ASSUMED AND EFFECTIVE PRIORS IN BAYESIAN MAP ESTIMATION

Mila Nikolova

Laboratoire Traitement et Communication de l'Information (CNRS–ENST)
ENST, 46, rue Barrault, 75634 Paris Cédex 13, France
E-mail : nikolova@tsi.enst.fr

ABSTRACT

Bayesian *maximum a posteriori* estimation (MAP) is a very popular way to recover unknown signals and images by using jointly observed data and priors formulated as a probability law. In a variational context, a MAP estimate minimizes an objective function where the priors are seen as a regularization or diffusion term.

Independently of such interpretations, MAP estimates are *implicit functions* of the data and of the *functions* expressing the priors. This point of view enabled the author to exhibit analytical relations between prior functions and the features of the relevant estimates. These results entail important consequences and questions which are the subject of this paper. Namely, they reveal an essential gap between prior models and the way these are *effectively* involved in a MAP estimate. Hence the question about the rationale of MAP estimation. At the same time, they give precious indications about the hyperparameters and suggest how to construct estimators which indeed respect the priors.

1. BAYESIAN MAP ESTIMATION

We address the problem of the estimation of a magnitude $\boldsymbol{x} \in \mathbf{R}^p$—a signal, an image—from observed data $\boldsymbol{y} \in \mathbf{R}^q$. The observation system which transforms $\boldsymbol{x}$ into $\boldsymbol{y}$ is typically composed of an operator (blurring, obscurations, nonlinear transforms, *etc.*) and involves random noise effects. The likelihood function $p(\boldsymbol{y}|\boldsymbol{x})$ resumes the information about the unknown $\boldsymbol{x}$ contained in teh observed $\boldsymbol{y}$. In many situations $\boldsymbol{x}$ cannot be determined unambigously from $p(\boldsymbol{y}|\boldsymbol{x})$ only. Nevertheless, it is often possible to extract priors about the unknown magnitude based on the context and the expectations (such as presence of edges, spikes, homogeneous zones, *etc.*) Numerous works are devoted to the construction of a prior probability density $p(\boldsymbol{x})$ based on such "diffuse" informations [1, 2]. Bayesian estimation [3] relies on the posterior law $p(\boldsymbol{x}|\boldsymbol{y})$ which contains all the information about $\boldsymbol{x}$ after having observed the data $\boldsymbol{y}$. A MAP estimate minimizes a 0-1 loss function and reads

$$\widehat{\boldsymbol{x}} = \arg\max p(\boldsymbol{x}|\boldsymbol{y}) = \arg\min\left[-\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{x})\right]$$

This paper is aimed at being a "counter-example" of Bayesian MAP estimation hence there is no loss of generality to focus on particular classes of models. Thus we consider data $\boldsymbol{y}$ obtained at the output of a linear system $\mathcal{A}$ and perturbed by white Gaussian noise, and priors $p(\boldsymbol{x})$ formalized using Markov models. Then $\widehat{\boldsymbol{x}}$, defined above, minimizes an energy function of the form:

$$\mathcal{E}(\boldsymbol{x}) = \|\mathcal{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \beta\Phi(\boldsymbol{x}) \qquad (1)$$

where Φ is the energy of a Markov random model and β reflects the noise variance. For definiteness, we focus on Markov chains defined over the differences $\boldsymbol{d}_k^T\boldsymbol{x}$ between neighbouring samples [4, 5]:

$$\Phi(\boldsymbol{x}) = \sum_k \varphi(\boldsymbol{d}_k^T\boldsymbol{x}) \qquad (2)$$

where φ is called potential function (PF) while typically $\boldsymbol{d}_k^T\boldsymbol{x} = x_k - x_{k+1}$ or $\boldsymbol{d}_k^T\boldsymbol{x} = 2x_k - x_{k-1} - x_{k+1}$. Recall that φ is symmetric and increasing on $[0, \infty[$ and $\mathcal{C}^2$ almost everywhere. Notice that Φ as given above does not define a proper probability measure since the partition function $Z = \int \exp[-\Phi(\boldsymbol{x})]d\boldsymbol{z}$ is finite only for $\boldsymbol{x}$ belonging to a bounded set. But this fact is not disturbing as far as we are mainly concerned with the differences rather than with the values of the samples. To avoid such complications, we put

$$t_k := \boldsymbol{d}_k^T\boldsymbol{x}$$

In the sequel we compare the priors expressed by

$$p(\boldsymbol{t}) = \prod_k p(t_k) \quad \text{with} \quad p(t_k) = \frac{\exp\left[-\varphi(t_k)\right]}{Z_k} \qquad (3)$$

and some persistant characteristics of $\widehat{\boldsymbol{x}}$ originating in the shape of φ.

