $\boldsymbol{y}$ [9, 10] or a marginalised posterior in order to remove λ from the model [10, 11]. In the context of ℓ_2-TVℓ_0 (also known as the ℓ_2-Potts model), a taylored procedure relying on a reparametrization of $\boldsymbol{x}$ has been derived in [12] and its performance have been assessed and validated numerically. All these approaches share the common idea that the problem amounts in considering a joint estimation of $\boldsymbol{x}$ and λ, thus requiring an additional penalization term, precluding the trivial choice $\lambda = 0$, and often designed from hierarchical Bayesian arguments.

Contributions and outline. Departing from [12], which is taylored to Gaussian noise (i.e., the ℓ_2-Potts model), this paper derives a parameter-free estimation procedure suited to Laplacian piecewise contant denoising. It consists in solving

$$(\widehat{\boldsymbol{x}}, \widehat{\sigma}, \widehat{\lambda}) \in \underset{\boldsymbol{x}\in\mathbb{R}^N, \sigma\geq 0, \lambda \geq 0}{\operatorname{Argmin}} \frac{1}{\sigma}\|\boldsymbol{y}-\boldsymbol{x}\|_1 + \frac{\lambda}{\sigma}\|L\boldsymbol{x}\|_0 + \phi(\sigma, \lambda) \tag{2}$$

after having carefully designed ϕ beforehand. To that end, this work relies on the following original key ingredients. First, a suitable reparametrization of $\boldsymbol{x}$ is proposed in Section 2. Then, the penalty ϕ is derived based on hierarchical Bayesian arguments described in Section 3. Moreover, we design in Section 4 an algorithmic procedure providing an approximate solution for (2) which benefits from numerically efficient dynamic programming techniques solving (1). Estimation performance are assessed in Section 5 and demonstrate that the proposed procedure provides estimates of the regularization parameter of the ℓ_1-TVℓ_0 problem that lead to performance close to that of the oracle estimator, at competitive computational cost. Section 6 concludes on this contribution.

2. PROBLEM FORMULATION

2.1. Problem statement

The underlying problem consists in estimating a piecewise constant signal $\overline{\boldsymbol{x}} \in \mathbb{R}^N$ from the noisy observations $\boldsymbol{y} = \overline{\boldsymbol{x}} + \boldsymbol{\epsilon}$. The noise samples $(\boldsymbol{\epsilon}_i)_{1\leq i\leq N}$ are assumed to be independent and identically distributed (i.i.d.) zero mean Laplace variables with common but unknown scale parameter σ, i.e., $\boldsymbol{\epsilon}|\sigma \sim \text{Laplace}\,(\boldsymbol{0}, \sigma \boldsymbol{I}_N)$.

2.2. Problem parametrization

Following [13, 14], any candidate solution $\boldsymbol{x}$ can be parametrized by the indicator vector $\boldsymbol{r} \in \{0,1\}^N$ of its change-points and the vector of values, $\boldsymbol{\mu} = (\mu_k)_{1\leq k\leq K}$, taken between each change-point. The indicator vector $\boldsymbol{r} = (r_i)_{1\leq i\leq N}$ is introduced as follows

$$r_i = \begin{cases} 1, & \text{if there is a change-point at time instant } i, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

By convention, $r_i = 1$ indicates that x_i is the last sample belonging to the current segment, and thus that x_{i+1} belongs to the next segment. Moreover, setting $r_N = 1$ ensures that $K = \sum_{i=1}^{N} r_i$.

For each $k \in \{1,\dots,K\}$, the set $\mathcal{R}_k \subset \{1,\dots,N\}$ represents the set of time indices associated to the k-th segment. Therefore, $\mathcal{R}_k \cap \mathcal{R}_{k'} = \{\emptyset\}$ for $k \neq k'$ and $\cup_{k=1}^{K}\mathcal{R}_k = \{1,\cdots,N\}$. Hereafter, the notation $K_{\boldsymbol{r}}$ will be adopted to emphasize the dependence of the number K of segments on the indicator vector $\boldsymbol{r}$, i.e., $K = \|\boldsymbol{r}\|_0$.

