SCALE INVARIANT MARKOV MODELS FOR BAYESIAN INVERSION OF LINEAR INVERSE PROBLEMS

Stéphane Brette, Jérôme Idier and Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes (CNRS-ESE-UPS)
École Supérieure d'Électricité,
Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France

ABSTRACT. In a Bayesian approach for solving linear inverse problems one needs to specify the prior laws for calculation of the posterior law. A cost function can also be defined in order to have a common tool for various Bayesian estimators which depend on the data and the hyperparameters. The Gaussian case excepted, these estimators are not linear and so depend on the scale of the measurements. In this paper a weaker property than linearity is imposed on the Bayesian estimator, namely the scale invariance property (SIP).

First, we state some results on linear estimation and then we introduce and justify a scale invariance axiom. We show that arbitrary choice of scale measurement can be avoided if the estimator has this SIP. Some examples of classical regularization procedures are shown to be scale invariant. Then we investigate general conditions on classes of Bayesian estimators which satisfy this SIP, as well as their consequences on the cost function and prior laws. We also show that classical methods for hyperparameters estimation (*i.e.*, Maximum Likelihood and Generalized Maximum Likelihood) can be introduced for hyperparameters estimation, and we verify the SIP property for them.

Finally we discuss how to choose the prior laws to obtain scale invariant Bayesian estimators. For this, we consider two cases of prior laws : *entropic prior laws* and *first-order Markov models*. In related preceding works [1, 2], the SIP constraints have been studied for the case of entropic prior laws. In this paper extension to the case of first-order Markov models is provided.

KEY WORDS : Bayesian estimation, Scale invariance, Markov modelling, Inverse Problems, Image reconstruction, Prior model selection

1. Introduction

Linear inverse problem is a common framework for many different objectives, such as reconstruction, restoration, or deconvolution of images arising in various applied areas [3]. The problem is to estimate an object $\boldsymbol{x}$ which is indirectly observed through a linear operator A, and is therefore noisy. We choose explicitly this linear model because its simplicity captures many of interesting features of more complex models without their computational complexity. Such a degradation models allows the following description:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{b}$ includes both the modeling errors and unavoidable noise of any physical observation system, and $\boldsymbol{A}$ represents the indirect observing system and depends on a particular application. For example, $\boldsymbol{A}$ can be diagonal or block-diagonal in deblurring, Toeplitz or bloc-Toeplitz in deconvolution, or have no special interesting form as in X-ray tomography.

In order to solve these problems, one may choose to minimize the quadratic residual error $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$. That leads to the classical linear system

$$\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{x}} = \boldsymbol{A}^t\boldsymbol{y}. \tag{2}$$

When mathematically exact solutions exist, they are too sensitive to unavoidable noise and so are not of practical interest. This fact is due to a very high condition number of A [3]. In order to have a solution of interest, we must mathematically qualify admissible solutions.

The Bayesian framework is well suited for this kind of problem because it could combine information from data $\boldsymbol{y}$ and prior knowledge on the solution. One needs then to specify the prior laws $p_x(\boldsymbol{x};\boldsymbol{\lambda})$ and $p_b(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x};\boldsymbol{\psi})$ for calculation of the posterior $p_{x|y}(\boldsymbol{x}|\boldsymbol{y}) \propto p_x(\boldsymbol{x})\, p_b(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})$ with the Bayes rules. Most of the classical Bayesian estimators, *e.g.*, Maximum *a posteriori* (MAP), Posterior Mean (PM) and Marginal MAP (MMAP), can be studied using the common tool of defining a cost function $C(\boldsymbol{x}^*, \boldsymbol{x})$ for each of them. It leads to the classical Bayesian estimator

$$\widehat{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{\theta}) = \arg\min_{\boldsymbol{x}} \left\{ \mathrm{E}_{\boldsymbol{x}^*|\boldsymbol{y}} \left\{ C(\boldsymbol{x}^*, \boldsymbol{x}) | \boldsymbol{y} \right\} \right\} \tag{3}$$

depending both on data $\boldsymbol{y}$ and hyperparameters $\boldsymbol{\theta}$.

Choosing a prior model is a difficult task. This prior model would include our prior knowledge. Some criteria based on information theory and maximum entropy principle, have been used for that. For example, when our prior knowledge are the moments of the image to be restored, application of maximum entropy principle leads Djafari & Demoment [4] to exact determination of the prior, including its parameters. Knowledge of the bounds (a gabarit) and the choice of a reference measure leads LeBesnerais [5, 6] to the construction of a model accounting for human shaped prior in the context of astronomic deconvolution.

We consider the case when there is no important and quantitative prior information such as the knowledge of moment or bounds of the solution. Then we propose to reduce the arbitrariness of the choice of prior model by application of constraint to the resulting Bayesian estimator. The major constraint for the estimator is to be scale invariant, that is, whichever the scale or physical unit we choose, estimation results must be identical. This desirable property will reduce the possible choice for prior models and make it independent of the unavoidable scale choice. In this sense, related works of Jaynes [7] or Box & Tiao [8] on non-informative prior are close to our statement, although in these works the ignorance is not limited to the measurement scale. In our work, qualitative information only is supposed to be known (positivity excepted), so we think of choosing a parametric family of probability laws as a usual and natural way in accounting for the prior. The parameters estimation in the chosen family of laws will be done according to the data, with a Maximun Likelihood (ML) or the Generalized Maximum Likelihood (GML) approach. These approaches are shown in this paper to be scale invariant.

One can criticize choosing the prior law from a desired property of the final estimator rather than from the available prior knowledge. We do not maintain having exactly chosen a model but just restricting the available choice. Then Gaussian or convex prior popularity is due likely to the tractability of the associated estimator rather than Gaussianity or convexity of the modeling process. Lastly, good as the model is, its use depends on the tradeoff between the good behavior of the final estimator and the quality of estimation.