0-7803-6293-4/00/$10.00 ©2000 IEEE.

2. NONSMOOTH AT ZERO PRIORS

Let us focus on a Markov model where the differences among neighbouring samples are realizations of a Laplacian distribution, *i.e.*

$$\varphi(t) = \alpha|t|$$

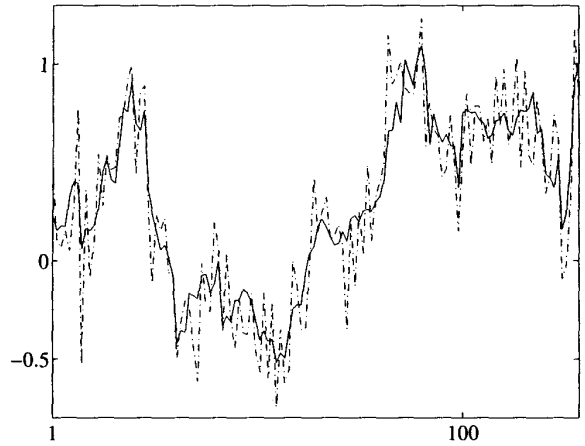

Figure 1: Markov chain $\boldsymbol{x}$ (—) where the differences $\boldsymbol{d}_k^T\boldsymbol{x} = x_k - x_{k+1}$ follow a Laplace distribution with $\alpha = 10$. Data $\boldsymbol{y}$ (-.-.-) obtained by adding to $\boldsymbol{x}$ white Gaussian noise with variance $\sigma^2 = 0.04$.

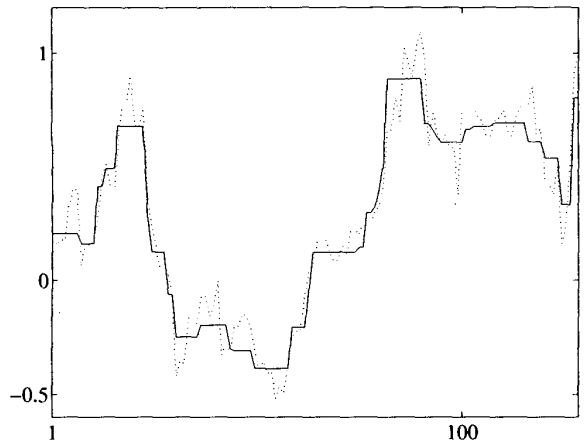

Figure 2: MAP estimate $\widehat{\boldsymbol{x}}$ (—) obtained using the true prior distribution and the true hyperparameters. It can be compared with the original chain $\boldsymbol{x}$ (....)

We perform the following experiment which is illustrated on Figs. 1-3. The original chain $\boldsymbol{x}$ is defined over differences between adjacent samples (Fig. 1) and its differences give rise to a rather flat histogram centered at zero (Fig. 3-up). Since the relevant $p(\boldsymbol{t})$ is continuous, the probability that $\boldsymbol{x}$ involves zero-valued differences is null. Then we generate data $\boldsymbol{y}$ (Fig. 1, mixed line) by adding to $\boldsymbol{x}$ white Gaussian noise with known variance σ^2. Thus we are placed in the ideal situation where we know the true prior distribution as well as the true values of the parameters. Equipped with all this knowledge, we perform a MAP estimation. The result $\widehat{\boldsymbol{x}}$, shown in Fig. 2 with (—), produces a striking visual effect: the obtained estimate contains *large constant zones, i.e.* it involves a large number of zero-valued differences. The presence of such constant zones constitutes a *highly organized structural information.* What is disturbing is that this so strong structural information was not modeled in the prior $p(\boldsymbol{x})$ but is entirely introduced by the estimator!