The values taken on each segment of $\boldsymbol{x}$ can be encoded by introducing the vector $\boldsymbol{\mu} = (\mu_k)_{1\leq k\leq K_{\boldsymbol{r}}}$ such that

$$(\forall k \in \{1,\dots,K_{\boldsymbol{r}}\})(\forall i \in \mathcal{R}_k) \quad x_i = \mu_k. \tag{4}$$

This parametrization leads to a first result whose derivation comes from straightforward yet tedious calculations not reported here for brevity, see also [12].

Proposition 2.1. *Let $\boldsymbol{y} \in \mathbb{R}^N$ and $\phi\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$. Problem* (2) *is equivalent to*

$$\underset{\substack{\{\boldsymbol{r},\boldsymbol{\mu}\}\in\{0,1\}^N\times\mathbb{R}^{K_{\boldsymbol{r}}} \\ \lambda\geq 0,\sigma\geq 0}}{\text{minimize}} \left\{ \frac{1}{\sigma}\sum_{k=1}^{K_{\boldsymbol{r}}}\sum_{i\in\mathcal{R}_k}|y_i-\mu_k| + \frac{\lambda}{\sigma}(K_{\boldsymbol{r}}-1) + \phi(\sigma,\lambda)\right\}. \tag{5}$$

3. BAYESIAN DRIVEN DERIVATION OF ϕ

It is well established that a minimization problem balancing data versus regularization terms can be related to the maximization of a posterior distribution (see e.g., [15]). A similar idea is used here, for a posterior distribution that is obtained with a hierarchical Bayesian formulation in order to derive a connection with (5).

3.1. Hierarchical Bayesian model

First, the joint likelihood function for $\boldsymbol{y}$ given the piecewise constant model $\{\boldsymbol{r},\boldsymbol{\mu}\}$, the noise model and the scale parameter σ follows a Laplace distribution, i.e.,

$$f_{\mathcal{L}}\left(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\mu},\sigma\right) = \prod_{k=1}^{K_{\boldsymbol{r}}}\prod_{i\in\mathcal{R}_k}\frac{1}{\sigma}\exp\Big(-\frac{|\mu_k-y_i|}{\sigma}\Big). \tag{6}$$

To derive the posterior distribution, prior distributions over $\boldsymbol{r}$, $\boldsymbol{\mu}$, and σ have to be specified. In the literature, it is often encountered to assume that $(r_i)_{1\leq i\leq N}$ are independent and identically distributed (i.i.d.) according to a Bernoulli distribution with hyper-parameter p [16–18]. The parameter p quantifies the prior probability of occurrence of a change, independently of the location:

$$f_{\mathcal{B}}(\boldsymbol{r}|p) = \prod_{i=1}^{N-1} p^{r_i}(1-p)^{1-r_i}. \tag{7}$$

From a hierarchical Bayesian perspective, a natural choice for the posterior distribution of segment amplitudes $(\mu_k)_{1\leq k\leq K_{\boldsymbol{r}}}$ would consist in electing independent conjugate Laplace prior distributions. However, such posterior distributions cannot be made explicit as simple tractable functions. To alleviate this problem, we choose a non-informative prior such as a uniform distribution on a fixed interval $[\mu^-,\,\mu^+]$, i.e.,

$$f_{\mathcal{U}}(\boldsymbol{\mu}|\boldsymbol{r}) = \prod_{k=1}^{K_{\boldsymbol{r}}}(\mu^+-\mu^-)^{-1}\imath_{[\mu^-,\mu^+]}(\mu_k), \tag{8}$$

where $\imath_A$ is the indicator function of the set A. In particular, if we choose $\mu^- < \min(\boldsymbol{y})$ and $\mu^+ > \max(\boldsymbol{y})$, then (8) recasts into

$$f_{\mathcal{U}}(\boldsymbol{\mu}|\boldsymbol{r}) = \left(\mu^+-\mu^-\right)^{-K_{\boldsymbol{r}}}\imath_{[\mu^-,\mu^+]^{K_{\boldsymbol{r}}}(\boldsymbol{\mu})}. \tag{9}$$

