MAXIMUM-A-POSTERIORI ESTIMATION WITH UNKNOWN REGULARISATION PARAMETERS

Marcelo Pereyra *

School of Mathematics
University of Bristol
Bristol, United Kingdom

José M. Bioucas-Dias, Mário A. T. Figueiredo†

Instituto de Telecomunicações, and
Instituto Superior Técnico,
Universidade de Lisboa, Portugal

ABSTRACT

This paper presents two hierarchical Bayesian methods for performing maximum-a-posteriori inference when the value of the regularisation parameter is unknown. The methods are useful for models with homogenous regularisers (i.e., prior sufficient statistics), including all norms, composite norms and compositions of norms with linear operators. A key contribution of this paper is to show that for these models the normalisation factor of the prior has a closed-form analytic expression. This then enables the development of Bayesian inference techniques to either estimate regularisation parameters from the observed data or, alternatively, to remove them from the model by marginalisation followed by inference with the marginalised model. The effectiveness of the proposed methodologies is illustrated on applications to compressive sensing using an ℓ_1-wavelet analysis prior, where they outperform a state-of-the-art SURE-based technique, both in terms of estimation accuracy and computing time.

Index Terms— regularisation parameters; maximum-a-posteriori estimation; hierarchical Bayesian inference; inverse problems; statistical signal processing.

1. INTRODUCTION

Bayesian inference methods have become ubiquitous in modern signal processing, machine learning, and computer vision. In particular, *maximum-a-posteriori* (MAP) estimation has been adopted as a standard approach for solving many high-dimensional inverse problems, mostly because MAP estimates can often be computed efficiently by optimisation. Indeed, the development of new Bayesian models and optimisation algorithms for high-dimensional inference has received a lot of attention in the late, leading to significant improvements in estimation accuracy and computing time [1].

An important aspect of MAP estimation is the selection of the so-called "regularisation parameters" (see [2–4] for an excellent introduction to this topic and [5] for a recent comprehensive literature review in the image restoration context). These hyper-parameters specify the prior distribution defining the Bayesian model and impact strongly MAP estimation results. Unfortunately, setting appropriate values for these parameters is generally difficult. Calibrating them a priori is challenging because values that perform well on some training data often generalise poorly to new data. Developing Bayesian inference techniques to adjust their value automatically is also difficult, as it requires evaluating a normalisation factor, which, for many relevant models, is computationally intractable [6].

This paper presents two hierarchical Bayesian methods for performing MAP inference when the value of the regularisation parameter is unknown. The paper is organised as follows: Section 2 highlights the difficulties associated with selecting regularisation parameters and gives a brief background on Bayesian and non-Bayesian techniques. Section 3 describes the proposed methodologies, which allow either estimating the value of the regularisation parameter directly from the observed data, or, alternatively, removing it from the model by marginalisation. An additional advantage of these methods is that they can be integrated straightforwardly to existing MAP optimisation algorithms, in which the regularisation parameter is assumed to be known and fixed. In Section 4, the two methodologies are illustrated on an application to compressive sensing of images and compared to a state-of-the-art approach. Conclusions and perspectives are finally reported in Section 5.

2. PROBLEM STATEMENT

Let $\boldsymbol{x} \in \mathbb{R}^n$ be an unknown signal of interest and $\boldsymbol{y}$ an observation related to $\boldsymbol{x}$ by a statistical model with likelihood function $p(\boldsymbol{y}|\boldsymbol{x}) = \exp\{-g_{\boldsymbol{y}}(\boldsymbol{x})\}$. Suppose that the recovery of $\boldsymbol{x}$ from $\boldsymbol{y}$ is ill-posed or ill-conditioned. Following a Bayesian approach, we address this difficulty by modelling $\boldsymbol{x}$ as a random vector with prior distribution

*MP holds a Marie Curie Intra-European Fellowship for Career Development. He also acknowledges support from the "SuSTaIn" program at the University of Bristol (EPSRC grant EP/D063485/1).

†This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), under grants UID/EEA/50008/2013 and PTDC/EEIPRO/ 1470/2012.