The paper is organized as follows. First, we state some known results on Gaussian estimators as well as introduce and justify the imposition of scale invariance property (SIP)

onto the estimator. This will be done in section 2 with various examples of scale invariant models. In section 3 we prove a general theorem for a Bayesian estimator to be scale invariant. This theorem states a sufficient condition on the prior laws which can be used for reducing the choice to admissible priors. For this, we consider two cases of prior laws : *entropic prior laws* and *first-order Markov models*. In related preceding works [1, 2], the SIP constraints has been studied for the case of entropic prior laws. In this paper we extend that work to the case of first-order Markov models.

2. Linearity and scale invariance property

In order to better understand the scale invariance property (SIP), in the next subsection we consider in detail the classical case of linear estimators. First, let us define linearity as combination of additivity:

$$\forall \boldsymbol{y}_1, \boldsymbol{y}_2, \quad \left\{ \begin{array}{l} \boldsymbol{y}_1 \mapsto \widehat{\boldsymbol{x}}_1 \\ \boldsymbol{y}_2 \mapsto \widehat{\boldsymbol{x}}_2 \end{array} \right. \implies \boldsymbol{y}_1 + \boldsymbol{y}_2 \mapsto \widehat{\boldsymbol{x}}_1 + \widehat{\boldsymbol{x}}_2, \tag{4}$$

and the scale invariance property (SIP):

$$\forall \boldsymbol{y}, \quad \boldsymbol{y} \mapsto \widehat{\boldsymbol{x}} \implies \forall k, \quad k\boldsymbol{y} \mapsto k\widehat{\boldsymbol{x}}. \tag{5}$$

Linearity includes the SIP and so is a stronger property. We show a particular case how the SIP is satisfied in these linear models.

2.1. Linearity and Gaussian assumptions

Linear estimators under Gaussian assumptions have been (and probably still are) the most studied Bayesian estimators because they lead to an explicit estimation formula. In a similar way their practical interest is due to their easy implementation, such as Kalman filtering. In all these cases, prior laws have the following form:

$$p_x(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_x)^t \boldsymbol{\Sigma}_x^{-1} (\boldsymbol{x} - \boldsymbol{m}_x)\right), \tag{6}$$

whereas the conditional additive noise is often a zero mean Gaussian process $\mathcal{N}(0, \Sigma_b)$.

Minimization of the posterior likelihood for all the three classical cost functions MAP, PM and MMAP is the same as those of a quadratic form. It leads to the general form of the solution:

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^t \boldsymbol{\Sigma}_b^{-1} \boldsymbol{A} + \boldsymbol{\Sigma}_x^{-1} \boldsymbol{I})^{-1} (\boldsymbol{A}^t \boldsymbol{\Sigma}_b^{-1} \boldsymbol{y} + \boldsymbol{\Sigma}_x^{-1} \boldsymbol{m}_x) \tag{7}$$

which is a linear estimator.

Some particular cases follow:

- Case where $\boldsymbol{\Sigma}_x^{-1} = 0$ and $\boldsymbol{\Sigma}_b = \sigma_b^2 \boldsymbol{I}$. This can be interpreted as degenerated uniform prior of the solution. The solution is the minimum variance one and is rarely suitable due to the high condition number of A.

- Case where $\boldsymbol{\Sigma}_b = \sigma_b^2 \boldsymbol{I}$ and $\boldsymbol{\Sigma}_x = \sigma_x^2 \boldsymbol{I}$. This leads to the classical Gaussian inversion formula:

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^t \boldsymbol{A} + \mu \boldsymbol{I})^{-1} (\boldsymbol{A}^t \boldsymbol{y} + \mu \boldsymbol{m}_x), \quad \text{with} \quad \mu = \sigma_b^2 / \sigma_x^2, \tag{8}$$

The Signal-to-noise ratio (SNR) $\mu = \sigma_x^2/\sigma_b^2$ appears explicitly and serves as a scale invariant parameter. It plays therefore the meaningful role of a hyperparameter.

- The Gauss-Markov regularization case, which considers a smooth prior of the solution, is specified by setting $\boldsymbol{\Sigma}_x^{-1} = \mu \boldsymbol{D}^t \boldsymbol{D} + \sigma_x^{-2} \boldsymbol{I}$, with $\boldsymbol{D}$ a discrete difference matrix.

For all these cases, estimate $\widehat{\boldsymbol{x}}$ depends on a scale. Let us look at the dependence. For that matter, suppose that we change the measurement scale. For example, if both $\boldsymbol{x}$ and $\boldsymbol{y}$ are optic images where each pixel represents the illumination (in Lumen) onto the surface of an optical device, we measure the number of photons coming into this device. (This could be of practical interest for X-ray tomography.) Then we convert $\boldsymbol{y}$ into the new chosen scale and simultaneously update our parameters $\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_b$ and $\boldsymbol{m}_x$. Estimation formula is then given by

$$\widehat{\boldsymbol{x}}_k = (\boldsymbol{A}^t k^{-2} \boldsymbol{\Sigma}_b^{-1} \boldsymbol{A} + k^{-2} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{I})^{-1} (\boldsymbol{A}^t k^{-2} \boldsymbol{\Sigma}_b^{-1} k \boldsymbol{y} + k^{-2} \boldsymbol{\Sigma}_x^{-1} k \boldsymbol{m}_x), \tag{9}$$

or, canceling the scale factor k:

$$\widehat{\boldsymbol{x}}_k = k(\boldsymbol{A}^t \boldsymbol{\Sigma}_b^{-1} \boldsymbol{A} + \boldsymbol{\Sigma}_x^{-1} \boldsymbol{I})^{-1} (\boldsymbol{A}^t \boldsymbol{\Sigma}_b^{-1} \boldsymbol{y} + \boldsymbol{\Sigma}_b^{-1} \boldsymbol{m}_x). \tag{10}$$

Thus, if we take care of hyperparameters, the two restored images are physically the same.

