Nonlinear estimators for the detection of small and rare features

Sylvain Sardy
Section de Mathématiques, University of Geneva, Switzerland

Abstract
We illustrate in three settings (i.e., wavelet smoothing, total variation density estimation and wavelet-based inverse problem) the need of nonlinear estimators to retrieve small or rare features hidden in data. Such nonlinear nonparametric methods could be specifically developed for inverse problems at CERN.

1 Introduction

Consider the regression setting

$$Y_n = \mu_n + \epsilon_n, \quad n = 1, \dots, N, \tag{1}$$

where Y_n are measurements of the signal μ_n with noise ϵ_n. In the following, we write vectors in bold, e.g., $\mathbf{Y} = (Y_1, \dots, Y_N)$. Suppose $\hat{\boldsymbol{\mu}}_\lambda$ is an estimator of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ indexed by a regularization parameter λ (which use will become clear). To measure the quality of this estimator, the risk of $\hat{\boldsymbol{\mu}}_\lambda$ is defined as $\mathrm{R}(\lambda) = \mathrm{E}[(\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu})^2]$, where E stands for expectation. In practice the risk is unknown, but can be estimated from the data. Importantly the risk has a bias-variance decomposition

$$\mathrm{R}(\lambda) = \mathrm{bias}^2(\lambda) + \mathrm{Var}(\lambda).$$

In some settings, estimators like the maximum likelihood estimator (MLE) or least squares (LS) have no bias, but have a very high variance; conversely, other estimators have no variance but a high bias. The goal of regularization is to propose appropriate ways to introduce bias and to control it well with a good selection of the regularization parameter λ.

We distinguish two regression problems to illustrate regularization.

1.1 Nonparametric estimation

If μ is a univariate function (or an image) observed at points x_n (in which case $\mu_n = \mu(x_n)$) then one can try to recover μ from the data Y_n without making any strong parametric assumption on μ. Hence the linear smoothing splines estimator [1] assumes μ belongs to a Sobolev space which only imposes a smoothness class. Such an estimator performs well to estimate smooth functions.

Recently Waveshrink [2] provides a nonlinear estimator capable of detecting small and sharp features such as peaks, discontinuities or small bursts. They assume the underlying signal μ expands linearly on N orthonormal wavelets, with corresponding wavelet coefficients $\boldsymbol{\alpha}$, which form a basis of Besov spaces which include Sobolev spaces as particular cases. One can extract an orthonormal regression matrix W of dimension $N \times N$ from this representation such that (1) becomes in vector and matrix notation

$$\boldsymbol{Y} = W\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

Assuming Gaussian independent noise $\boldsymbol{\epsilon}$, the maximum likelihood estimate is obtained by applying the discrete wavelet transform (DWT) to the data: $\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}} = W^{\mathrm{T}}\boldsymbol{Y}$. It has no bias but high variance. Importantly, projected on a wavelet basis, most functions μ have a sparse wavelet representation (i.e., most entries of $\boldsymbol{\alpha}$ are zero). So [2] propose to regularize the MLE by applying componentwise a nonlinear function that enforces sparsity, for instance, the so-called soft-thresholding function

$$\hat{\boldsymbol{\alpha}}_\lambda = \left\{1 - \frac{\lambda}{|\hat{\alpha}_n^{\mathrm{MLE}}|}\right\}_+ \hat{\alpha}_n^{\mathrm{MLE}}, \quad n = 1, \dots, N, \tag{2}$$

where $\{x\}_+ = 0$ if x is negative. The smoothing parameter λ controls the bias-variance trade-off: if $|\hat{\alpha}_n^{\mathrm{MLE}}|$ is abnormally large with respect to λ, it will be kept as a significant coefficient; otherwise, it will be seen as noise and set to zero by the thresholding function. Then the estimator of the underlying signal is $\hat{\boldsymbol{\mu}}_\lambda = W\hat{\boldsymbol{\alpha}}_\lambda$. Note that for $\lambda = 0$, no thresholding/regularization is performed and we get back the MLE. Reference [3] derives near minimax results for Waveshrink.

