COMPARING BAYESIAN MODELS IN THE ABSENCE OF GROUND TRUTH

Marcelo Pereyra[*]

School of Mathematics
University of Bristol
Bristol, UK.
marcelo.pereyra@bristol.ac.uk

Steve McLaughlin[†]

School of Engineering and Physical Sciences
Heriot Watt University
Edinburgh, UK.
s.mclaughlin@hw.ac.uk

ABSTRACT

Modern signal processing methods rely strongly on Bayesian statistical models to solve challenging problems. This paper considers the objective comparison of two alternative Bayesian models, for scenarios with no ground truth available, and with a focus on model selection. Existing model selection approaches are generally difficult to apply to signal processing because they are unsuitable for models with priors that are improper or vaguely informative, and because of challenges related to high dimensionality. This paper presents a general methodology to perform model selection for models that are high-dimensional and that involve proper, improper, or vague priors. The approach is based on an additive mixture meta-model representation that encompasses both models and which concentrates on the model that fits the data best, and relies on proximal Markov chain Monte Carlo algorithms to perform high-dimensional computations efficiently. The methodology is demonstrated on a series of experiments related to image resolution enhancement with a total-variation prior.

Index Terms— Statistical signal processing; Bayesian inference; model selection; Markov chain Monte Carlo; computational imaging

1. INTRODUCTION

Modern signal processing (SP) methods rely on statistical models and inference methods to solve SP problems (we use SP here to cover all relevant statistical signal, image and multimedia processing problems). Most problems considered in the literature involve inference on an unknown signal of interest from some observed raw data. Despite the wide range of different problems and of different applications considered, all statistical SP methods proceed in essentially the same way: first the specification of a statistical model relates the observation with the unknown, followed by a statistical inference procedure conditional on the modelling assumptions (typically point estimation by using maximum likelihood or Bayesian strategies). Of course, the choice of the statistical model has a significant impact on the estimation results, and hence the development of ever more accurate and application-specific models is the focus of perpetual research effort.

An important related question, which we investigate in this paper, is how to objectively compare two competing models and decide which one is the most suitable for analysis of a given dataset [1]. The statistical SP community has traditionally addressed this issue by conducting benchmark experiments and reporting measures of estimation accuracy with respect to some ground truth (e.g. mean-squared-error of the estimation). This approach has some important limitations, particularly it cannot be applied in real data settings where no ground truth is available (as well as some other minor issues such as potentially conflicting performance metrics) [2].

This paper considers the problem of objectively comparing two alternative Bayesian models without reference to ground truth. The Bayesian framework provides several ways of performing model selection or model comparisons (see [3, Sec. 1] for a survey on this topic and [4] for a recent application in signal processing). However, to the best of our knowledge these are unsuitable for many important SP problems because they cannot be applied to models involving improper priors or vaguely informative priors, which are ubiquitous in modern SP (this difficulty arises because the model's evidence or marginal likelihood is either not defined or takes arbitrary values). The method presented in this paper is based on the new Bayesian hypothesis testing framework recently introduced in [3], which allows the use of improper and vaguely informative priors.

2. PROBLEM STATEMENT

Consider the estimation of a high-dimensional signal of interest $\boldsymbol{x} \in \mathbb{R}^n$ from some observation $\boldsymbol{y} \in \mathbb{R}^p$, related to $\boldsymbol{x}$ by a statistical model with likelihood function $f_0(\boldsymbol{y}|\boldsymbol{x})$. Suppose that the recovery of $\boldsymbol{x}$ from $\boldsymbol{y}$ is ill-posed or ill-

[*]Marcelo Pereyra holds a Marie Curie Intra-European Research Fellowship for Career Development.

[†]Steve McLaughlin acknowledges the support of EPSRC via grant EP/J015180/1.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

conditioned, and that we address this difficulty by following a Bayesian approach [1]. More precisely, we model $\boldsymbol{x}$ as a random vector with prior distribution $p_0(\boldsymbol{x})$ promoting solutions with expected structural or regularity properties (e.g., sparsity or smoothness), and combine observed and prior information by using Bayes' theorem, leading to the posterior distribution

$$\mathcal{M}_0: \quad p_0(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_0(\boldsymbol{y}|\boldsymbol{x})p_0(\boldsymbol{x})}{\int_{\mathbb{R}^n} f_0(\boldsymbol{y}|\boldsymbol{u})p_0(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}, \tag{1}$$

which represents the knowledge about $\boldsymbol{x}$ after observing $\boldsymbol{y}$ under the assumption of the Bayesian model $\mathcal{M}_0 = \{f_0, p_0\}$.