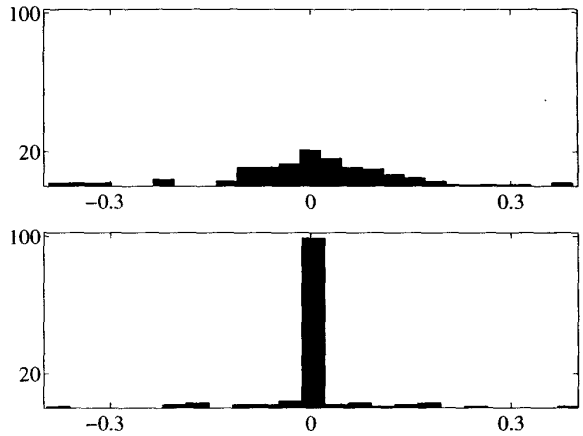

Figure 3: *Up*: histogram of the differences $x_k - x_{k+1}$ of original chain $\boldsymbol{x}$. *Down*: histogram of the differences of the MAP estimate $\widehat{\boldsymbol{x}}$— it contains 96 null differences.

This phenomenon can be explained thanks to [6]. The key to understand it is the following result [6, 7]:

Theorem. *Consider an energy function $\mathcal{E}(.,\boldsymbol{y})$ as given in (1-2) with φ a PF which is $\mathcal{C}^2$ continuous beyond zero and nonsmooth at zero. Let $\widehat{\boldsymbol{x}}$ be a local minimizer of $\mathcal{E}(.,\boldsymbol{y})$ which (i) gives rise to a locally continuous implicit minimizer function $\mathcal{X}$ (so $\widehat{\boldsymbol{x}} = \mathcal{X}(\boldsymbol{y})$), and (ii) involves a (large) set of zero-valued differences, say $\widehat{J} := \{k : \boldsymbol{d}_k^T\widehat{\boldsymbol{x}} = 0\}$.*

Then $\boldsymbol{y}$ is contained in an open neighbourhood $\mathcal{N}(\boldsymbol{y})$ whose data $\boldsymbol{y}' \in \mathcal{N}(\boldsymbol{y})$ yield local minimizers $\widehat{\boldsymbol{x}}' = \mathcal{X}(\boldsymbol{y}')$ which involve exactly the same zero-valued differences as $\widehat{\boldsymbol{x}}$. In other words, $\{k : \boldsymbol{d}_k^T\widehat{\boldsymbol{x}}' = 0\} = \widehat{J}$ for all $\boldsymbol{y}' \in \mathcal{N}(\boldsymbol{y})$.

In [6, 7] the zones of $\boldsymbol{x}$ relevant to zero-valued differences $\boldsymbol{d}_k^T\boldsymbol{x} = 0$ were said to be *strongly homogeneous.* The above Theorem says that the data domain $\mathbf{R}^q$ contains sets of positive Lebesgue measure, which are composed of data which yield minimizers shareing the same strongly homogeneous zones. Re-

ciprocally, we get minimizers—and in particular MAP estimates—involving large strongly homogeneous zones (as illustrated in Figs. 1-3) independently of the fact that such zones are, or are not, present in the original $\boldsymbol{x}$, and independently of the noise corrupting the data. Moreover, the phenomenon evoked above is produced in both signals and images under prety general conditions: ($\bullet$) *any* C^2-smooth log-likelihood function, ($\bullet$) *any* set of linear operators $\{\boldsymbol{d}_k, k \geq 1\}$, ($\bullet$) *any* potential function φ which is *nonsmooth at zero.*

Let us return back to our MAP estimation. A prior distribution $p(\boldsymbol{x})$ of the form of (2-3), with φ continuous nonsmooth at zero, provides a "diffuse" prior information which can be expected to fit to a broad range of signals or images. However, *the MAP estimator transforms this "diffuse" prior into a highly structured strong prior saying that the reconstructed signals contain large strongly homogeneous zones!* We can state that the latter is the *effective prior* conveyed by an MAP estimator whenever $p(\boldsymbol{x})$ is nonsmooth at the origin. So, if $p(\boldsymbol{x})$ is a *bona fide* prior distribution—whose realizations exhibit useful features of the unknown $\boldsymbol{x}$—then the MAP estimation will fail to use correctly this prior. Conversely, if we have to segment a signal, or an image, into strongly homogeneous zones, it is sufficient to introduce in (2) a PF φ which is nonsmooth at zero. Following this idea, the possibility to recover quasi-binary images using convex nonsmooth PFs is explored in [8].

3. PIECEWISE GAUSSIAN MODELS

We now focus on priors defined using a truncated quadratic potential function

$$\varphi(t) = \begin{cases} \lambda^2 t^2 & \text{if } |t| < \theta, \\ \alpha & \text{if } |t| \geq \theta, \end{cases} \quad \text{where } \theta = \frac{\sqrt{\alpha}}{\lambda} \qquad (4)$$

with parameters $\alpha > 0$, $\lambda > 0$. Such priors can be seen as piecewise Gaussian and involve a line-process [9, 10].