This property is rarely stated in the Gaussian case, which can be explained by the use of SNR as a major tool of reasoning. Thus if we set the SNR, then $\widehat{\boldsymbol{x}}_k$ and $k\widehat{\boldsymbol{x}}$ are equal.

In many cases Gaussian assumptions are fulfilled, often leading to fast algorithms for calculating the resulting linear estimator. We focus on the case where Gaussian assumptions are too strong. It is the case when Gauss-Markov models are used, leading to smoother restoration than wanted. It might be explained by the short probability distribution tails which make discontinuity rare and which prevent appearing of wide homogeneous areas into the restored image.

2.2. Scale invariance basics

Although the particular case considered above may appear obvious, it is at the base of the scale invariance axiom. In order to estimate or to compare physical parameters, we must choose a scale measurement. This can have a physical meaningful unit or only a grey-level scale in computerized optics. Anyway we have to keep in mind that a physical unit or scale is just a practical but arbitrary tool, both common and convenient. As a consequence of this remark we state the following axiom of scale invariance:

Estimation results must not depend on the arbitrary choice of the scale measurement.

This is true when scale measurement depends on time exposure (astronomic observations, Positron emission tomography, X-ray tomography, etc.). Estimation results with two different values of time exposure must be coherent. SIP is also of practical interest when exhaustive tests are required for the validation.

Let us have a look on some regularized criteria for Bayesian estimation. In all the cases, the MAP criterion is used, and the estimators take the following form:

$$\widehat{\boldsymbol{x}}(\boldsymbol{y}; \boldsymbol{\psi}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{x}} \left\{ -\log p_b(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}; \boldsymbol{\psi}) - \log p_x(\boldsymbol{x}; \boldsymbol{\lambda}) \right\}. \tag{11}$$

L_p–norm estimators: General form of those criteria involves an L_p–norm rather than a quadratic norm. Then, the noise models and prior models take the following form:

$$p_b(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}; \psi) \propto \exp\left[\psi \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_p\right] \tag{12}$$

and

$$p_x(\boldsymbol{x}; \lambda) \propto \exp\left[\lambda \|\boldsymbol{M}\boldsymbol{x}\|_q\right], \tag{13}$$

where $\boldsymbol{M}$ can be a difference matrix as used by Bouman & Sauer and Besag on the Generalized Gauss-Markov Models [9], and L_1–Markov models [10]. Finally, with $q = 1$ and $\boldsymbol{M}$ an identity matrix it leads to a L_1–deconvolution algorithm in the context of seismic deconvolution [11].

According to the scale transformation $x \mapsto kx$ and $y \mapsto ky$, the models change in the following way:

$$p_b(k\boldsymbol{y} - Ak\boldsymbol{x}; \psi) \propto \exp\left[k^p \psi \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_p\right] \tag{14}$$

and

$$p_x(k\boldsymbol{x}; \lambda) \propto \exp\left[k^q \lambda \|\boldsymbol{M}\boldsymbol{x}\|_q\right]. \tag{15}$$

If we set $(\psi_k, \lambda_k) = (k^p\psi, k^q\lambda)$, the two estimates are scale invariant. Moreover, if $p = q$, we can drop the scale k in the MAP criteria (*eq. 11*) which becomes scale invariant. This is done in [9] [11], but it makes the choice of the prior and the noise models mutually dependent. We can also remark that ψ^q/λ^p is scale invariant and can be interpreted as a generalized SNR.

Maximum Entropy methods: Maximum Entropy reconstruction methods have been extensively used in the last decade. A helpful property of these methods is positivity of the restored image. In these methods, the noise is considered zero-mean Gaussian $\mathcal{N}(0, \Sigma_b)$, while the Log-prior take different forms which look like an "Entropy measure" of Burg or Shannon. Three different forms which have been used in practical problems are considered below.

- First, in a Fourier synthesis problem, Wernecke & D'Addario [12] used the following form:

$$p_x(\boldsymbol{x}; \lambda) \propto \exp\left[-\lambda \sum_i \log x_i\right]. \tag{16}$$

  Changing the scale in this context just modifies the partition function which is not important in the MAP criterion (*eq. 11*). As the noise is considered Gaussian, these authors show that if we update the λ parameter in a proper way ($\lambda_k = k^2\lambda$), then the ME reconstruction maintain linearity with respect to the measurement scale k. Thus, this ME solution is scale invariant, although nonlinear.

- In image restoration, Burch & *al.* [13] , consider a prior law of the form

$$p_x(\boldsymbol{x}; \lambda) \propto \exp\left[-\lambda \sum_i x_i \log x_i\right]. \tag{17}$$

Applying our scale changing yields:

$$p_x(k\boldsymbol{x};\lambda) \propto \exp\left[-\lambda \sum_i k\, x_i \log x_i + k \log k \sum_i x_i\right], \tag{18}$$

which does not satisfy the scale invariance property due to the $k \log k \sum_i x_i$ term. It appears from their later papers that they introduced a data pre-scaling before the reconstruction. Then, the modified version of their entropy becomes

$$p_x(\boldsymbol{x};\lambda) \propto \exp\left[-\lambda \sum_i \frac{x_i}{s} \log\left(\frac{x_i}{s}\right)\right], \tag{19}$$

where s is the pre-scaling parameter.

- Modification of the above expression with natural parameters for exponential family leads to the "entropic laws" used later by Gull & Skilling. [14] and Djafari [15]:

$$p_x(\boldsymbol{x};\lambda) \propto \exp\left[-\lambda_1 \sum_i x_i \log x_i - \lambda_2 \sum_i x_i\right]. \tag{20}$$

The resulting estimator is scale invariant for the reasons stated above.