From a methodological viewpoint, once the posterior $p_0(\boldsymbol{x}|\boldsymbol{y})$ is properly specified, the estimation of $\boldsymbol{x}$ can be straightforwardly addressed by Bayesian point estimation [e.g., by computing the minimum-mean-squared-error estimator $\hat{\boldsymbol{x}}_{MMSE} = \int_{\mathbb{R}^n} \boldsymbol{x} p_0(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{x}$, or the maximum-a-posteriori estimator $\hat{\boldsymbol{x}}_{MAP} = \operatorname{argmax}_{\boldsymbol{x}\in\mathbb{R}^n} p_0(\boldsymbol{x}|\boldsymbol{y})$] [1].

Now consider an alternative Bayesian model $\mathcal{M}_1$, with likelihood function $f_1(\boldsymbol{y}|\boldsymbol{x}')$, prior $p_1(\boldsymbol{x}')$, and posterior distribution given by

$$\mathcal{M}_1: \quad p_1(\boldsymbol{x}'|\boldsymbol{y}) = \frac{f_1(\boldsymbol{y}|\boldsymbol{x}')p_1(\boldsymbol{x}')}{\int_{\mathbb{R}^n} f_1(\boldsymbol{y}|\boldsymbol{u})p_1(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}. \tag{2}$$

Note that we use $\boldsymbol{x} \in \mathbb{R}^n$ in (1) and $\boldsymbol{x}' \in \mathbb{R}^n$ in (2) to represent explicitly that each model may involve a different parametrisation of the signal of interest, in addition to different likelihood and prior distributions for the unknown.

This paper considers the problem of objectively comparing the two Bayesian models $\mathcal{M}_0$ and $\mathcal{M}_1$ in the case where there is no ground truth available, $\boldsymbol{x}$ and $\boldsymbol{x}'$ are high-dimensional, and the priors $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x}')$ are either improper or vaguely informative. As explained previously, Bayesian model comparison for these types of models, ubiquitous in statistical SP, is a very challenging problem.

3. BAYESIAN MODEL COMPARISON METHOD

This section presents a new Bayesian approach for comparing intrinsically (i.e., without ground truth available) two alternative high-dimensional Bayesian models $\mathcal{M}_0$ and $\mathcal{M}_1$. The approach is based on the Bayesian hypothesis testing framework recently introduced in [3], and on the proximal Markov chain Monte Carlo computation framework [5]. For simplicity, in this paper we focus on models with priors with k-homogenous regularisers, and in which the elements of $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{x}$ or $\boldsymbol{x}'$. However, the results presented hereafter can be easily generalised to models involving dependent observations, as well as other types of priors (see [3] for more details).

3.1. Specification of $\mathcal{M}_0$ and $\mathcal{M}_1$

The model comparison approach considered in this paper is useful for pairs $\mathcal{M}_0$ and $\mathcal{M}_1$ that verify the following conditions. First, we assume that the observations $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{x}'$, and consequently that the likelihoods $f_0(\boldsymbol{y}|\boldsymbol{x})$ and $f_1(\boldsymbol{y}|\boldsymbol{x}')$ factorise as follows

$$f_0(\boldsymbol{y}|\boldsymbol{x}) = \prod_{j=1}^{p} f_0(y_j|\boldsymbol{x}), \quad f_1(\boldsymbol{y}|\boldsymbol{x}') = \prod_{j=1}^{p} f_1(y_j|\boldsymbol{x}') \tag{3}$$

where y_j denotes the j-th element of $\boldsymbol{y}$, and $f_0(y_j|\boldsymbol{x})$ and $f_1(y_j|\boldsymbol{x}')$ its marginal likelihood under model $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. This property holds for all models involving independent (i.e., white) noise. Second, we assume that the priors $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x}')$ can be expressed as follows

$$\begin{aligned} p_0(\boldsymbol{x}) &\propto \exp\{-\lambda_0 h(\boldsymbol{x})\}, \\ p_1(\boldsymbol{x}') &\propto \exp\{-\lambda_1 h(\boldsymbol{x}')\}, \end{aligned} \tag{4}$$