$p(\boldsymbol{x}|\lambda) = \exp\{-\lambda h(\boldsymbol{x})\}/C(\lambda)$ promoting solutions with some application-specific structural or regularity properties (which are encoded in the sufficient statistic $h(\boldsymbol{x})$). This prior is parametrised by a "regularisation" (hyper-) parameter $\lambda \in \mathbb{R}^+$ that controls the relative weight of $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x}|\lambda)$, and therefore the balance between observed and prior information. Once $p(\boldsymbol{x}, \boldsymbol{y}|\lambda) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\lambda)$ is properly specified, $\boldsymbol{x}$ can be estimated, for example, by computing the MAP estimator

$$\hat{\boldsymbol{x}}_\lambda = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\, g_{\boldsymbol{y}}(\boldsymbol{x}) + \lambda h(\boldsymbol{x}), \tag{1}$$

which we assume to be computationally tractable and unique (conditionally to a given value of λ).

In this paper, we assume that the value of λ is unknown, making the estimation of $\boldsymbol{x}$ from $\boldsymbol{y}$ even more challenging. We note that λ can impact MAP estimates significantly because the problem is not well-posed. The Bayesian framework provides a number of principled strategies for estimating $\boldsymbol{x}$ without specifying the value of λ [6]. However, to use them it is necessary to know the normalisation factor of $p(\boldsymbol{x}|\lambda)$, which is given by

$$C(\lambda) = \int_{\mathbb{R}^n} \exp\{-\lambda h(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x}. \tag{2}$$

For most models of interest, the evaluation of $C(\lambda)$ is a reputedly intractable problem due to the integration over $\mathbb{R}^n$ [7].

Bayesian methods typically address this difficulty by replacing $C(\lambda)$ with an approximation; for example pseudo-likelihood approximations [8], variational approximations [9], and Monte Carlo approximations [10]. Of course, the performance of these methods depends on the approximation accuracy, which is generally difficult to assess. Non-Bayesian methods have traditionally set regularisation parameters by generalised cross-validation by using discrepancy criterions (see [2, 3] for a detailed analysis and connections to Bayesian methods when h is quadratic). In particular, a very promising approach that has regained attention lately is to set λ by minimising a surrogate of the mean-square-error of $\hat{\boldsymbol{x}}_\lambda$ (typically a *Stein's unbiased risk estimate* (SURE) [11–13]). In this case, the performance depends mainly on the accuracy of the surrogate, which may also be difficult to assess a priori.

3. PROPOSED BAYESIAN METHODS

This section presents two inference methods for estimating $\boldsymbol{x}$ when the value of λ is unknown. More precisely, we propose two hierarchical Bayesian techniques that allow either estimating λ from $\boldsymbol{y}$ jointly with $\boldsymbol{x}$, or, alternatively, removing λ from the model by marginalisation and then estimating $\boldsymbol{x}$ with the marginalised model. A remarkable property of the techniques is that they are "exact", in the sense that they use the correct normalisation factor $C(\lambda)$ without any approximation error. Also, they can be easily integrated into existing Bayesian algorithms that assume that λ is known. The proposed techniques are useful when the regulariser h (i.e., the sufficient statistic of $p(\boldsymbol{x}|\lambda)$) is a k-homogenous function. This scenario arises in many applications, and comprises all norms and pseudo-norms.

3.1. Priors with k-homogenous sufficient statistics

We say that the regulariser h in (1) is a k-homogeneous function if there exists $k \in \mathbb{R}^+$ such that

$$h(\eta\boldsymbol{x}) = \eta^k h(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \forall \eta > 0. \tag{3}$$

Notice that (3) holds for most models used in modern signal image processing. In particular, all norms (e.g., ℓ_1, ℓ_2, total-variation, nuclear, etc.), composite norms (e.g., $\ell_1 - \ell_2$), and compositions of norms with linear operators (e.g., analysis terms of the form $\|\Psi\boldsymbol{x}\|_1$) are 1-homogenous. Similarly, powers of norms with exponent q are q-homogenous. Property (3) also holds for all models belonging to the general framework described in [14].