Markovian models: A new Markovian model [16] has appeared from I-divergence considerations on small translation of an image in the context of astronomic deconvolution. This model can be rewritten as Gibbs distribution in the following form:

$$p_x(\boldsymbol{x};\lambda) \propto \exp\left[-\lambda \sum_{(s,r)\in\mathcal{C}} (x_s - x_r) \log\left(\frac{x_s}{x_r}\right)\right]. \tag{21}$$

If we change the scale of the measurement, the scale factor k vanishes in the logarithm, and

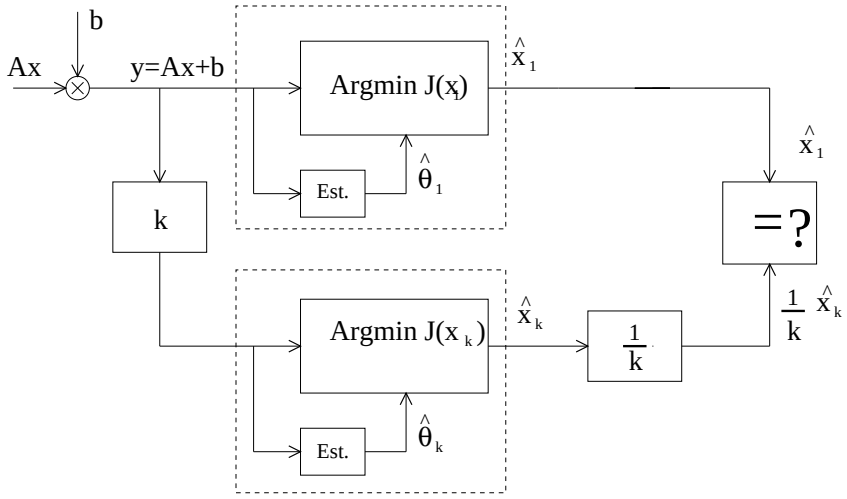
$$p_x(k\boldsymbol{x};\lambda) \propto \exp\left[-k\lambda \sum_{(s,r)\in\mathcal{C}} (x_s - x_r) \log\left(\frac{x_s}{x_r}\right)\right]. \tag{22}$$

Thus this particular Markov random field leads to a scale invariant estimator if we update the parameter λ according to $\lambda\sigma_b$ constant (the noise is assumed Gaussian-independent). In the same way as in the L_p norm example, $\lambda\sigma_b$ can be considered as a generalized SNR.

These examples show that the family of scale invariant laws is not a duck-billed platypus family. It includes many already employed priors on the context of image estimation. We have shown in a related work that other scale invariant prior laws exist, both in the Markovian prior family [17] and in the uncorrelated prior [2] family.

3. Scale invariant Bayesian estimator

Before further developing the scale invariance constraint for the estimator, we want to emphasize the role of the hyperparameters $\boldsymbol{\theta}$ (*i.e.*, parameters of the prior laws) and to sketch their estimation from the data which is very important in real-world applications. The estimation problem is considered globally. By globally we mean that, although we are interested on the estimation of $\boldsymbol{x}$ we want also to take into account the estimation of the



Scheme 1: Global scale invariance property for an estimator

hyperparameters $\boldsymbol{\theta}$. To summarize the SIP of an estimator, we illustrate it by the following scheme:

For more detail, let us define a scale invariant estimator in the following way:

Definition 1 *An estimator* $\widehat{\boldsymbol{x}}(\boldsymbol{y};\boldsymbol{\theta})$ *is said to be scale invariant if there exists function* $\boldsymbol{\theta}_k = \boldsymbol{f}_k(\boldsymbol{\theta})$ *such that*

$$\forall(\boldsymbol{y},\boldsymbol{\theta},k>0), \quad \widehat{\boldsymbol{x}}(k\boldsymbol{y},\boldsymbol{\theta}_k) = k\,\widehat{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{\theta}) \tag{23}$$

or in short

$$\boldsymbol{y} \mapsto \widehat{\boldsymbol{x}} \Longrightarrow \forall k>0, \quad k\boldsymbol{y} \mapsto k\widehat{\boldsymbol{x}}. \tag{24}$$

In this paper, we focus only on priors which admit density laws. We define then the scale invariant property for those laws as follows:

Definition 2 *A probability density function* $p_u(\boldsymbol{u};\boldsymbol{\theta})$ *[resp., a conditional density* $p_{u|v}(\boldsymbol{u}|\boldsymbol{v};\boldsymbol{\theta})$*,] is said to be scale invariant if there exists function* $\boldsymbol{\theta}_k = \boldsymbol{f}_k(\boldsymbol{\theta})$ *such that*

$$\forall(\boldsymbol{u},\boldsymbol{\theta},k>0), \quad p_u(k\boldsymbol{u};\boldsymbol{\theta}_k) = k^{-N} p_u(\boldsymbol{u};\boldsymbol{\theta}), \tag{25}$$

[resp., $\forall(\boldsymbol{u},\boldsymbol{\theta},k>0), \quad p_{u|v}(k\boldsymbol{u}|k\boldsymbol{v};\boldsymbol{\theta}_k) = k^{-N} p_{u|v}(\boldsymbol{u}|\boldsymbol{v};\boldsymbol{\theta}),$*]*
where $N = dim(\boldsymbol{u})$.
If $\boldsymbol{f}_k = Id$, i.e. *; if* $\boldsymbol{\theta}_k = \boldsymbol{\theta}$ *then* $p_u(\boldsymbol{u};\boldsymbol{\theta})$ *is said to be strictly scale invariant.*

The above property for density laws specifies that these laws are a part of a family of the laws which is closed relative to scale transformation. Thus, in this class, a set of pertinent parameters exists for each chosen scale.