for some $\lambda_0 \in \mathbb{R}$ and $\lambda_1 \in \mathbb{R}$, and a common prior sufficient statistic or regulariser $h : \mathbb{R}^n \to \mathbb{R}$ that is k-homogenous (we say that h is k-homogenous if there exists $k \in \mathbb{R}^+$ such that $h(\eta\boldsymbol{x}) = \eta^k h(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \forall \eta > 0$). Notice that (4) holds for all norms (e.g., ℓ_1, ℓ_2, total-variation, nuclear, spectral, etc.), composite norms (e.g., $\ell_1 - \ell_2$), and compositions of norms with linear operators (e.g., analysis regularisers of the form $\|\Psi\boldsymbol{x}\|_1$), all of which are 1-homogenous [6]. Similarly, power of norms with exponent q are q-homogenous.

Finally, for computational tractability, we assume that the joint densities $f_0(\boldsymbol{y}|\boldsymbol{x})p_0(\boldsymbol{x})$ and $f_1(\boldsymbol{y}|\boldsymbol{x}')p_1(\boldsymbol{x}')$ admit a decomposition

$$\begin{aligned} f_0(\boldsymbol{y}|\boldsymbol{x})p_0(\boldsymbol{x}) &= \exp\{s_0(\boldsymbol{x}) - g_0(\boldsymbol{x})\}, \\ f_1(\boldsymbol{y}|\boldsymbol{x}')p_1(\boldsymbol{x}') &= \exp\{s_1(\boldsymbol{x}') - g_1(\boldsymbol{x}')\}, \end{aligned} \tag{5}$$

such that s_0 and s_1 are continuously differentiable functions, and g_0 and g_1 are lower-semicontinuous convex functions with tractable proximal operators[1].

Notice that assumptions (3), (4), and (5) hold for many Bayesian models that are widely used in statistical SP. For example, they hold for all models of the form

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\{-\|\boldsymbol{y} - A\boldsymbol{x}\|_2^2/2\sigma^2 - \lambda\phi(\boldsymbol{x})\},$$

where $A \in \mathbb{R}^{p\times n}$ is a linear observation operator, $\sigma > 0$, and ϕ is a k-homogenous convex regulariser whose proximal operator is known analytically or can be computed efficiently with a specialised algorithm.

3.2. Bayesian comparison of $\mathcal{M}_0$ and $\mathcal{M}_1$

In a manner akin to [3], we first reparametrise $\mathcal{M}_1$ such that $\boldsymbol{x}$ and $\boldsymbol{x}'$ have the same prior distribution (this enables the use improper and vaguely informative priors). Because $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x}')$ verify (4) this can be achieved straightforwardly

[1] The proximal operator of a lower-semicontinuous convex functions $g : \mathbb{R}^n \to (-\infty, \infty]$ at a point $\boldsymbol{x} \in \mathbb{R}^n$ is defined as $prox_g(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} g(\boldsymbol{u}) - \|\boldsymbol{x} - \boldsymbol{u}\|_2^2$ [7].

by using the scaling transformation $\boldsymbol{x}' = (\lambda_0/\lambda_1)^{-k}\boldsymbol{x}$. We denote the reparametrised likelihood function and prior for $\mathcal{M}_1$ as $f_1(\boldsymbol{y}|\boldsymbol{x})$ and $p_1(\boldsymbol{x})$ respectively.

Following on from this, we define a single Bayesian meta-model encompassing $\mathcal{M}_0$ and $\mathcal{M}_1$. We introduce a model-selection auxiliary variable $\alpha \in [0, 1]$ and construct the augmented posterior

$$p(\boldsymbol{x}, \alpha|\boldsymbol{y}) \propto f_{01}(\boldsymbol{y}|\boldsymbol{x}, \alpha)p(\boldsymbol{x})p(\alpha) \tag{6}$$

where the likelihood $f_{01}(\boldsymbol{y}|\boldsymbol{x}, \alpha)$ arises from the additive mixture representation

$$f_{01}(\boldsymbol{y}|\boldsymbol{x}, \alpha) = \prod_{j=1}^{p}[(1-\alpha)f_0(y_j|\boldsymbol{x}) + \alpha f_1(y_j|\boldsymbol{x})],$$