A central contribution of this paper is to show that if property (3) holds, then $C(\lambda)$ takes the following form:

Proposition 3.1 *If h, the sufficient statistic of $p(\boldsymbol{x}|\lambda)$, is k-homogenous, then, the normalisation factor has the form*

$$C(\lambda) = D\lambda^{-n/k},$$

where $D = C(1)$ is a constant independent of λ.

The proof follows straightforwardly by using the change of variables $\boldsymbol{u} = \lambda^{1/k}\boldsymbol{x}$ and (3) to rearrange (2) as a product of a function of λ and the constant $D = \int_{\mathbb{R}^n} \exp\{-h(\boldsymbol{u})\}\mathrm{d}\boldsymbol{u}$.

3.2. Hierarchical Bayesian inference

We are now ready to describe the proposed Bayesian inference strategies, which use Proposition 3.1 to estimate $\boldsymbol{x}$ when λ is unknown. Following a hierarchical Bayesian approach, we represent λ as an additional unknown quantity in our model and assign it the gamma hyper-prior (a natural choice, since λ plays the role of a scale parameter [15])

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp\{-\beta\lambda\}\mathbf{1}_{\mathbb{R}^+}(\lambda),$$

with fixed parameters α and β. When n is large, the exact values of α and β generally have little impact on the inferences; without loss of generality we use $\alpha = 1$ and $\beta = 1$ in our experiments.

3.2.1. Joint maximum-a-posteriori estimation

A natural extension of (1) to the case of unknown λ is to compute a joint MAP estimator. Suppose that $(\hat{\boldsymbol{x}}^*, \lambda^*) \in \mathbb{R}^{n+1}$ is a maximiser of $p(\boldsymbol{x}, \lambda|\boldsymbol{y})$, then

$$\mathbf{0}_{n+1} \in \partial_{\boldsymbol{x},\lambda} \log p(\hat{\boldsymbol{x}}^*, \lambda^*|\boldsymbol{y}),$$

where $\mathbf{0}_j$ denotes the j-dimensional null vector and $\partial_{\boldsymbol{s}} f(\boldsymbol{s}^*)$ the set of subgradients of a function $f(\boldsymbol{s})$ at a point $\boldsymbol{s}^*$ [16], which in turn implies that

$$\mathbf{0}_n \in \partial_{\boldsymbol{x}} \log p(\hat{\boldsymbol{x}}^*, \lambda^* | \boldsymbol{y}), \tag{4}$$

and that

$$0 \in \partial_{\lambda} \log p(\hat{\boldsymbol{x}}^*, \lambda^* | \boldsymbol{y}). \tag{5}$$

It follows from (4) that

$$\hat{\boldsymbol{x}}^* = \hat{\boldsymbol{x}}_{\lambda^*} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} g_{\boldsymbol{y}}(\boldsymbol{x}) + \lambda^* h(\boldsymbol{x}).$$

Also, by developing (5) and using Proposition 3.1, we obtain

$$\lambda^* = \frac{n/k + \alpha - 1}{h(\hat{\boldsymbol{x}}_{\lambda^*}) + \beta}. \tag{6}$$

The values λ^* satisfying (6) can be identified by solving a one-dimensional root-finding problem. We emphasise at this point that in all our experiments we observed that $p(\boldsymbol{x}, \lambda | \boldsymbol{y})$ is unimodal, with only one λ^* (and $\hat{\boldsymbol{x}}^*$) satisfying (6). A detailed theoretical analysis of the conditions for existence and uniqueness of λ^* and $\hat{\boldsymbol{x}}^*$ is beyond the scope of this paper.

Finally, to compute $\hat{\boldsymbol{x}}^*$ and λ^* we maximise $p(\boldsymbol{x}, \lambda | \boldsymbol{y})$ alternatively w.r.t. $\boldsymbol{x}$ and λ with the following scheme

$$\begin{aligned} \boldsymbol{x}^{(t)} &= \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} g_{\boldsymbol{y}}(\boldsymbol{x}) + \lambda^{(t-1)} h(\boldsymbol{x}), \\ \lambda^{(t)} &= \frac{n/k + \alpha - 1}{h(\boldsymbol{x}^{(t)}) + \beta}, \end{aligned} \tag{7}$$

which in our experiments converged within 5 to 10 iterations. Note that without guarantees of uniqueness, the solution of (7) could potentially depend on initialisation, though we have not observed this in practice.