Following usual prior choices in hierarchical approaches [14], a scale-invariant non-informative Jeffreys prior is assigned to the scale parameter σ in order to account for the absence of prior knowledge on σ, i.e.,

$$f_{\mathcal{J}}\left(\sigma\right) \propto \frac{1}{\sigma}, \tag{10}$$

while a conjugate Beta distribution with fixed parameters α_0 and α_1 is assigned to the unknown hyper-parameter p

$$f_{\beta}(p) = \frac{\Gamma\left(\alpha_0+\alpha_1\right)}{\Gamma\left(\alpha_0\right)\Gamma\left(\alpha_1\right)}p^{\alpha_1-1}(1-p)^{\alpha_0-1}. \tag{11}$$

Assuming that the parameters $\boldsymbol{r}$, $\boldsymbol{\mu}$ and σ are a priori independent, the joint posterior distribution can be derived as

$$f(\boldsymbol{\Theta}|\boldsymbol{y}) \propto f_{\mathcal{L}}\left(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\mu},\sigma\right) f_{\mathcal{U}}\left(\boldsymbol{\mu}|\boldsymbol{r}\right) f_{\mathcal{B}}(\boldsymbol{r}|p) f_{\mathcal{J}}\left(\sigma\right) f_{\beta}(p) \tag{12}$$

with $\boldsymbol{\Theta} = \{\boldsymbol{r},\boldsymbol{\mu},\sigma,p\} \in \mathcal{Q} = \{0,1\}^N\times\mathbb{R}^{K_{\boldsymbol{r}}}\times\mathbb{R}_+\times[0,1]$. Finally, the maximum a posteriori (MAP) estimator can be computed by minimizing the negative log-posterior distribution (12), which leads to the problem

$$\begin{aligned}\underset{\boldsymbol{\Theta}=\{\boldsymbol{r},\boldsymbol{\mu},\sigma^2,p\}\in\mathcal{Q}}{\text{minimize}}\ & \frac{1}{\sigma}\sum_{k=1}^{K_{\boldsymbol{r}}}\sum_{i\in\mathcal{R}_k}|y_i-\mu_k| \\ & + (K_{\boldsymbol{r}}-1)\left(\log\left(\frac{1-p}{p}\right)+\log(\mu^+-\mu^-)\right) \\ & + N\log(2\sigma) - (N-1)\log(1-p) + \log\sigma \\ & - (\alpha_1-1)\log p - (\alpha_0-1)\log(1-p) \\ & + \log(\mu^+-\mu^-).\end{aligned} \tag{13}$$

3.2. Problem equivalence and derivation of ϕ

The core idea to derive ϕ consists in drawing an equivalence between problems (5) and (13).

Proposition 3.1. *The minimization problems* (5) *and* (13) *lead to a similar solution with the following parametrization of λ*

$$\frac{\lambda}{\sigma} = \left(\log\left(\frac{1-p}{p}\right)+\log(\mu^+-\mu^-)\right) \tag{14}$$

and choice of ϕ

$$\begin{aligned}\phi(\sigma,\lambda) &= N\log(2\sigma)+\log(\sigma) \\ &- \frac{\lambda}{\sigma}(N+\alpha_0-2) + (N+\alpha_0-1)\log(\mu^+-\mu^-) \\ &+ (N+\alpha_0+\alpha_1-3)\log\left(1+\exp\left(\frac{\lambda}{\sigma}-\log(\mu^+-\mu^-)\right)\right).\end{aligned} \tag{15}$$

Algorithm 1 Bayesian driven resolution of the ℓ_1-TVℓ_0 problem

Input: Observed signal $\boldsymbol{y} \in \mathbb{R}^N$.
The predefined set of regularization parameters Λ.
Hyperparameters $\Phi = \{\alpha_0, \alpha_1, \mu^+ - \mu^-\}$.