1.2 Parametric estimation

Here we also assume that covariates $(x_1, \ldots, x_P)_n$ are observed along with Y_n for $n = 1, \ldots, N$. Linear parametric regression assumes $\mu_n = \sum_{p=1}^P \alpha_p x_{np}$, which in vector form is

$$\mathbf{Y} = X\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

The least squares estimate may have a high variance if the matrix X is badly conditioned, so the linear ridge regression estimator [4] adds a quadratic penalty to control the bias, and estimate the coefficients by solving

$$\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\eta^\eta \tag{3}$$

for $\eta = 2$, where $\|\cdot\|_\eta$ stands for the ℓ_η norm, e.g., $\|\boldsymbol{\alpha}\|_2 = \sqrt{\sum_{p=1}^P \alpha_p^2}$ and $\|\boldsymbol{\alpha}\|_1 = \sum_{p=1}^P |\alpha_p|$. Recently [5] developed the nonlinear lasso estimator for $\eta = 1$; interestingly, lasso performs model selection in the sense that the solution to (3) is a sparse vector (the larger λ the more sparse the estimated vector). Moreover, when the matrix X is orthonormal (e.g., a wavelet matrix), then (3) has a closed form solution via the soft thresholding function (2). More recently, adaptive lasso [6] is a variation of lasso that is oracle in the sense that it selects the right model with a high probability and is root-N consistent for the non-zero coefficients.

2 Nonlinear estimation to detect sharp and rare features

Based on Sections 1.1 and 1.2, we address estimation and detection of rare and sharp features in more complex settings that may be of interest towards solving inverse problems in particle physics at CERN.

2.1 Wavelet smoothing from several captors

Gravitational wave bursts are rare events expected to be produced by energetic cosmic phenomena such as the collapse of a supernova [7]. The signal-to-noise ratio is believed to be low, so that only the joint information recorded by Q captors at a high frequency of 5MHz may help prove the existence of such wave bursts. The noise is colored and possibly non-Gaussian. A good model for these data is (1) for Q signals and for $n = t$ (for time), namely

$$Y_t^{(q)} = \mu_t^{(q)} + \epsilon_t^{(q)}, \quad t = 1, \ldots, T,\ q = 1 \ldots, Q \tag{4}$$

where the noises $\epsilon^{(q)}$ and $\epsilon^{(q')}$ are independent between captors $q \neq q'$. Importantly, most of the time the underlying signal $\mu^{(q)}(t) = 0$ for all q, but, if $\mu^{(q)}(t) \neq 0$ for a given time t and captor q, then $\mu^{(q')}(t) \neq 0$ for all other captors q' at the same time t. Moreover when a wave burst occurs, we may not have $\mu^{(q)}(t) = \mu^{(q')}(t)$, but only a proportionality constant relates them, because the incoming wave burst may not hit the captor with the same angle, or the captors may not have the same sensitivity.

Assuming a wavelet representation of each $\mu^{(q)}$ for captors $q = 1, \ldots, Q$, one can estimate the wavelet coefficients from the data by

$$\hat{\boldsymbol{\alpha}}^{(q)} = W^{\mathrm{T}}\mathbf{Y}^{(q)},$$

where the DWT also has a decorrelating property [8]. Letting $\hat{\boldsymbol{\alpha}}_n = (\hat{\alpha}_n^{(1)}, \ldots, \hat{\alpha}_n^{(Q)})$ be the block of Q wavelet coefficients corresponding to the nth wavelet used in the linear expansion of the Q underlying