3.2.2. Marginalisation

An alternative approach is to remove λ from the model by marginalisation and then estimate $\boldsymbol{x}$ by maximising the marginalised posterior. Precisely, we integrate the posterior $p(\boldsymbol{x}, \lambda | \boldsymbol{y})$ w.r.t. λ and compute the marginal MAP estimator; that is,

$$\begin{aligned} \hat{\boldsymbol{x}}^\dagger &= \operatorname*{argmax}_{\boldsymbol{x} \in \mathbb{R}^n} \int_0^\infty p(\boldsymbol{x}, \lambda | \boldsymbol{y}) \mathrm{d}\lambda, \\ &= \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} g_{\boldsymbol{y}}(\boldsymbol{x}) + (n/k + \alpha) \log\{h(\boldsymbol{x}) + \beta\}, \end{aligned} \tag{8}$$

which incorporates the uncertainty about λ in the inferences. To study (8), we construct the following majorant of $\log\{h(\boldsymbol{x}) + \beta\}$, based on the concavity of the logarithm:

$$\begin{aligned} q(\boldsymbol{x} | \boldsymbol{x}_0) &= \log\{h(\boldsymbol{x}_0) + \beta\} + \frac{h(\boldsymbol{x}) - h(\boldsymbol{x}_0)}{h(\boldsymbol{x}_0) + \beta} \\ &\geq \log\{h(\boldsymbol{x}) + \beta\}, \end{aligned} \tag{9}$$

such that $\hat{\boldsymbol{x}}^\dagger$ is also the unique minimiser of $g_{\boldsymbol{y}}(\boldsymbol{x}) + (\alpha + n/k) q(\boldsymbol{x} | \hat{\boldsymbol{x}}^\dagger)$, which is a majorant of (8). By developing this result we obtain that

$$\hat{\boldsymbol{x}}^\dagger = \hat{\boldsymbol{x}}_{\lambda^\dagger} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} g_{\boldsymbol{y}}(\boldsymbol{x}) + \lambda^\dagger h(\boldsymbol{x}),$$

with

$$\lambda^\dagger = \frac{n/k + \alpha}{h(\hat{\boldsymbol{x}}_{\lambda^\dagger}) + \beta}. \tag{10}$$

Again, the values $\lambda^\dagger$ satisfying (10) can be identified by solving a one-dimensional root-finding problem. However, in all our experiments we observed that $p(\boldsymbol{x} | \boldsymbol{y})$ is unimodal with only one $\lambda^\dagger$ and $\hat{\boldsymbol{x}}^\dagger$ satisfying (10). A detailed theoretical analysis of this property is beyond the scope of this paper.

Finally, in a manner akin to (7), we compute these values with the iterative scheme fixed-point scheme

$$\begin{aligned} \boldsymbol{x}^{(t)} &= \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} g_{\boldsymbol{y}}(\boldsymbol{x}) + \lambda^{(t-1)} h(\boldsymbol{x}), \\ \lambda^{(t)} &= \frac{n/k + \alpha}{h(\boldsymbol{x}^{(t)}) + \beta}, \end{aligned} \tag{11}$$

which can be interpreted as a majorisation-minimisation algorithm to solve (8) using the majorant $q(\boldsymbol{x} | \boldsymbol{x}_0)$ [8], or equivalently as an expectation-maximisation algorithm [17]. Again, in all our experiments (11) converged in 5 to 10 iterations.

It is worth noticing that typically $\lambda^* \approx \lambda^\dagger$ because $n/k \gg 1$; thus we can expect the Bayesian estimators $\hat{\boldsymbol{x}}^*$ and $\hat{\boldsymbol{x}}^\dagger$ to be practically equivalent from an inferential viewpoint. This observation is relevant because $\hat{\boldsymbol{x}}^*$ arises by setting λ to its most likely value, whereas $\hat{\boldsymbol{x}}^\dagger$ integrates information from all possible values of λ. The fact that for large n both estimators coincide suggests that the joint density $p(\boldsymbol{x}, \lambda | \boldsymbol{y})$ becomes very concentrated around λ^*; which is also in agreement with observations that $p(\boldsymbol{x}, \lambda | \boldsymbol{y})$ and $p(\boldsymbol{x} | \boldsymbol{y})$ are unimodal.