Iterations:

1: **for** $\lambda \in \Lambda$ **do**
2: Compute $\widehat{\boldsymbol{x}}_\lambda = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{y}-\boldsymbol{x}\|_1 + \lambda\|L\boldsymbol{x}\|_0$.
3: Compute $\widehat{\sigma}_\lambda = \|\boldsymbol{y}-\widehat{\boldsymbol{x}}_\lambda\|_1/N$.
4: **end for**

Output: Solution $\{\widehat{\boldsymbol{x}}_{\widehat{\lambda}}, \widehat{\lambda}, \widehat{\sigma}_{\widehat{\lambda}}\}$ with $\widehat{\lambda} = \arg\min_{\lambda\in\Lambda} F(\widehat{\boldsymbol{x}}_\lambda, \lambda, \widehat{\sigma}_\lambda)$

The principle of the proof consists in observing that both (5) and (13) contain the same data fidelity term, a term proportional to $(K_{\boldsymbol{r}} - 1)$ which leads to (14) by identification, and an additional term independent of $\boldsymbol{r}$, namely

$$\begin{aligned}\phi(\sigma, p) = {} & N\log(2\sigma) - (N-1)\log(1-p) + \log\sigma \\ & - (\alpha_1 - 1)\log p - (\alpha_0 - 1)\log(1-p) \\ & + \log(\mu^+ - \mu^-).\end{aligned} \tag{16}$$

By inverting (14), ϕ can be parametrized in terms of (σ, λ), as in (15).

4. ALGORITHMIC SOLUTION

Now that an explicit expression (15) has been derived for ϕ, we aim to solve the nonconvex optimization problem (5) in a deterministic setting. We propose to use a predefined and fixed discrete grid Λ for λ and to solve, $\forall\lambda \in \Lambda$

$$(\widehat{\boldsymbol{x}}_\lambda, \widehat{\sigma}_\lambda) \in \operatorname*{Argmin}_{\boldsymbol{x}\in\mathbb{R}^N, \sigma\geq 0} \underbrace{\frac{1}{\sigma}\|\boldsymbol{y}-\boldsymbol{x}\|_1 + \frac{\lambda}{\sigma}\|L\boldsymbol{x}\|_0 + \phi(\sigma,\lambda)}_{F(\boldsymbol{x},\lambda,\sigma)}, \tag{17}$$

which we estimate as

$$(\forall\lambda\in\Lambda) \quad \begin{cases} \widehat{\boldsymbol{x}}_\lambda & = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{y}-\boldsymbol{x}\|_1 + \lambda\|L\boldsymbol{x}\|_0, \\ \widehat{\sigma}_\lambda & = \|\boldsymbol{y}-\widehat{\boldsymbol{x}}_\lambda\|_1/N. \end{cases} \tag{18}$$

The solution triplet $\{\widehat{\boldsymbol{x}}_{\widehat{\lambda}}, \widehat{\lambda}, \widehat{\sigma}_{\widehat{\lambda}}\}$ is then selected such that

$$\widehat{\lambda} = \arg\min_{\lambda\in\Lambda} F(\widehat{\boldsymbol{x}}_\lambda, \lambda, \widehat{\sigma}_\lambda). \tag{19}$$

The corresponding algorithm steps are reported in Algorithm 1. This approach permits to use the dynamic programming algorithm *Pottslab* developed in [7,8] to solve the problem (1) for any $\lambda \in \Lambda$.