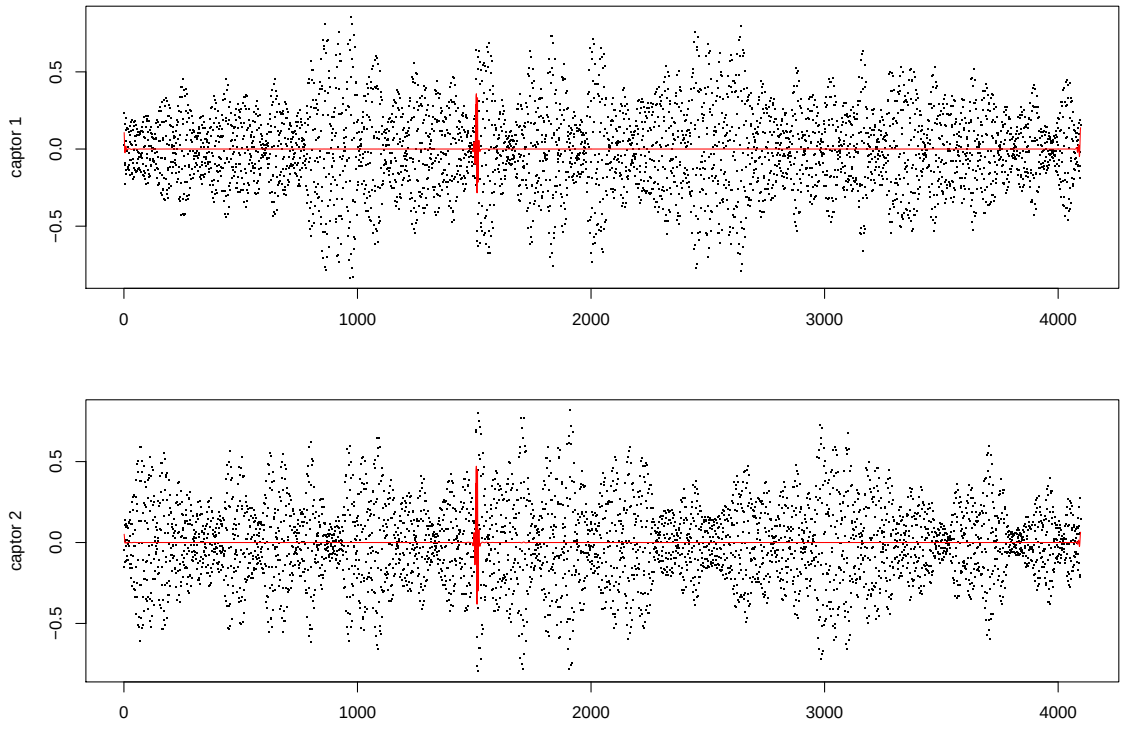

Fig. 1: Gravitational wave burst detection: concomitant and independent noisy signals recorded on $Q = 2$ captors (black dots), and the block thresholded estimator (red line).

signals μ_q, $q = 1 \ldots, Q$, we want to decrease the variance of this block by thresholding it towards zero. If this vector is abnormally large, then we will believe it contains an important feature; otherwise, all of its components will be set to zero concomitantly. To enforce this concomitant sparsity while preserving abnormally large blocks, we generalize (2) by applying the following block soft-thresholding function:

$$\hat{\boldsymbol{\alpha}}_{n,\lambda} = \left\{1 - \frac{\lambda}{\|\hat{\boldsymbol{\alpha}}_n\|_2}\right\}_+ \hat{\boldsymbol{\alpha}}_n, \quad n = 1, \ldots, N, \tag{5}$$

where $\|\boldsymbol{\alpha}\|_2 = \sqrt{\alpha_1^2 + \ldots + \alpha_Q^2}$. We can then estimate the underlying signal on each captor using the inverse DWT. Figure 1 shows typical time series recorded by the two captors (dots) in which an artificial signal resembling a wave burst has been "injected;" the red curve is the estimate based on (5). Figure 2 zooms around the time of the injection. We observe that the artificial wave burst is well detected and that the noise is well removed otherwise, although the signal to noise ratio was small.

2.2 Density estimation

Density estimation is an old problem in statistics [9–11]. Suppose a sample of size N from a density function f has been collected, and let $x_1, \ldots, x_N$ be the corresponding order statistics. The goal is to estimate f from the data x_n, $n = 1, \ldots, N$. The histogram is the commonly used nonparametric estimator, but is unstable to the choice of the binwidth and the left point. Moreover the histogram can show too many modes/bumps, as illustrated on the top graph of Figure 3. Taut string [12] is a more recent nonparametric estimator that controls the numbers of modes and that has some connection with the total variation estimator [13]. That latter estimator regularizes the likelihood with an ℓ_1-based penalty