We need also to set two properties for scale invariant density laws. Both concern the conservation of the SIP, one after marginalization, the other after application of the Bayes rules.

Lemma 1 *If $p_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ is scale invariant, then the marginalized $p_y(\boldsymbol{y}; \boldsymbol{\theta})$ is also scale invariant.*

Lemma 2 *If $p_x(\boldsymbol{x}; \boldsymbol{\lambda})$ and $p_{y|x}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\psi})$ are scale invariant, then the joint law $p_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\psi})$ is also scale invariant.*

Proofs are straightforward and are found in Appendix A.

Using these two definitions, we prove the following theorem which summarizes sufficient conditions for an estimator to be scale invariant:

Theorem 1 *If the cost function $C(\boldsymbol{x}^*, \boldsymbol{x})$ of a Bayesian estimator satisfies the condition:*

$$\forall k > 0, \exists (a_k \in I\!R, b_k > 0) \text{ such that } \forall (\boldsymbol{x}^*, \boldsymbol{x}), \quad C(\boldsymbol{x}_k^*, \boldsymbol{x}_k) = a_k + b_k C(\boldsymbol{x}^*, \boldsymbol{x}), \tag{26}$$

and if the posterior law is scale invariant, i.e.*, there exists function $\boldsymbol{\theta}_k = \boldsymbol{f}_k(\boldsymbol{\theta})$ such that:*

$$\forall k > 0, \forall (\boldsymbol{x}, \boldsymbol{y}), \quad p(k\boldsymbol{x}|k\boldsymbol{y}; \boldsymbol{\theta}_k) = k^{-dim(\boldsymbol{x})} p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}), \tag{27}$$

then, the resulting Bayesian estimator is scale invariant, i.e.,

$$\widehat{\boldsymbol{x}}(k\boldsymbol{y}, \boldsymbol{\theta}_k) = k\, \widehat{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{\theta}). \tag{28}$$

See the appendix B for the proof. It is also shown there that the cost functions of the three classical Bayesian estimators, *i.e.*; MAP, PM and the MMAP, satisfy the first constraint.

Remark: In this theorem, the SIP is applied to the posterior law $p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta})$. However, we can separate the hyperparameters $\boldsymbol{\theta}$ in two sets $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$, where $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$ are the parameters of the prior laws $p_x(\boldsymbol{x}; \boldsymbol{\lambda})$ and $p_b(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}; \boldsymbol{\psi})$. In what follows, we want to make the choice of p_x and p_b independent. From the lemma 1 and 2, if p_x and p_b satisfy the SIP then the posterior $p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta})$ satisfies the SIP. As a consequence $\boldsymbol{\theta}_k$ must be separated according to $\boldsymbol{\theta}_k = [\boldsymbol{\lambda}_k, \boldsymbol{\psi}_k] = [\boldsymbol{g}_k(\boldsymbol{\lambda}), \boldsymbol{h}_k(\boldsymbol{\psi})]$.

4. Hyperparameters estimation

In the above theorem, we assumed that the hyperparameters $\boldsymbol{\theta}$ are given. Thus, given the data $\boldsymbol{y}$ and the hyperparameters $\boldsymbol{\theta}$, we can calculate $\widehat{\boldsymbol{x}}$. Now, if the scale factor k of the data has been changed, we have first to update the hyperparameters [18] according to $\boldsymbol{\theta}_k = \boldsymbol{f}_k(\boldsymbol{\theta})$, and then we can use the SIP:

$$\widehat{\boldsymbol{x}}(k\boldsymbol{y}, \boldsymbol{\theta}_k) = k\, \widehat{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{\theta}). \tag{29}$$

Now, let us see what happens if we have to estimate both $\boldsymbol{x}$ and $\boldsymbol{\theta}$, either by Maximum or Generalized Maximum Likelihood.

- Maximum likelihood (ML) method estimates first $\boldsymbol{\theta}$ by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{L(\boldsymbol{\theta})\}, \tag{30}$$

where

$$L(\boldsymbol{\theta}) = p(\boldsymbol{y};\boldsymbol{\theta}) \tag{31}$$

and then $\widehat{\boldsymbol{\theta}}$ is used to estimate $\boldsymbol{x}$. At a scale k,

$$\widehat{\boldsymbol{\theta}}_k = \arg\max_{\boldsymbol{\theta}_k} \{L_k(\boldsymbol{\theta}_k)\}. \tag{32}$$

Application of lemma 1 implies that

$$L_k(\boldsymbol{\theta}_k) = k^{\dim(\boldsymbol{y})} L(\boldsymbol{\theta}), \tag{33}$$

thus, the Maximum Likelihood estimator satisfies the condition

$$\widehat{\boldsymbol{\theta}}_k = \boldsymbol{f}_k(\widehat{\boldsymbol{\theta}}). \tag{34}$$

The likelihood function (*eq.* 31) has rarely an explicit form, and a common algorithm for its locally maximization is the EM algorithm which is an iterative algorithm described briefly as follows:

$$\left\{ \begin{array}{rcl} Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(i)}) & = & \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}^{(i)}} \{\ln p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\} \\ \widehat{\boldsymbol{\theta}}^{(i+1)} & = & \arg\max_{\boldsymbol{\theta}} \left\{ Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(i)}) \right\}. \end{array} \right. \tag{35}$$

At a scale k,

$$\begin{array}{rcl} Q_k(\boldsymbol{\theta}_k;\widehat{\boldsymbol{\theta}}_k^{(i)}) & = & \mathrm{E}_{k\boldsymbol{x}|k\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k^{(i)}} \{\ln p(k\boldsymbol{y}|k\boldsymbol{x};\boldsymbol{\theta}_k)\} \\ & = & -M\ln k + \mathrm{E}_{k\boldsymbol{x}|k\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k^{(i)}} \{\ln p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\} \\ & = & -M\ln k + k^{-\dim(\boldsymbol{y})} Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(i)}). \end{array} \tag{36}$$

Thus, if we initialize this iterative algorithm with the value $\widehat{\boldsymbol{\theta}}_k^{(0)} = f_k(\widehat{\boldsymbol{\theta}}^{(0)})$, then we have

$$\widehat{\boldsymbol{\theta}}_k^{(1\cdots l)} = \boldsymbol{f}_k\left(\widehat{\boldsymbol{\theta}}^{(1\cdots l)}\right). \tag{37}$$

Then the scale invariance coherence of hyperparameters is ensured during the optimization steps.