5. ESTIMATION PERFORMANCE

5.1. Experimental setting

Data $\boldsymbol{y}$ are synthesized in two steps. First, the change-point locations of $\overline{\boldsymbol{x}}$ are drawn i.i.d. with the given change-point probability $\overline{p}$, and then the value taken on each segment is uniformly drawn between a minimal value $\overline{x}_{\min}$ and a maximal value $\overline{x}_{\max}$ also given beforehand. Second, we generate $\boldsymbol{\epsilon} \sim \text{Laplace}\,(\boldsymbol{0}, \sigma\boldsymbol{I}_N)$ and form $\boldsymbol{y} = \overline{\boldsymbol{x}} + \boldsymbol{\epsilon}$. Therefore the mean amplitude between successive segments is about $(\overline{x}_{\max} - \overline{x}_{\min})/3$ and its comparison w.r.t. the noise standard deviation $\sqrt{2}\overline{\sigma}$ is conducted thanks to the so-called amplitude-to-noise-ratio defined as $\text{ANR} = (\overline{x}_{\max} - \overline{x}_{\min})/(3\sqrt{2}\overline{\sigma})$.

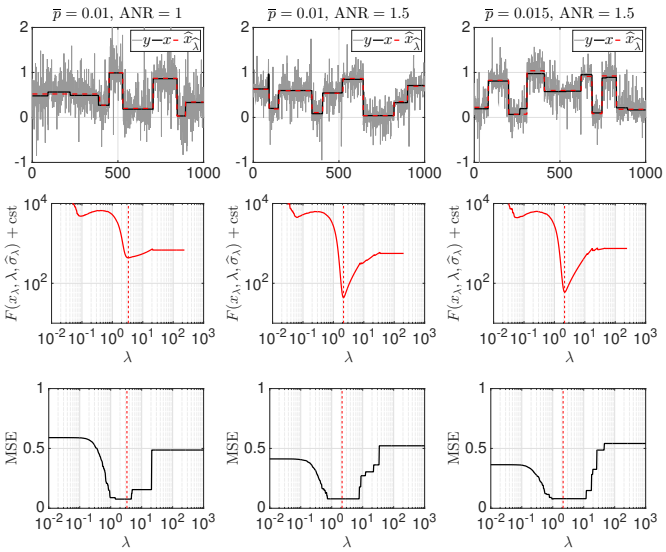

Fig. 1. **Illustration of the proposed method.** Three configurations are examined depending on $\overline{p}$ and ANR. The proposed criterion $F(\widehat{\boldsymbol{x}}_\lambda, \lambda, \widehat{\sigma}_\lambda)$ is displayed in red in the second row as a function of λ while the relative MSE between $\widehat{\boldsymbol{x}}_\lambda$ and $\overline{\boldsymbol{x}}$ is displayed in the third row. The estimate $\widehat{\lambda}$ is indicated with a vertical dashed line and the corresponding solution $\widehat{\boldsymbol{x}}_{\widehat{\lambda}}$ is reported in red in the first row.

The hyperparameters are chosen as follows. A non-informative prior is used for the prior probability p by setting $\alpha_0 = \alpha_1 = 1$ so that the Beta distribution in (11) reduces to a uniform distribution over $(0, 1)$. In addition, μ^+ and μ^- are chosen such that $\mu^- < \min \boldsymbol{y}$, $\mu^+ > \max \boldsymbol{y}$ and $\mu^+ - \mu^- = 10^4$. This choice is further discussed in Section 5.3.

In our experiments, the set Λ of discrete values for λ has been composed of 500 values equally spaced, in a $\log_{10}$-scale, between 10^{-5} and 10^5.

5.2. Illustration of the automatic selection of λ

Three different observations $\boldsymbol{y}$ (grey) are represented in each column of Fig. 1 as representatives of different configurations of change-point probability $\overline{p}$ and scale parameter $\overline{\sigma}$. The true $\overline{\boldsymbol{x}}$ that we aim to recover is displayed in black.

The proposed criterion $F(\widehat{x}_\lambda, \lambda, \widehat{\sigma}_\lambda)$ is displayed in the second row in solid red as function of λ. The position of its minimum $\widehat{\lambda}$ (see (19)) is indicated by a vertical dashed red line. The corresponding solution $\widehat{\boldsymbol{x}}_{\widehat{\lambda}}$ is reported in dashed red lines in the first row. It matches exactly or satisfactorily the location of minimum MSE (between $\widehat{\boldsymbol{x}}_\lambda$ and $\overline{\boldsymbol{x}}$) and appears visually as a good estimate of $\overline{\boldsymbol{x}}$.