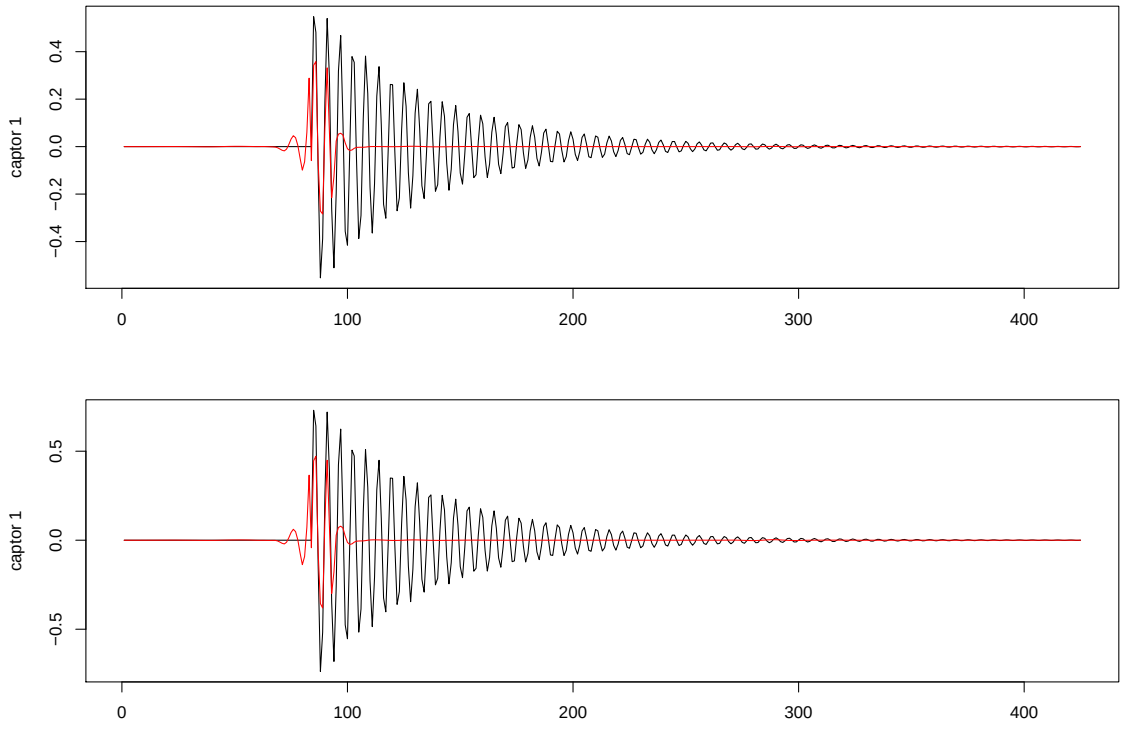

Fig. 2: Gravitational wave burst detection: zoom around the time of an "injection" (black line), and the block thresholded estimator (red line).

by solving

$$\min_{\mathbf{f}\in\mathbb{R}^N} -\sum_{i=1}^{N} \log f_i + \lambda \sum_{i=2}^{N} |f_n - f_{n-1}|, \quad \text{s.t.} \quad \mathbf{a}^T\mathbf{f} = 1, \tag{6}$$

where the equality constraint forces the estimated function to integrate to one (to be a density), and where $a_1 = (x_2 - x_1)/2$, $a_N = (x_N - x_{N-1})/2$ and $a_n = (x_{n+1} - x_{n-1})/2$ for $n = 2, \ldots, N-1$. Here λ controls the smoothness of the estimate. The total variation estimate (middle graph) of Figure 3 illustrates its ability to estimate the underlying density without unnecessary bumps.

2.3 Inverse problem

Likewise in the inverse problem, one can retrieve bumps from data quite well by developing an appropriate nonlinear estimator. Suppose the sample $Y_1, \ldots, Y_N$ measures with noise an unknown function f through a known linear operator $\mathcal{K}$ at known locations $\boldsymbol{t} = (t_1, \ldots, t_N)$ in Ω according to

$$Y_n = \mathcal{K}f(t_n) + \epsilon_n, \quad n = 1, \ldots, N. \tag{7}$$

We propose to expand f linearly on a wavelet basis W and regularize the least squares problem with lasso, i.e., (3) with $\eta = 1$ to enforce a sparse wavelet estimation. Hence we solve

$$\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - KW\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

where the smoothing parameter λ is chosen to minimize an estimate of the risk. Figure 4 illustrates the power of this nonlinear estimator (red curve) to retrieve peaks (the green curve is the underlying function to retrieve) from a blurred and noisy signal (black line). Some peaks that had disappeared with the blurring can be retrieved surprisingly well.