and where $f_0(y_j|\boldsymbol{x})$ and $f_1(y_j|\boldsymbol{x})$ are the marginal likelihoods of y_j given $\boldsymbol{x}$ under each model [notice that $f_{01}(\boldsymbol{y}|\boldsymbol{x}, \alpha = 0) = f_0(\boldsymbol{y}|\boldsymbol{x})$ and $f_{01}(\boldsymbol{y}|\boldsymbol{x}, \alpha = 0) = f_1(\boldsymbol{y}|\boldsymbol{x})$]. The prior $p(\boldsymbol{x}) = p_0(\boldsymbol{x}) = p_1(\boldsymbol{x})$ because of the reparametrisation of $\mathcal{M}_1$. In order to achieve model selection -as opposed to model averaging- the prior on α should concentrate on 0 and 1. Following [3], we use the symmetric beta prior $\alpha \sim \text{Beta}(a_0, a_0)$ with $a_0 \ll 1$ (in our experiments we used $a_0 = 0.1$).

Finally, in order to perform model selection we focus on the marginal posterior distribution

$$p(\alpha|\boldsymbol{y}) = \int p(\boldsymbol{x}, \alpha|\boldsymbol{y})\mathrm{d}\boldsymbol{x}, \tag{7}$$

which measures the degree of support of $\boldsymbol{y}$ for $\mathcal{M}_0$ (if $\alpha|\boldsymbol{y}$ concentrates on 0), for model $\mathcal{M}_1$ (if $\alpha|\boldsymbol{y}$ concentrates on 1), or for both (if $\alpha|\boldsymbol{y}$ concentrates equally on 0 and 1) [3].

3.3. Computation of $p(\alpha|\boldsymbol{y})$

The marginal density (7) is generally computationally intractable due to the high dimensionality of $\boldsymbol{x}$. In this paper we address this difficulty by using a stochastic simulation algorithm to generate a set of samples $\alpha^{(1)}, \ldots, \alpha^{(M)}$ distributed according to (7), which then enables the approximation of posterior probabilities and expectations w.r.t. $p(\alpha|\boldsymbol{y})$ by Monte Carlo integration [8, 9, Ch. 3]. More precisely, we use a Metropolis-Hastings (MH) Markov chain Monte Carlo (MCMC) algorithm to generate a set of samples $(\boldsymbol{x}, \alpha)^{(1)}, \ldots, (\boldsymbol{x}, \alpha)^{(M)}$ distributed according to the joint posterior $p(\boldsymbol{x}, \alpha|\boldsymbol{y})$, and then obtain samples from $p(\alpha|\boldsymbol{y})$ by implicit marginalisation by only recording the coordinate associated with α (this also greatly reduces the algorithm's memory footprint) [9].

The algorithm proceeds in an iterative manner as follows. First, following a Gibbs scheme, at each iteration $\boldsymbol{x}$ and α are updated alternatively by using two different MH kernels targeting the conditional distributions $p(\alpha|\boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{y}, \alpha)$ respectively. The kernel for $p(\alpha|\boldsymbol{y}, \boldsymbol{x})$ is an independent MH step using the prior $p(\alpha)$ as proposal [8], which has most of its mass close to 0 and to 1. The choice of the kernel for $p(\boldsymbol{x}|\boldsymbol{y}, \alpha)$ is significantly more challenging because of the high dimensionality of $\boldsymbol{x}$. An important aspect of the proposed approach is the use of a proximal MH kernel based on the following proposal

$$\begin{aligned} q(\cdot|\boldsymbol{x}) = &\frac{1}{2}\mathcal{N}\left[prox_{\delta g_1}(\boldsymbol{x}) + \delta\nabla s_1(\boldsymbol{x}), 2\delta\boldsymbol{I}_n\right] \\ &+ \frac{1}{2}\mathcal{N}\left[prox_{\delta g_2}(\boldsymbol{x}) + \delta\nabla s_2(\boldsymbol{x}), 2\delta\boldsymbol{I}_n\right], \end{aligned} \tag{8}$$

that combines the two proximal Langevin proposals associated with $\mathcal{M}_0$ and $\mathcal{M}_1$ [5]. Since we expect that $\alpha|\boldsymbol{y}$ will concentrate on 0 or on 1, and that the proximal Langevin proposals are very efficient for simulating individually from $\mathcal{M}_0$ and $\mathcal{M}_1$ in high-dimensional scenarios, we anticipate that the scheme will perform well. The proposal parameter $\delta > 0$ is adjusted by cross-validation to achieve a stationary acceptance probability of approximately 25% to optimise the overall convergence properties of the algorithm [5].