- In Generalized Maximum Likelihood (GML) method, one estimates both $\boldsymbol{\theta}$ and $\boldsymbol{x}$ by

$$\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{x}}\right) = \arg\max_{(\boldsymbol{\theta},\boldsymbol{x})} \{p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})\}. \tag{38}$$

Applying the same demonstration as above to the joint laws rather than to the marginalized one leads to

$$\left(\widehat{\boldsymbol{\theta}}_k,\widehat{\boldsymbol{x}}_k\right) = \left(\boldsymbol{f}_k(\widehat{\boldsymbol{\theta}}), k\widehat{\boldsymbol{x}}\right). \tag{39}$$

However, this holds if and only if the GML has a maximum. This may not be always the case and this is a major drawback in GML. Also, in GML method, direct resolution

is rarely possible and sub-optimal techniques lead to the classical two-step estimation scheme:

$$\widehat{\boldsymbol{x}}^{(i)} = \arg\max_{\boldsymbol{x}} \left\{ p(\boldsymbol{x}, \boldsymbol{y}; \widehat{\boldsymbol{\theta}}^{(i)}) \right\}, \tag{40}$$

$$\widehat{\boldsymbol{\theta}}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\widehat{\boldsymbol{x}}^{(i)}, \boldsymbol{y}; \boldsymbol{\theta}) \right\}. \tag{41}$$

We see that, in each iteration, the $\boldsymbol{\theta}$ estimation step may be considered as the ML estimation of $\boldsymbol{\theta}$ if $\boldsymbol{x}^{(i)}$ is supposed to be a realization of the prior law. Thus the coherence of estimated hyperparameters at different scales is fulfilled during the both optimization steps, and

$$\left(\widehat{\boldsymbol{\theta}}_k^{(1\cdots l)}, \widehat{\boldsymbol{x}}_k^{(1\cdots l)} \right) = \left(\boldsymbol{f}_k(\widehat{\boldsymbol{\theta}}^{(1\cdots l)}), k\widehat{\boldsymbol{x}}^{(1\cdots l)} \right). \tag{42}$$

Thus, if we consider the whole estimation problem (with a ML or GML approach), the SIP of the estimator is assured in both cases. It is also ensured during the iterative optimization schemes of ML or GML.

5. Markovian invariant distributions

Markovian distributions as priors in image processing allow to introduce local characteristics and inter-pixels correlations. They are widely used but there exist many different Markovian models and very few model selection guidelines exist. In this section we apply the above scale invariance considerations to the prior model selection in the case of first order homogeneous MRFs.

Let $X \in \Omega$ be a homogeneous Markov random field defined on the subset $[1 \dots N] \times [1 \dots M]$ of $\mathbf{Z}^2$. The Markov characteristic property is:

$$p_X(x_i | x_{\mathcal{S}-i}) = p_X(x_i | x_{\delta i}), \tag{43}$$

where δi is the neighbourhood of site i, and $\mathcal{S}$ is the set of pixels. Hammersley-Clifford theorem for the first order neighbourhood reads:

$$p_X(\boldsymbol{x}; \lambda) \propto \exp\left(-\lambda \sum_{\{r,s\} \in \mathcal{C}} \phi(x_s, x_r) \right), \tag{44}$$

where $\mathcal{C}$ is the clique set, and $\phi(x, y)$ the clique potential. In most works [9, 19, 20, 21] a simplified model is introduced under the form $\phi(x, y) = \phi(x - y)$. Here we keep a general point of view. Application of the scale invariance condition to the Markovian prior laws $p_X(\boldsymbol{x}, \lambda)$ leads to the two following theorems:

Theorem 2 *A familly of Markovian distribution is scale invariant if and only if there exist two functions $f(k, \lambda)$ and $\beta(k)$ such that clique potential $\phi(x_s, x_r)$ satisfies:*

$$f(k, \lambda)\ \phi(kx_s, kx_r) = \lambda \phi(x_s, x_r) + \beta(k). \tag{45}$$

Theorem 3 *A necessary and sufficient condition for a Markov random fields to be scale invariant is that exists a triplet (a, b, c) such as the clique potential $\phi(x_s, x_r)$ verifies the linear partial differential equation (PDE) :*

$$a\phi(x_s, x_r) + b\left(x_s \frac{\partial \phi(x_s, x_r)}{\partial x_s} + x_r \frac{\partial \phi(x_s, x_r)}{\partial x_r}\right) = c.$$

Finally, enforcing symmetry of the clique potentials $\phi(x_s, x_r) = \phi(x_r, x_s)$ the following theorem provides the set of scale invariant clique potentials:

Theorem 4 *$p_X(\boldsymbol{x}, \lambda)$ is scale invariant if and only if $\phi(x_s, x_r)$ is chosen from one of the following vector spaces:*

$$\mathcal{V}_0 = \left\{\phi(x_s, x_r) \mid \exists \psi(.) \text{ even and } p \in I\!R, \phi(x_s, x_r) = \psi\left(\log\left|\frac{x_s}{x_r}\right|\right) - p\log|x_s x_r|\right\} \quad (46)$$

$$\mathcal{V}_1(p) = \left\{\phi(x_s, x_r) \mid \exists \psi(.) \text{ even }, \phi(x_s, x_r) = \psi\left(\log\left|\frac{x_s}{x_r}\right|\right)|x_s x_r|^p\right\} \quad (47)$$

Moreover, $\mathcal{V}_0$ is the subspace of strictly scale invariant clique potentials.