4. APPLICATION TO COMPRESSIVE SENSING

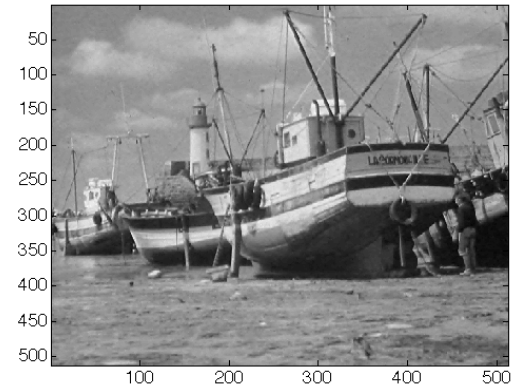
In this section, we illustrate the proposed methodologies with an application to compressive sensing reconstruction using an ℓ_1-wavelet analysis prior. For comparison we also report obtained with SUGAR [13], a recent state-of-the-art SURE-based technique that seeks to minimise the estimation mean-squared-error (MSE), and with an oracle that knows the value of λ that minimises the estimation MSE. To make comparisons fair, all methods were implemented with the same generalised forward-backward solver for (1) (we used the MATLAB implementation of [13]). The methods are compared by computing the peak signal-to-noise ratio ($\mathrm{PSNR}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 10 \log_{10}(255^2 / \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2)$) and the structural similarity index (SSIM) [18] between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$.

For this experiment we suppose that an unknown image $\boldsymbol{x} \in \mathbb{R}^n$, of size $n = 512 \times 512$, is observed through a noisy measurement $\boldsymbol{y} = \Phi \boldsymbol{x} + \boldsymbol{w} \in \mathbb{R}^p$ of size $p = n/2$,

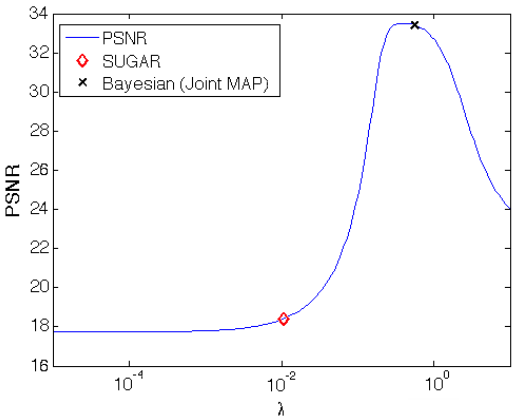
where $\Phi \in \mathbb{R}^{p\times n}$ denotes a compressive sensing random matrix and $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_p)$ additive Gaussian noise with $\sigma^2 = 10$. In a manner akin to [13], we assign $\boldsymbol{x}$ the *analysis* prior $p(\boldsymbol{x}|\lambda) = \exp\{-\lambda\|\boldsymbol{\Psi x}\|_1\}/C(\lambda)$, where $\boldsymbol{\Psi}$ represents the horizontal and vertical (but not the diagonal) components of a *Daubechies 4* wavelet transform. Notice that $\|\boldsymbol{\Psi}(\alpha \boldsymbol{x})\|_1 = \alpha\|\boldsymbol{\Psi x}\|_1$ for all $\alpha > 0$, and therefore $h(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_1$ is k-homogeneous with $k = 1$. We report experiments with the two widely used test images `Boat` and `Mandrill`.

Notice that the value of λ defining $p(\boldsymbol{x}|\lambda)$ remains unspecified and will be 1) estimated jointly with $\boldsymbol{x}$ by joint maximisation using algorithm (7); 2) removed from the model by marginalisation using algorithm (11); and 3) selected using SUGAR. We note that algorithms (7) and (11) were initialised with $\boldsymbol{x}^{(0)} = \Phi^T(\Phi\Phi^T)^{-1}\boldsymbol{y}$, and that SUGAR was implemented using the prediction MSE (see [13] for more details).