4. EXPERIMENTS

4.1. Bayesian image deconvolution model

In this section we illustrate the proposed methodology on a Bayesian image deconvolution problem with a total-variation prior. In this canonical inverse problem the goal is to recover a high-resolution image $\boldsymbol{x} \in \mathbb{R}^n$ from a blurred and noisy observation $\boldsymbol{y} \in \mathbb{R}^n$ related to $\boldsymbol{x}$ by $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{w}$, where H is a blurring operator and $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$. This inverse problem is ill-posed, a difficulty that Bayesian image deconvolution methods address by exploiting prior knowledge about $\boldsymbol{x}$. Here we use a total-variation norm prior [10], which is a widely used model for this type of problem [11]. The resulting posterior density is log-concave and is given by

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{\exp\left[-(\|\boldsymbol{y} - H\boldsymbol{x}\|^2/2\sigma^2 + \lambda\|\nabla_d\boldsymbol{x}\|_{1-2})\right]}{Z_{\boldsymbol{y}}}, \tag{9}$$

where $\|\cdot\|_{1-2}$ is the composite $\ell_1 - \ell_2$ norm, ∇_d is the two-dimensional discrete gradient operator, and $Z_{\boldsymbol{y}}$ is the distribution's normalising constant. Notice that the total-variation prior is improper because it is invariant to additive shifts of $\boldsymbol{x}$. It is easy to check that (9) verifies the three assumptions (3), (4), and (5) necessary to apply the proposed method: the likelihood factorises as $f(\boldsymbol{y}|\boldsymbol{x}) = \prod_{j=1}^{p} f(y_j|\boldsymbol{x})$ with

$$f(y_j|\boldsymbol{x}) = (2\pi\sigma_0^2)^{-1/2}\exp\{-(y_j - \boldsymbol{h}_j\boldsymbol{x})^2/2\sigma_0^2\},$$

where $\boldsymbol{h}_j \in \mathbb{R}^{1\times n}$ is j-th row of H; the regulariser $h(\boldsymbol{x}) = \|\nabla_d\boldsymbol{x}\|_{1-2}$ is 1-homogenous; and (9) admits the decomposition (5) with $s(\boldsymbol{x}) = -\|\boldsymbol{y} - H\boldsymbol{x}\|^2/2\sigma^2$ Lipschitz continuously differentiable and $g(\boldsymbol{x}) = \lambda\|\nabla_d\boldsymbol{x}\|_{1-2}$ lower-semicontinuous convex, with proximal operator available by using the specialised algorithm [12].

4.2. Comparison of Bayesian and SURE selection of regularisation parameters

In this first experiment we illustrate the proposed model comparison approach in the context of methods to select regularisation parameters. The goal is to recover the `Boat` image of size $n = 256 \times 256$, which we degraded with a uniform blur operator of size 5×5 and additive white Gaussian noise with $\sigma = 0.5$ (this blurred image is depicted in Fig. 1(a)). We consider a comparison between two instances of model (9) involving different values for the total-variation regularisation parameter λ. In model $\mathcal{M}_0$ we use the value $\lambda_0 = 0.0045$ obtained by using the Bayesian approach recently proposed in [6], whereas in model $\mathcal{M}_1$ we use the value $\lambda_1 = 0.0032$ obtained with the state-of-the-art SUGAR method [13] (which is based on the minimisation of a surrogate of the estimation mean-squared-error). Fig.1(b) and Fig.1(c) show the maximum-a-posteriori estimate obtained with each model, which we computed by using a generalised forward-backward optimisation algorithm [14] (we used the MATLAB implementation of [13]). As expected, we observe that both models have restored the fine detail and sharp edges in the image, confirming the good performance of the methods.