The distribution $p(t_k)$ induced by φ is defined on a bounded interval, say $[-T, T]$ with $T > \theta$. It has a Gaussian shape on $[-\theta, \theta]$ and is uniform beyond it. The differences in $\boldsymbol{x}$ with values in $[-\theta, \theta]$ belong to homogeneous zones whereas those beyond it form edges. A t_k generated according to $p(t_k)$ may thus take any value on $[-T, T]$.

Let us envisage an experiment similar to the latter one. We generate an original chain $\boldsymbol{x}$ (see Fig. 4) using $p(\boldsymbol{t})$. The noisy data, plotted on the same Figure with (-.-.-), are corrupted by additive white Gaussian noise. Then we perform a MAP estimation where we use the true prior distribution and the true parameters. The obtained MAP estimate (Fig. 5) involves very neat edges. It was observed that piecewise Gaussian priors always give rise to MAP estimates which always have *very neat edges.* Besides, this intuitively corresponds to the notion of line process. By [11, 12], the fact which underlines this behaviour is that the threshold θ is placed in the interior of an interval where the differences of the global minimizer of $\mathcal{E}$ cannot be placed. In a more formal way, we have the following result:

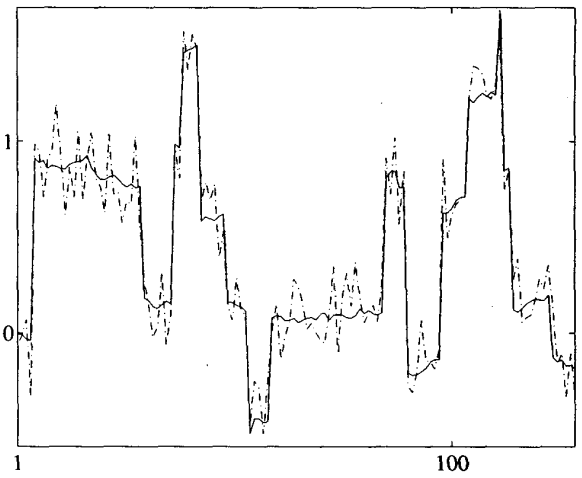

Figure 4: Markov chain $\boldsymbol{x}$ (—) whose differences $t_k = x_k - x_{k+1}$ are generated according to (3,4) with $(\alpha, \lambda) = (6, 50)$. Data $\boldsymbol{y}$ (-.-.-) are corrupted by additive white Gaussian noise with variance $\sigma^2 = 0.0225$.

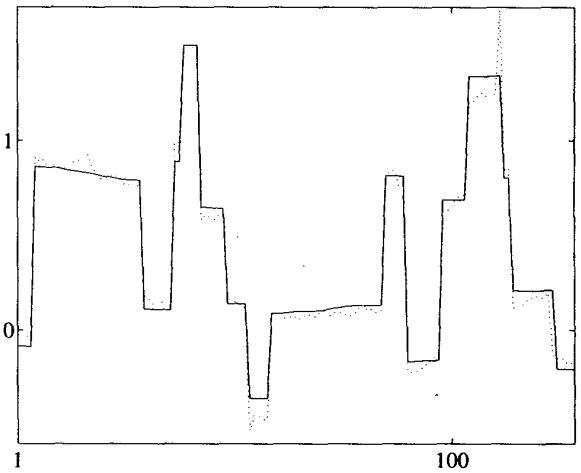

Figure 5: The MAP estimate $\widehat{\boldsymbol{x}}$ (—) is calculated using the true prior and the true parameters. The original signal is recalled with (....).

Theorem. *Consider $\mathcal{E}(., \boldsymbol{y})$ defined using a truncated quadratic PF (4). Let $\widehat{\boldsymbol{x}}$ be a global minimizer of $\mathcal{E}(., \boldsymbol{y})$.*

Then with each k there is associated a constant $\Gamma_k \in [0, 1[$ such that $\boldsymbol{d}_k^T \widehat{\boldsymbol{x}}$ satisfies the following alternative:

either $|d_k^T \widehat{x}| \leq \theta\Gamma_k$ *or* $|d_k^T \widehat{x}| \geq \theta/\Gamma_k$ *where the second possibility exists only if* $\Gamma_k > 0$.