Performance are further quantified in term of relative mean squared error (MSE) between $\widehat{\boldsymbol{x}}_\lambda$ and $\overline{\boldsymbol{x}}$ as a function of λ in the third row of Fig. 1. For $\overline{p} = 0.01$ and $\text{ANR} = 1.5$ (center column), the method automatically selects one $\widehat{\lambda}$ such as the solution benefits from a lower MSE than for any other $\lambda \in \Lambda$. However, when the ANR decreases (left column) or $\overline{p}$ increases (right column), the solution may not be optimal in terms of MSE but still maintains very good estimation performance close to the minimum MSE.

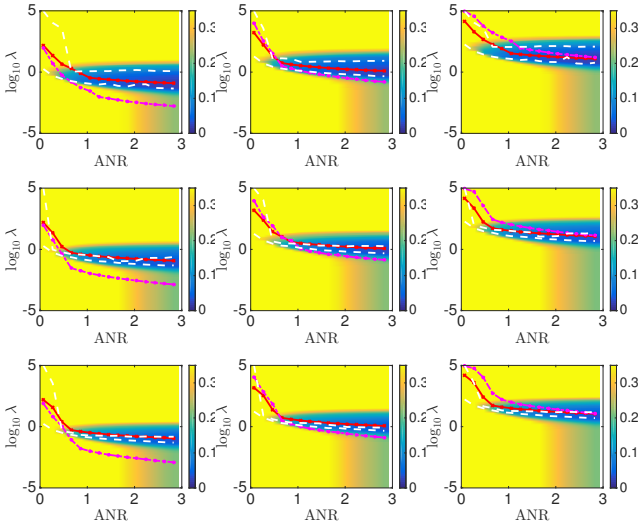

Fig. 2. **Estimation performance ($\widehat{\lambda}$ vs. ANR).** From top to bottom: $\overline{p} = 0.005, 0.010$ and 0.015. Different dynamics are examined from left to right: $\overline{x}_{\max} - \overline{x}_{\min} = 0.1$, 1, and 10. The average proposed estimate $\widehat{\lambda}$ is displayed in red as a function of the ANR and is compared w.r.t $\widehat{\lambda}^{(\ell_2)}$ (mixed magenta) and the MSE oracle estimator Λ_{MSE} whose range is delimited by dashed white lines. Overall, the proposed solution (red line) remains between the dashed white lines, thus showing that $\widehat{\boldsymbol{x}}_\lambda$ for $\lambda = \widehat{\lambda}$ achieves the best performance in terms of relative MSE than for any other λ.

5.3. Quantification of estimation performance

In this section, we further examine the estimation performance with respect to (w.r.t.) the dynamic $\overline{x}_{\max} - \overline{x}_{\min} \in \{0.1, 1, 10\}$, the change-point probability $\overline{p} \in \{0.005, 0.01, 0.015\}$ and the ANR. In Fig. 2, the regularization parameter selected by the proposed method ($\widehat{\lambda}$, red) is compared against the similar method developed in [12] for Gaussian noise ($\widehat{\lambda}^{(\ell_2)}$, magenta). In addition, we also report the oracle estimate (Λ_{MSE}, dashed white) for which $\widehat{\boldsymbol{x}}_{\lambda \in \Lambda_{\text{MSE}}}$ yields the lowest MSE and whose range is delimited by two dashed white lines. Results presented here are averaged over 25 realizations.