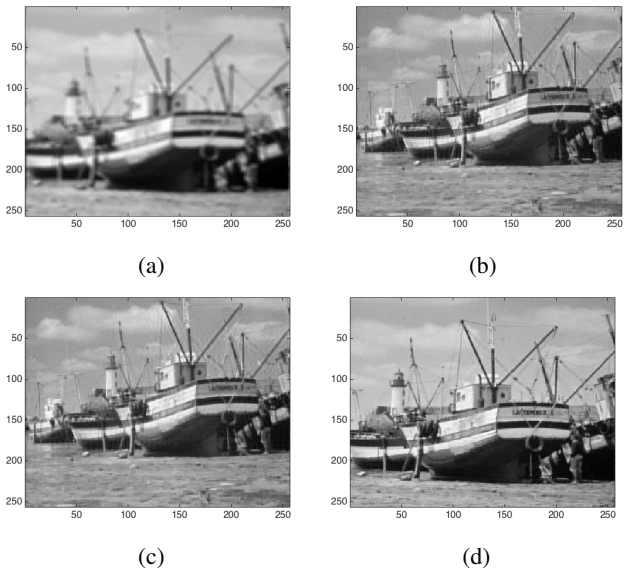

Fig. 1. (a) Blurred `Boat` image of size 256×256 pixels (uniform 5×5 blur, $\sigma = 0.5$), (b) MAP estimate of $\mathcal{M}_0$ given by (9) with the true blur and with $\lambda_0 = 0.0045$ (calculated with the method [6]), (c) MAP estimate of $\mathcal{M}_1$ given by (9) with the true blue and with $\lambda_1 = 0.0032$ (calculated with SUGAR [13]), (d) MAP estimate of $\mathcal{M}_1$ given by (9) with a misspecified uniform blur of size 6×6 pixels and $\lambda_1 = 0.0030$ (calculated with the method [6]).

We argue that the estimate obtained with $\mathcal{M}_0$ by using the Bayesian approach [6] is more accurate than the one obtained with $\mathcal{M}_1$ by using SUGAR [13] (this is related to the fact that SUGAR minimises a projection of the mean-squared-error). However, it is very challenging to provide evidence to support this claim without the availability of a ground truth. Currently the predominant approach is to compare the models $\mathcal{M}_0$ and $\mathcal{M}_1$ objectively, and in this case the two techniques to select the value of λ, compute an estimate of the performance metric with respect to the ground truth. For example, in this experiment $\mathcal{M}_0$ achieved a peak-signal-to-noise-ratio (PSNR) value of 32.7dB, whereas $\mathcal{M}_1$ produced a PSNR value of 32.2dB, supporting the claim that $\mathcal{M}_0$ is more appropriate for data $\boldsymbol{y}$.

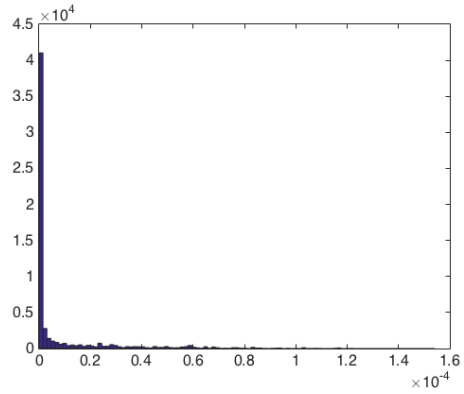

Fig. 2. Histogram approximation of $p(\alpha|\boldsymbol{y})$ for the experiment comparing the methods [6] and [13] to select the value of λ in (9). Notice that $\alpha|\boldsymbol{y}$ concentrates on 0, supporting the hypothesis that the model associated with [6] is more accurate than that of [13].

We now suppose that there is no ground truth available and we use the proposed model comparison approach to evaluate $\mathcal{M}_0$ and $\mathcal{M}_1$ intrinsically. We reparametrise $\mathcal{M}_1$, construct the mixture representation described in Section 3.2, and use the MCMC approach of Section 3.3 to generate $M = 100\,000$ samples $\alpha^{(1)}, \ldots, \alpha^{(M)}$ distributed according to $p(\alpha|\boldsymbol{y})$ (these samples were generated from a chain of $150\,000$ iterations including an initial burn-in stage of $50\,000$ iterations which were discarded). Fig. 2 shows the histogram approximation of $p(\alpha|\boldsymbol{y})$. We observe that $\alpha|\boldsymbol{y}$ is strongly concentrated on 0 (posterior median $\hat{\alpha} = 6 \times 10^{-8}$), clearly supporting the hypothesis that model $\mathcal{M}_0$ is more appropriate to analyse data $\boldsymbol{y}$ than $\mathcal{M}_1$.