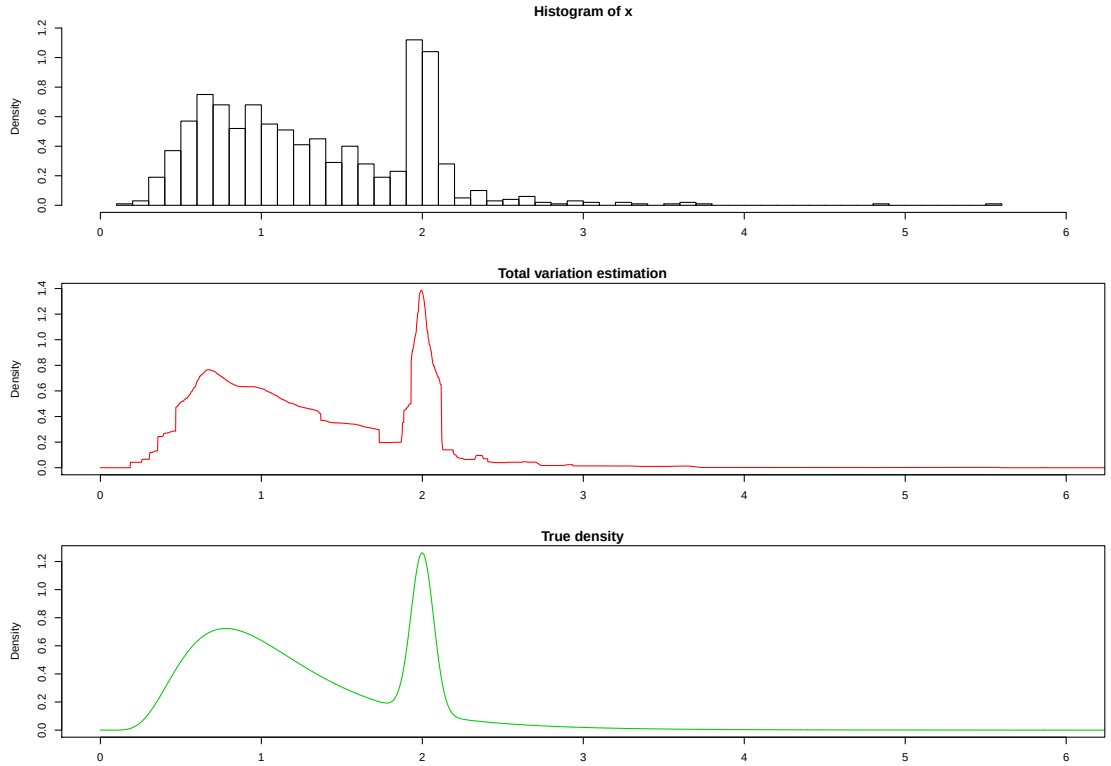

Fig. 3: Looking for bumps in a density. Histogram (top), total variation estimate (middle), true density (bottom).