For the proof of these theorems see [22].

Among the most common models in use for image processing purposes, only few clique potentials fall into the above set. Let us give two examples:

First, the GGMRFs proposed by Bouman & Sauer [9] were built by a similar approach of scale invariance but under the restricted assumption that $\phi(x_s, x_r) = \phi(x_s - x_r)$. The yielded expression $\phi(x_s, x_r) = |x_s - x_r|^p$ can be factored according to $\phi(x_s, x_r) = |x_s\, x_r|^{p/2}|2\mathrm{sh}\left(\log(x_s/x_r)/2\right)|^p$ which shows that it falls in $\mathcal{V}_1(p)$.

The second example of potential does not reduce to the single variable function $\phi(x_s - x_r)$: $\phi(x_s, x_r) = (x_s - x_r)\log(x_s/x_r)$. It has recently been introduced from I-divergence penalty considerations in the field of image estimation problem (optic deconvolution) by O'Sullivan [16]. Factoring $|x_s x_r|^{\frac{1}{2}}$ leads to:

$$\phi(x_s, x_r) = |x_s x_r|^{\frac{1}{2}}\psi\left(\log(x_s/x_r)\right), \quad (48)$$

where $\psi(X) = 2X\mathrm{sh}(X/2)$ is even. It shows that $\phi(x_s, x_r)$ is in $\mathcal{V}_1(1/2)$ and is scale invariant. As $\phi(x_s, x_r)$ is defined only on $\mathbf{R}^2_{*+}$ it applies to positive quantities. This feature is very useful in image processing where prior positivity applies to many physical quantities.

6. Conclusions

In this paper we have outlined and justified a weaker property than linearity that is desired for the Bayesian estimators to have. We have shown that this scale invariance property (SIP) helps to avoid an arbitrary choice for the scale of the measurement. Some models already employed in Bayesian estimation, including Markov prior Models [9, 16], Entropic prior [23, 2] and Generalized Gaussian models [11], have demonstrated the existence and usefulness of scale invariant models. Then we have given general conditions for a Bayesian estimator to be scale invariant. This property holds for most Bayesian estimators such as MAP, PM, MMAP under the condition that the prior laws are also scale invariant. Thus,

imposition of the SIP can assist in the model selection. We have also shown that classical hyperparameters estimation methods satisfy the SIP property for estimated laws.

Finally we discussed how to choose the prior laws to obtain scale invariant Bayesian estimators. For this, we considered two cases: *entropic prior laws* and *first-order Markov models*. In related preceding works [1, 2, 24], the SIP constraints have been studied for the case of entropic prior laws. In this paper we extended that work to the case of first-order Markov models and showed that many common Markov models used in image processing are special cases.

A SIP property inheritance

- **Proof of the Lemma 1:**

  Let $p_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ have the scale invariance property, then if there exists $\boldsymbol{\theta}_k = \boldsymbol{f}_k(\boldsymbol{\theta})$ such that

  $$p_{x,y}(k\boldsymbol{x}, k\boldsymbol{y}; \boldsymbol{\theta}_k) = k^{-(M+N)} p_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}),$$

  where $N = \dim(\boldsymbol{x})$ and $M = \dim(\boldsymbol{y})$, then, marginalizing with respect to $\boldsymbol{x}$, we obtain

  $$p_y(k\boldsymbol{y}; \boldsymbol{\theta}_k) = k^{-(M+N)} \int p_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) k^{-N} d\boldsymbol{x} = k^{-M} p_y(\boldsymbol{y}; \boldsymbol{\theta}),$$

  which completes the proof.

- **Proof of the Lemma 2:**

  The definition of SIP for density laws and direct application of the Bayes rule lead to

  $$p_{x,y}(k\boldsymbol{x}, k\boldsymbol{y}; \boldsymbol{\theta}_k) = k^{-N} p_x(\boldsymbol{x}; \boldsymbol{\lambda})\, k^{-M} p_{y|x}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\psi}) = k^{-(M+N)} p_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}),$$

  which concludes the proof.

B SIP conditions for Bayesian estimator

- **Proof of the Theorem 1:**

  Since a Bayesian estimator is defined by

  $$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\{ \int C(\boldsymbol{x}^*, \boldsymbol{x})\, p(\boldsymbol{x}^*|\boldsymbol{y}; \boldsymbol{\theta})\, d\boldsymbol{x}^* \right\},$$

  then

  $$\begin{aligned}
  \widehat{\boldsymbol{x}}_k &= \arg\min_{\boldsymbol{x}_k} \left\{ \int C(\boldsymbol{x}_k^*, \boldsymbol{x}_k)\, p(\boldsymbol{x}_k^*|k\boldsymbol{y}; \boldsymbol{\theta}_k) d(\boldsymbol{x}_k^*) \right\} \\
  &= k \arg\min_{\boldsymbol{x}} \left\{ \int C(k\boldsymbol{x}^*, k\boldsymbol{x})\, p(k\boldsymbol{x}^*|k\boldsymbol{y}; \boldsymbol{\theta}_k)\, k^N \, \mathrm{d}\boldsymbol{x}^* \right\} \\
  &= k \arg\min_{\boldsymbol{x}} \left\{ \int [a_k + b_k C(\boldsymbol{x}^*, \boldsymbol{x})] k^{-N}\, p(\boldsymbol{x}^*|\boldsymbol{y}; \boldsymbol{\theta})\, k^N \, \mathrm{d}\boldsymbol{x}^* \right\} = k\, \widehat{\boldsymbol{x}},
  \end{aligned}$$

  which proves the Theorem 1.