Behavior of $\widehat{\lambda}$. Each plot in Fig. 2 illustrates how $\widehat{\lambda}$, $\widehat{\lambda}^{(\ell_2)}$ and Λ_{MSE} vary w.r.t. the ANR. For comparison purposes, the relative MSE is also superimposed. Overall, the proposed solution (red line) remains between the dashed white lines, thus showing that $\widehat{\boldsymbol{x}}_\lambda$ for $\lambda = \widehat{\lambda}$ achieves the best performance in terms of relative MSE than for any other λ. However, we observe that the performance deteriorate as $\overline{p}$ increases (see Fig. 2 from top to bottom) as the estimation problem is more difficult when a larger number of segments needs to be detected. This is quantified by the shrinking of the oracle range Λ_{MSE}.Further investigations show that $\widehat{\lambda}$ scales properly when the dynamic varies (see translations of red and white lines in Fig. 2 from left to right): For two observations that are identical up to a scale factor, the proposed method outputs identical solutions up to that same scale factor. This does not hold for $\widehat{\lambda}^{(\ell_2)}$.

Comparison of MSE. In addition, estimation performance are compared in terms of relative MSE between $\widehat{\boldsymbol{x}}_\lambda$ and $\overline{\boldsymbol{x}}$ for $\lambda = \widehat{\lambda}$ (red), $\widehat{\lambda}^{(\ell_2)}$ (magenta) and any $\lambda \in \Lambda_{\text{MSE}}$ (black) in Fig. 3. Results illustrate that taking into account the Laplacian nature of the noise (red) rather than assuming it Gaussian (magenta) achieves systematically lower relative MSE. In addition, both methods exhibit equivalent performance for sufficiently large ANR. Overall, the proposed method provides MSE performance close to the oracle estimate as $\overline{p}$ is small and ANR is large.

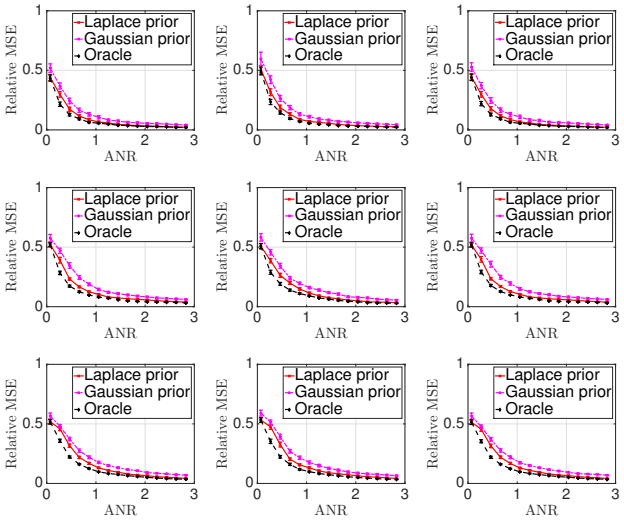

Fig. 3. **Estimation performance (MSE vs. ANR).** From top to bottom: $\overline{p} = 0.005, 0.010$ and 0.015. Different dynamics are examined from left to right: $\overline{x}_{\max} - \overline{x}_{\min} = 0.1$, 1, and 10. For each configuration, relative MSE $\|\widehat{\boldsymbol{x}}_\lambda - \overline{\boldsymbol{x}}\|/\|\overline{\boldsymbol{x}}\|$ are compared for $\lambda = \widehat{\lambda}$ (red), $\widehat{\lambda}^{(\ell_2)}$ (magenta) and any $\lambda \in \Lambda_{\text{MSE}}$ (black). Results show that including the knowledge that the noise is Laplacian permits to yield lower MSE.

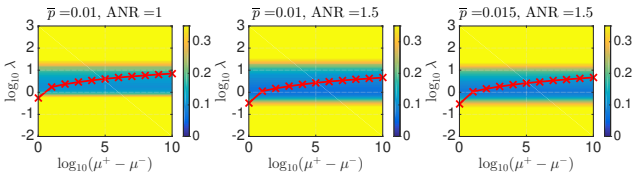

Fig. 4. **Impact of hyperparameters (MSE vs. $\mu^+ - \mu^-$).** For each configuration, $\widehat{\lambda}$ (red) is plotted as a function of the hyperameter value $\mu^+ - \mu^-$. Results show satisfactory and similar performances for any choice such as $10 \leq \mu^+ - \mu^- \leq 10^6$.