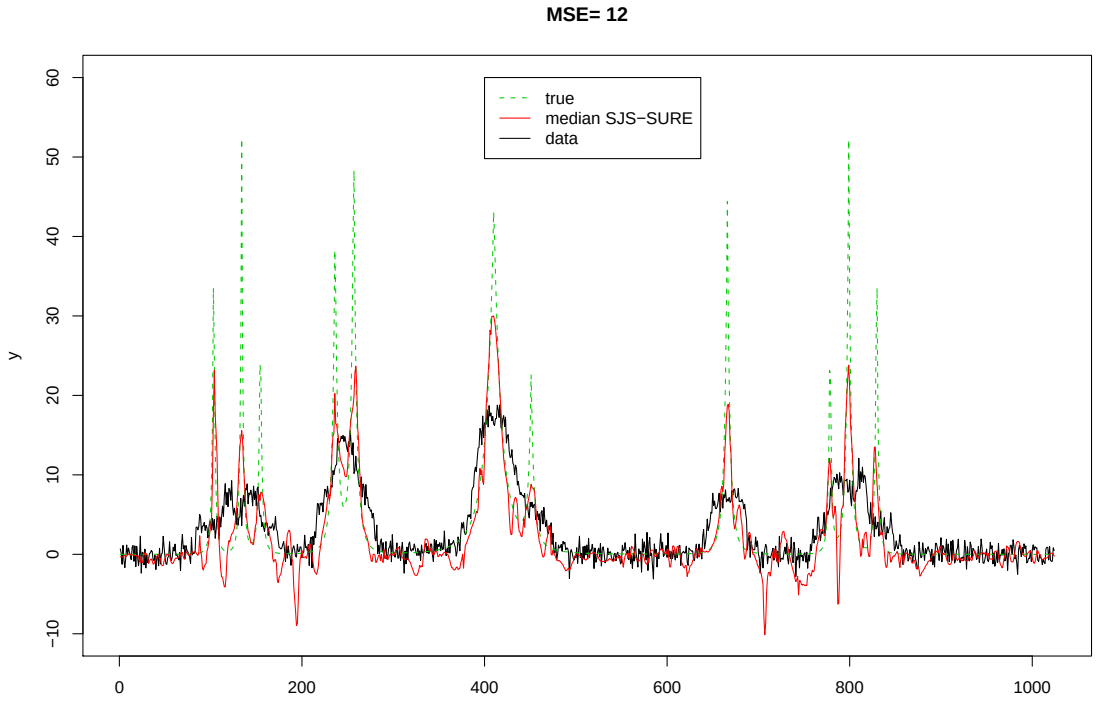

Fig. 4: Looking for bumps in a blurring inverse problem: true underlying function (green), data (black) and nonlinear estimate (red).

3 Conclusion

The three settings considered (i.e., regression, density estimation and inverse problem) illustrate the ability of nonlinear nonparametric estimators to retrieve sharp and rare features from data. Developing such estimators for the specificities of inverse problems encountered at CERN is challenging and will reveal whether these estimators can enhance discoveries on CERN real applications.

4 Acknowledgements

We thank David Hand for his comments on a first draft, and Stefano Foffa, Roberto Terenzi and the ROG group for providing a sample of the astrophysics data in Figure 1.

References

[1] Grace Wahba, *Spline Models for Observational Data*, SIAM, 1990.

[2] David Donoho and Iain Johnstone, Ideal Spatial Adaptation via Wavelet Shrinkage, Biometrika, p. 425–455, 1994.

[3] David Donoho, Iain Johnstone, Gérard Kerkyacharian and Dominique Picard, Wavelet Shrinkage: Asymptopia? (with discussion), Journal of the Royal Statistical Society, Series B: Methodological, p. 301–369, 1995.

[4] Arthur Hoerl and Robert Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, p. 55–67, 1970.

[5] Robert Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society, Series B: Methodological, p. 267–288, 1996.

[6] Hui Zou, The Adaptive LASSO and Its Oracle Properties, Journal of the American Statistical Association, vol. 101, p. 1418–1429, 2006.

[7] Sergey Klimenko and Guenakh Mitselmakher, A wavelet method for detection of gravitational wave bursts, Classical and Quantum Gravity, p. 1819–1830, 2004.

[8] Iain Johnstone and Bernard Silverman, Wavelet Threshold Estimators for Data with Correlated Noise, Journal of the Royal Statistical Society, Series B: Methodological, p. 319–351, 1997.

[9] David Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, 1992.

[10] Bernard Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.

[11] Jeffrey Simonoff, *Smoothing Methods in Statistics*, Springer-Verlag, 1996.

[12] Laurie Davies and Arne Kovac, Densities, spectral densities and modality, The Annals of Statistics, p. 1093–1136, 2004.

[13] Sylvain Sardy and Paul Tseng, Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion, Scandinavian Journal of Statistics, p. 321–337, 2010.