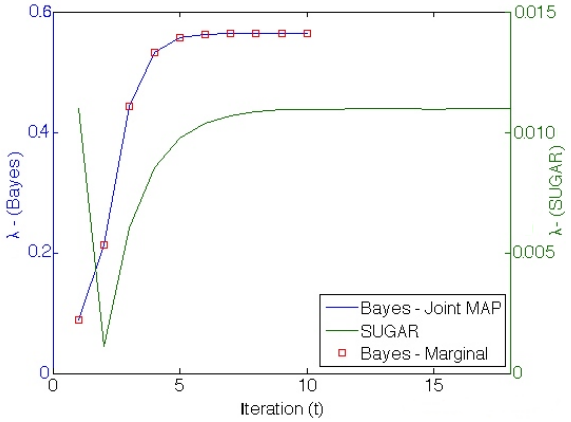
Tables 1 and 2 show the values of λ, performance indicators and computing times for each method and for the two test images `Boat` and `Mandrill`. For completeness, we also report the results obtained by least-squares estimation (i.e., $\hat{\boldsymbol{x}}_{LS} = \Phi^T(\Phi\Phi^T)^{-1}\boldsymbol{y}$), and with an oracle that knows the optimal value of λ. We observe that the proposed Bayesian techniques produced very good results, and that they achieved reconstruction accuracies that are close to the oracle performance for both images. These results are interesting because they reveal that the prior $p(\boldsymbol{x}|\lambda)$, in spite of being simplistic, captures enough knowledge about $\boldsymbol{x}$ to promote values of λ that are in agreement with what we would select by visual cross-validation (and that also lead to excellent PSNR and SSIM performances). Tables 1 and 2 also show that SUGAR performed worse, particularly for the `Boat` image where it achieved an estimation accuracy comparable to that of the least-squares estimate. Moreover, notice that the Bayesian techniques were approximately four times faster than SUGAR in both experiments. This difference in computing time is due to the facts that: 1) the Bayesian techniques converged in approximately half as many iterations as SUGAR, and 2) the Bayesian techniques require solving (1) once per iteration, whereas each iteration of SUGAR computes (1) twice. Finally, for illustration, Fig. 1(a) shows the Bayesian joint MAP estimate (7) corresponding to the `Boat` experiment; Fig. 1(b) the estimation PSNR as a function λ; and Fig. 1(c) the evolution of the iterates $\lambda^{(t)}$ for the two Bayesian methods and for SUGAR.



(a) Joint MAP ($\lambda^* = 56.4$, PSNR=33.4)



(b) Estimation accuracy (PSNR) vs λ



(c) Evolution of the iterates $\lambda^{(t)}$

Fig. 1. Compressive sensing experiment with the `Boat` image: (a) Bayesian joint MAP estimate (7). (b) Estimation PSNR as a function of λ. (c) Evolution of the iterates $\lambda^{(t)}$ for the proposed Bayesian methods (7) and (11) (left axis) and for SUGAR (right axis).

5. CONCLUSION

In this paper, we proposed two hierarchical Bayesian methods for performing maximum-a-posteriori inference when the value of the scalar regularisation parameter is unknown. The methods are useful for Bayesian models with priors of the form $p(\boldsymbol{x}|\lambda) = \exp\{-\lambda h(\boldsymbol{x})\}/C(\lambda)$, and for which h is k-homogenous. This includes many important priors, including all norms, composite norms, and compositions of norms with linear operators, which are 1-homogenous. A

Table 1. Values of λ, estimation accuracy (PSNR and SSIM), and computing times for the `Boat` experiment.

	λ	PSNR	SSIM	time [sec]
Joint MAP (7)	0.56	33.4	0.96	299
Marginal MAP (11)	0.56	33.4	0.96	299
SUGAR [13]	0.01	18.4	0.55	1137
Oracle	0.38	33.5	0.96	n/a
Least-squares	n/a	17.7	0.52	0.04

Table 2. Values of λ, estimation accuracy (PSNR and SSIM), and computing times for the `Mandrill` experiment.