4.3. Comparison of two blurring operators

In this second we consider a comparison between two image deconvolution models (9) with different point spread functions. Model $\mathcal{M}_0$ remains unchanged from Section 4.2, but $\mathcal{M}_1$ is now defined with a misspecified linear operator H_1 associated with a uniform blur of size 6×6 pixels (as opposed to H_0 which corresponds to a uniform blur of size 5×5 pixels), and with the regularisation parameter $\lambda_1 = 0.0030$ obtained by using the Bayesian technique [6]. Fig. 1(d) shows the maximum-a-posteriori estimate obtained with $\mathcal{M}_1$ (recall that the estimator for $\mathcal{M}_0$ is depicted in Fig. 1(b)). We observe

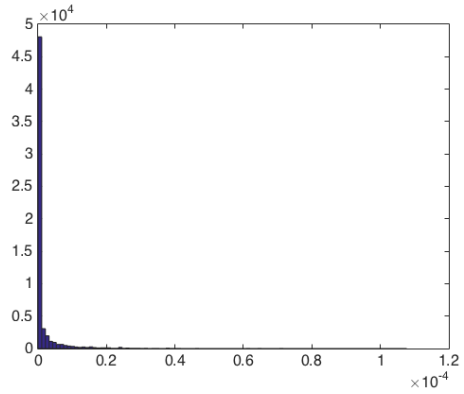

Fig. 3. Histogram approximation of $p(\alpha|\boldsymbol{y})$ for the experiment comparing two different blur operators and values of λ in (9). Notice that $\alpha|\boldsymbol{y}$ concentrates strongly on 0, which corresponds to the model with the true blur operator used to generate $\boldsymbol{y}$.

that despite the fact that the likelihood is not correct, $\mathcal{M}_1$ has nevertheless restored the fine detail in the image, showing that the Bayesian model is robust to mild misspecification and that identifying the correct model is challenging.

Assuming that there is no ground truth available, and using the same setup as in Section 4.2, we use the proposed approach to compare $\mathcal{M}_0$ and $\mathcal{M}_1$ intrinsically. Fig. 3(b) shows the histogram approximation of $p(\alpha|\boldsymbol{y})$ for this comparison. We observe that $\alpha|\boldsymbol{y}$ is correctly concentrated on 0 (posterior median $\hat{\alpha} = 1 \times 10^{-8}$), the model with the correct likelihood, and provides strong evidence that model $\mathcal{M}_0$ is more appropriate to analyse data $\boldsymbol{y}$ than $\mathcal{M}_1$, which has a likelihood that is mildly misspecified.

5. CONCLUSION

This paper has presented a Bayesian methodology to objectively compare two alternative or competing Bayesian models, where no ground truth is available with a focus on model choice. An important advantage of the method is that it can be applied to Bayesian models with proper, improper, or vaguely informative priors with convex regularisers that are k-homogenous, a class of priors that is widely used in statistical signal processing and is beyond the scope of other Bayesian model comparison approaches. The method can be easily applied to high-dimensional models by using modern proximal Markov chain Monte Carlo simulation techniques. The proposed approach was illustrated on two challenging model comparisons related to image deconvolution, where it correctly selected the most accurate models without using any form of ground truth. Future work will focus on a detailed theoretical analysis of the proposed method and inference algorithm, on generalisations to K-model comparisons by using a K-mixture model representation, and on applications to computational imaging problems in which there is significant uncertainty about the model that should be used to analyse the observed raw data.

6. REFERENCES

[1] C. P. Robert, *The Bayesian Choice (second edition)*, Springer Verlag, New-York, 2001.

[2] Zhou Wang and A.C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[3] K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau, "Testing hypotheses via a mixture estimation model," *ArXiv e-prints*, Dec. 2014.

[4] J.-F. Giovannelli and A. Giremus, "Bayesian noise model selection and system identification based on approximation of the evidence," in *IEEE Workshop on Statistical Signal Processing (SSP) 2014*, June 2014, pp. 125–128.

[5] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Statistics and Computing*, 2015, open access paper, http://dx.doi.org/10.1007/s11222-015-9567-4.

[6] M. Pereyra, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Maximum-a-posteriori estimation with unknown regularisation parameters," in *Proc. European Signal Proc. Conf. (EUSIPCO), Nice, France, Sep. 2015.*, Aug 2015, pp. 230–234.

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

[8] C.P. Robert and G. Casella, *Monte-Carlo Statistical Methods*, Springer Verlag, New-York, 1999.

[9] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A.O. Hero, and S. Mclaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE. J. Selected Topics in Signal Process.*, vol. 10, no. 2, pp. 224–241, Mar. 2016.

[10] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992.

[11] 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.

[12] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging Vis.*, vol. 20, no. 1-2, pp. 89–97, 2004.

[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] Hugo Raguet, Jalal Fadili, and Gabriel Peyr, "A generalized forward-backward splitting," *SIAM J. Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.