Impact of hyperparameter $\mu^+ - \mu^-$. For the same configurations as those presented in Fig. 1, the results displayed in Fig. 4 show that the tuning of $\mu^+ - \mu^-$ does not require a complicated procedure as the estimation performance are satisfactory and very similar for $10 \leq \mu^+ - \mu^- \leq 10^6$.

Computational cost. In the experiments presented here, simulations took around 40 seconds for $|\Lambda| = 500$ and $N = 10^3$.

6. CONCLUSION

Elaborating on previous work [12], the present contribution promoted the use of a Bayesian-driven criterion to estimate the regularization parameter inherent to the ℓ_1-TVℓ_0 minimization problem. The criterion was derived by drawing an equivalence between the variational problem and a hierarchical Bayesian formulation of the change-point detection problem. The equivalence also permitted to design a numerically efficient algorithm which benefits from low computational costs. The good performance of the procedure were evaluated in terms of MSE and compared to oracle estimates.

7. REFERENCES

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

[2] M. Little and N. 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] M. Nikolova, "A variational approach to remove outliers and impulse noise," *J. Math. Imag. Vis.*, vol. 20, no. 1-2, pp. 99–120, 2004.

[4] A. Mincholé, L. Sörnmo, and P. Laguna, "Detection of body position changes from the ECG using a Laplacian noise model," *Biomed. Signal Process. Contr.*, vol. 14, pp. 189–196, 2014.

[5] R. J. Marks, G. L. Wise, D. G. Haldeman, and J. L. Whited, "Detection in Laplace noise," *IEEE Trans. on Aerospace and Electronic Systems*, vol. AES-14, no. 6, pp. 866–872, Nov 1978.

[6] M. Wu and W. J. Fitzgerald, "Analytical approach to change-point detection in Laplacian noise," *IEEE Proc. Vision, Image and Signal Processing*, vol. 142, no. 3, pp. 174–180, Jun 1995.

[7] M. Storath, A. Weinmann, and L. Demaret, "Jump-sparse and sparse recovery using Potts functionals," *IEEE Trans. Signal Process.*, vol. 62, no. 14, pp. 3654–3666, July 2014.

[8] M. Storath, A. Weinmann, and M. Unser, "Exact algorithms for l^1-TV regularization of real-valued or circle-valued signals," *J. Sci. Comput.*, vol. 38, no. 1, pp. 614–630, 2016.

[9] L. Chaari, J.-C. Pesquet, J.-Y. Tourneret, and P. Ciuciu, "Parameter estimation for hybrid wavelet-total variation regularization," in *Proc. IEEE Workshop Stat. Sign. Proc.*, Nice, France, June, 28-30 2011.

[10] M. Pereyra, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Maximum-a-posteriori estimation with unknown regularisation parameters," in *Proc. Eur. Sig. Proc. Conference*, Nice, France, Aug 2015, pp. 230–234.

[11] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Adaptive total variation image deblurring: a majorization–minimization approach," *Signal Process.*, vol. 89, no. 9, pp. 1683–1693, 2009.

[12] J. Frecon, N. Pustelnik, N. Dobigeon, H. Wendt, and P. Abry, "Bayesian selection for the regularization parameter in TVℓ_0 denoising problems," 2016, arXiv preprint arXiv:1608.07739.

[13] M. Lavielle, "Optimal segmentation of random processes," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1365–1373, 1998.

[14] N. Dobigeon, J.-Y. Tourneret, and M. Davy, "Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1251–1263, Apr. 2007.

[15] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, and J.-C. Pesquet, "Wavelet-based image deconvolution and reconstruction," *Wiley Encyclopedia of EEE*, 2016.

[16] M. Lavielle and E. Lebarbier, "An application of MCMC methods for the multiple change-points problem," *Signal Process.*, vol. 81, no. 1, pp. 39–53, Jan. 2001.

[17] E. Punskaya, C. Andrieu, A. Doucet, and W. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 747–758, Mar. 2002.

[18] 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.