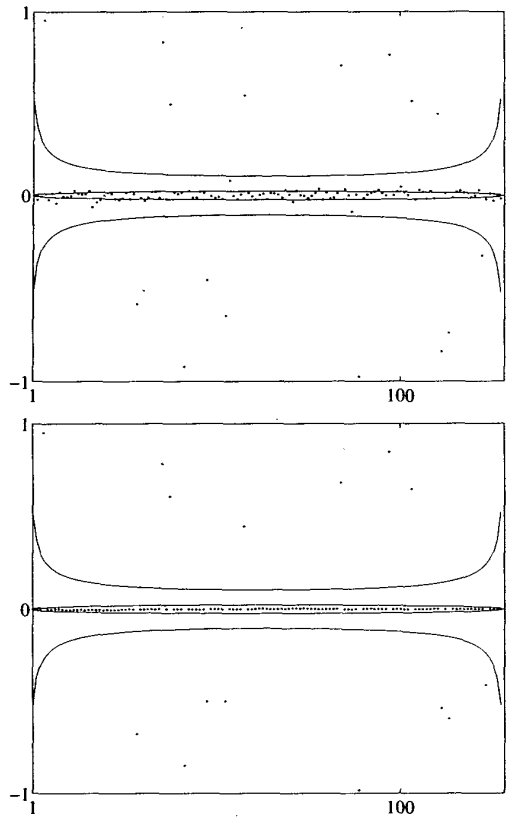

Figure 6: The thresholds $\pm\theta\Gamma_k$ and $\pm\theta/\Gamma_k$ are plotted with (—). X-axis: positions of the differences for $k = 1, \ldots, 127$. Y-axis: a dot at position k is the value of the k-th difference of the relevant signal. *Up*: distribution of the differences of the original chain $\boldsymbol{x}$ (they can be located everywhere on $[-T, T]$). *Down*: distribution of the differences of the MAP estimate $\widehat{\boldsymbol{x}}$ (their magnitudes are inevitably beyond $]\theta\Gamma_k, \theta/\Gamma_k[$).

In other words, this estimator implies a very hard thresholding where the magnitude of the differences between neighbouring samples at a global minimizer are either smaller than a first threshold, or larger than a second threshold which is strictly larger than the first threshold. Conversely, no difference corresponding to a global minimizer can be placed among these thresholds, for any data—this is seen in Fig. 6. This constitutes the *effective prior* applied by the MAP estimator. As previously, the MAP estimation introduces additional structural information which is not present in the prior distribution.

4. CONCLUSIONS

Based on some recent analytical results about the minimizers of regularized objective functions, we perform a critical analysis of Bayesian MAP estimation. More precisely, we reveal an essential gap between prior models and the way these are effectively involved in a MAP estimate. At the same time, this knowlegde can be used to construct estimators which do respect the priors.

5. REFERENCES

[1] J. E. Besag, "On the statistical analysis of dirty pictures (with discussion)", *Journal of the Royal Statistical Society B*, vol. 48, no. 3, pp. 259–302, 1986.

[2] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, Chapman et Hall, London, 1995.

[3] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[5] J. E. Besag, "Digital image processing : Towards Bayesian image analysis", *Journal of Applied Statistics*, vol. 16, no. 3, pp. 395–407, 1989.

[6] M. Nikolova, "Local strong homogeneity of a regularized estimator", *SIAM Journal on Applied Mathematics*, to appear, 1999.

[7] M. Nikolova, "Reconstruction of locally homogeneous images", *(submitted to SIAM)*, 1999.

[8] M. Nikolova, "Estimation of binary images using convex criteria", in *Proceedings of the International Conference on Image Processing*, 1998, vol. 2, pp. 108–112.

[9] F. Jeng and J. Woods, "Compound Gauss-Markov random fields for image estimation", *IEEE Transactions on Signal Processing*, vol. SP-39, no. 3, pp. 683–697, Mar. 1991.

[10] J. Marroquin, "Deterministic interactive particle models for image processing and computer graphics", *Computer Vision and Graphics and Image Processing*, vol. 55, no. 5, pp. 408–417, 1993.

[11] A. Blake and A. Zisserman, *Visual reconstruction*, The MIT Press, Cambridge, 1987.

[12] M. Nikolova, "Thresholding implied by truncated quadratic regularization", Tech. Rep., TSI–ENST, Paris, France, 1999.