- **Conditions for cost functions**:

  The three classical Bayesian estimators, MAP, PM and MMAP, satisfy the condition of the cost function:

  - *Posterior mean* (PM): $C(\boldsymbol{x}_k^*, \boldsymbol{x}_k) = (\boldsymbol{x}_k^* - \boldsymbol{x}_k)^t \boldsymbol{Q} (\boldsymbol{x}_k^* - \boldsymbol{x}_k) = k^2\, C(\boldsymbol{x}^*, \boldsymbol{x})$.
  - *Maximum a posteriori* (MAP): $C(\boldsymbol{x}_k^*, \boldsymbol{x}_k) = 1 - \delta(\boldsymbol{x}_k^* - \boldsymbol{x}_k) = C(\boldsymbol{x}^*, \boldsymbol{x})$.
  - *Marginal Maximum a Posteriori* (MMAP):

$$C(\boldsymbol{x}_k^*, \boldsymbol{x}_k) = \sum_i (1 - \delta([\boldsymbol{x}_k^*]_i - [\boldsymbol{x}_k]_i)) = C(\boldsymbol{x}^*, \boldsymbol{x}).$$

References

[1] A. Mohammad-Djafari and J. Idier, "Maximum entropy prior laws of images and estimation of their parameters," in *Maximum Entropy and Bayesian Methods in Science and Engineering* (T. Grandy, ed.), (Dordrecht, The Netherlands), MaxEnt Workshops, Kluwer Academic Publishers, 1990.

[2] A. Mohammad-Djafari and J. Idier, "Scale invariant Bayesian estimators for linear inverse problems," in *Proc. of the First ISBA meeting*, (San Fransisco, USA), Aug. 1993.

[3] G. Demoment, "Image reconstruction and restoration : Overview of common estimation structure and problems," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 37, pp. 2024–2036, Dec. 1989.

[4] A. Mohammad-Djafari and G. Demoment, "Estimating priors in maximum entropy image processing," in *Proceedings of IEEE ICASSP*, pp. 2069–2072, IEEE, 1990.

[5] G. Le Besnerais, J. Navaza, and G. Demoment, "Aperture synthesis in astronomical radio-interferometry using maximum entropy on the mean," in *SPIE Conf., Stochastic and Neural Methods in Signal Processing, Image Processing and Computer Vision* (S. Chen, ed.), (San Diego), p. 11, July 1991.

[6] G. Le Besnerais, J. Navaza, and G. Demoment, "Synthèse d'ouverture en radio-astronomie par maximum d'entropie sur la moyenne," in *Actes du 13ème colloque GRETSI*, (Juan-les-Pins, France), pp. 217–220, Sept. 1991.

[7] E. Jaynes, "Prior probabilities," *IEEE Transactions on Systems Science and Cybernetics*, vol. SSC-4, pp. 227–241, Sept. 1968.

[8] G. Box and T. G.C., *Bayesian inference in statistical analysis.* Addison-Wesley publishing, 1972.

[9] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Transactions on Medical Imaging*, vol. MI-2, no. 3, pp. 296–310, 1993.

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

[11] D. Oldenburg, S. Levy, and K. Stinson, "Inversion of band-limited reflection seismograms: theory and practise," *Procedings of IEEE*, vol. 74, p. 3, 1986.

[12] S. Wernecke and L. D'Addario, "Maximum entropy image reconstruction," *IEEE Transactions on Computers*, vol. C-26, pp. 351–364, Apr. 1977.

[13] S. Burch, S. Gull, and J. Skilling, "Image restoration by a powerful maximum entropy method," *Computer Vision and Graphics and Image Processing*, vol. 23, pp. 113–128, 1983.

[14] S. Gull and J. Skilling, "Maximum entropy method in image processing," *Proceedings of the IEE*, vol. 131-F, pp. 646–659, 1984.

[15] A. Mohammad-Djafari and G. Demoment, "Maximum entropy reconstruction in X ray and diffraction tomography," *IEEE Transactions on Medical Imaging*, vol. 7, no. 4, pp. 345–354, 1988.

[16] J. A. O'Sullivan, "Divergence penalty for image regularization," in *Proceedings of IEEE ICASSP*, vol. V, (Adelaide), pp. 541–544, Apr. 1994.

[17] S. Brette, J. Idier, and A. Mohammad-Djafari, "Scale invariant Markov models for linear inverse problems," in *Fifth Valencia Int. Meeting on Bayesian Statistics*, (Alicante, Spain), June 1994.

[18] 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.

[19] 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, p. 2, 1984.

[20] S. Geman and G. Reynolds, "Constrained restoration and recovery of discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-14, pp. 367–383, 1992.

[21] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society B*, vol. 48, p. 1, 1986.

[22] S. Brette, J. Idier, and A. Mohammad-Djafari, "Scale invariant Markov models for linear inverse problems," in *Second ISBA meeting*, vol. Bayesian Statistics, (Alicante, Spain), ISBA, American Statistical Association, June 1994.

[23] S. F. Gull, "Developments in maximum entropy data analysis," in *Maximum Entropy and Bayesian Methods* (J. Skilling, ed.), pp. 53–71, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1989.

[24] A. Mohammad-Djafari and J. Idier, "A scale invariant Bayesian method to solve linear inverse problems," in *Maximum Entropy and Bayesian Methods* (G. Heidbreder, ed.), (Dordrecht, The Netherlands), The 13th Int. MaxEnt Workshops, Santa Barbara, USA, Kluwer Academic Publishers, 1993.