	λ	PSNR	SSIM	time [sec]
Joint MAP (7)	0.20	25.3	0.87	229
Marginal MAP (11)	0.20	25.3	0.87	229
SUGAR [13]	0.10	22.9	0.80	984
Oracle	0.50	26.1	0.90	n/a
Least-squares	n/a	18.6	0.22	0.04

main contribution of the paper was showing that for all these models, the normalising factor of the prior takes the form $C(\lambda) = D\lambda^{-n/k}$. Based on this, we developed two hierarchical Bayesian inference techniques to estimate λ from $\boldsymbol{y}$ jointly with $\boldsymbol{x}$, or to remove λ from the joint model by marginalisation. The proposed methodologies were finally illustrated on an application to compressive sensing using a ℓ_1-wavelet analysis prior and compared to the state-of-the-art SURE-based technique SUGAR [13].

It is worth emphasising that, although this paper focused exclusively on MAP estimators, knowledge of $C(\lambda)$ enables the complete spectrum of Bayesian analysis techniques, including other estimators, credibility/confidence sets, Bayesian hypothesis tests, etc. Similarly, our results are also useful for other types of inference algorithms, for example expectation-maximisation, variational Bayes, and Markov chain Monte Carlo [19]. A detailed theoretical analysis of the Bayesian estimators and optimisation algorithms described in Section 3 is currently under investigation. Future works will also focus on the development of *empirical Bayesian* techniques that set λ by maximum likelihood estimation [4, 20], and on the application of the proposed techniques to other canonical inverse problems such as image deconvolution.

Finally, preliminary experiments (not reported in this paper) suggest that Bayesian techniques generally outperform SURE methods when the projection and prediction MSEs are poor surrogates of the true MSE (e.g., for problems involving rank-deficient observation operators). On the other hand, SURE techniques perform better for denoising problems. We do not see Bayesian and SURE techniques as competitors; instead we hope and anticipate that future methods will use both approaches in a complementary manner.

REFERENCES

[1] J.-F. Giovannelli and J. Idier, *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*, ISTE Ltd and John Wiley and Sons, 2015.

[2] A.M. Thompson, J.C. Brown, J.W. Kay, and D.M. Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 326–339, Apr. 1991.

[3] G. Archer and D.M. Titterington, "On some Bayesian/regularization methods for image restoration," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 989–995, Jul. 1995.

[4] R. Molina, A. Katsaggelos, K. Aggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 231–246, 1999.

[5] M. Figueiredo M. Almeida, "Parameter estimation for blind and non-blind deblurring using residual whiteness measures," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2751–2763, 2013.

[6] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, New York, New York, third edition, 2013.

[7] A. Gelman and X.-L. Meng, "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling," *Statist. Sci.*, vol. 13, no. 2, pp. 163–185, May 1998.

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

[9] G. Celeux, F. Forbes, and N. Peyrard, "EM procedures using mean field-like approximations for Markov model-based image segmentation," *Pattern Recognition*, vol. 36, no. 1, pp. 131 – 144, 2003.

[10] M. Pereyra, N. Dobigeon, H. Batatia, and J.-Y. Tourneret, "Estimating the granularity parameter of a Potts-Markov random field within an MCMC algorithm," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2385–2397, June 2013.

[11] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, 11 1981.

[12] J.-C. Pesquet, A. Benazza-Benyahia, and C. Chaux, "A SURE approach for digital signal/image deconvolution problems," *IEEE Trans. Sig. Process.*, vol. 57, no. 12, pp. 4616–4632, Dec. 2009.

[13] C. Deledalle, S. Vaiter, G. Peyré, and J. Fadili, "Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection," *SIAM J. Imaging Sci.*, vol. 7, no. 4, pp. 2448–2487, 2014.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[15] G. Box and G. Tiao, *Bayesian Inference in Statistical Analysis*, John Wiley and Sons, 1973.

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

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[18] W. Zhou, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[19] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, New York, second edition, 2004.

[20] D. JC. MacKay, "Hyperparameters: Optimize, or integrate out?," in *Maximum entropy and Bayesian methods*, pp. 43–59. Springer, 1996.