A Classroom Example of Semi-Blind Deconvolution Using Sample-Based Inference

Colin Fox and Richard A. Norton
Electronics Group
University of Otago
Dunedin, NZ
WWW: http://elec.otago.ac.nz/w/index.php/Colin_Fox

***Abstract*—We present an assignment example that asks the students to perform a blind deconvolution task using sample-based Bayesian inference, as well as some extensions to the structuring of the sampling task that provide substantial computational speed up. This example has the role of providing a direct comparison with the regularized solution that students have previously evaluated; for this example the resulting mean deconvolved image has the same visual quality as achieved by regularization methods while demonstrating that the (effective) regularizing parameter can determined within the inferential framework, unlike regularization where an extra *ad hoc* procedure is required. The extension we present restructures the probabilistic calculation by integrating over the latent field (unknown image) to give exactly the same inference result, but reduces computational cost to *less than* that required for regularized inversion, when selection of the regularizing parameter is accounted for.**

I. Introduction

Bayesian inference generalizes the traditional signal processing methods to allow quantification of uncertainties in estimates, as well as providing a framework in which data processing may be informed by physical models. Historically, these advantages have been offset by greater computational cost of implementing inference compared to signal processing, since the latter has been developed around highly efficient algorithms, such as the fast Fourier transform. However, ever since the remarkable four-volume treatise on *Detection, Estimation and Modulation Theory* by Harry Van Trees [1]–[4], dating from 1968 (!), it has been evident that Bayesian statistics provides a comprehensive and principled framework that generalizes traditional 'estimator-based' methods for signal processing, allowing for model-based inference with quantified uncertainties. Over the past 20 years there has been intensive research into developing new computational tools for Bayesian inference, so that now there are examples of inference that are actually more computationally efficient than the traditional estimator-based approaches, including the example presented here.

Inference has been taught in the Electronics 4th-year at Otago since 2009, is now a component of just about every research project undertaken within the group, whether student graduate project or one of the funded research projects, and compliments the more traditional 'circuits and systems' topics in the undergraduate electronics syllabus. Collectively, these topics facilitate the style of measurement devices that we build at Otago, that one could characterize as "small front-end, large back-end". Inference provides the means of constructing the large back-end that analyses measurements (made by the small front-end) whereby outputs can be informed by physical models of the world. A good example is our 'fast fix' GPS unit that requires just 2 milli-seconds of radio signal to make a GPS fix, precisely because the fix is performed by inference, in contrast to signal processing methods that require around 30 seconds to make a fix. That reduction in radio-on time translates into smaller battery requirements and longer operational lifetime.

Examples in class and assignments are often taken from image recovery, such as image deblurring, to take advantage of the intuitive notion of 'quality' of a reconstructed image that most people have, as well as the inherent interest that an image can provide. Small images, such as a 256×256 pixel image, present a manageable computational task for the non-specialist, while being large enough that computational efficiency does need to be considered if computation time is to not be annoying.

Techniques covered for image deblurring include the regularization methods, that are robust versions of the estimators used in classical signal processing, and also Bayesian inference that provides quantified errors in solutions in the (ubiquitous) presence of noise or other uncertainties. A university-mandated change to 400-level paper structure in 2013 motivated us to split the original paper titled 'Inverse Problems and Inference' into two collectively larger modules, one focusing on inverse problems while the other expanded on the theme of computational Bayesian inference. In this paper we present an assignment example that asks the students to perform a blind deconvolution task using sample-based Bayesian inference, as well as some extensions to the structuring of the sampling task that provide substantial computational speed up for large-scale linear inverse problems.

In this paper we present that assignment example to demonstrate that teaching Bayesian inference is quite feasible, and that the computing required is quite manageable. Of course this example comes after the required theory of probability has been covered, though we find that graduate students in electronics have little difficulty with the operational aspects of Bayesian probability and statistics.

II. An example requiring blind deconvolution

Figure 1 contains a photograph of Jupiter taken in the methane band (780nm) on a grid of size 256×256 pixels, each takes an integer value from 0 to 255. As can be seen, the image is somewhat blurry, so the task is to recover a deblurred

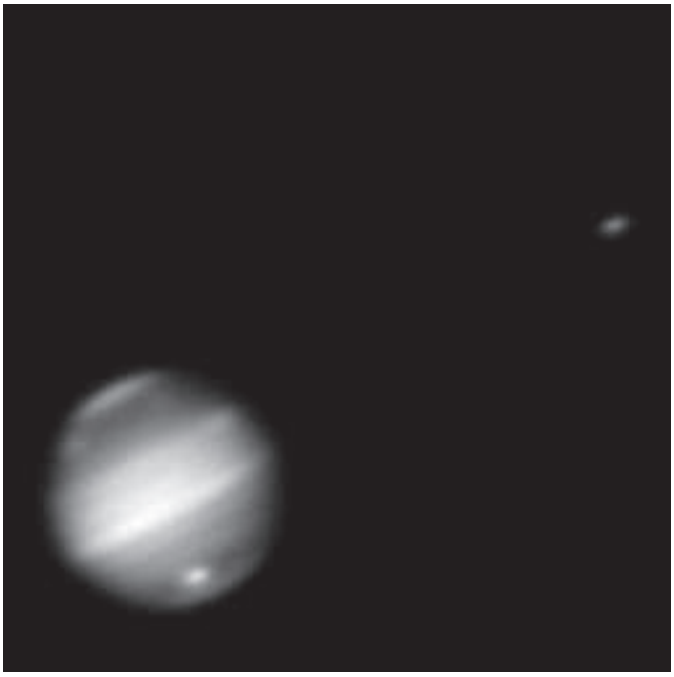

Fig. 1. A blurry photograph of Jupiter taken in the methane band (780nm).

version of this image. Since we are presented with just the blurry image and not the point-spread function, this problem is often called *blind* deconvolution.

The upper right-hand portion of the photograph shows one of the Galilean satellites. This satellite is small enough to be considered close to a point source, and so we can use that region of the photograph as an approximation to the point-spread function. For this reason the title calls this *semi-blind* deconvolution; a challenge question in the assignment asks student to instead model and infer the point-spread function to implement true blind deconvolution, though we do not consider that here. We would note, however, that in all case of successful blind deconvolution that we know of, the blurred photograph contains some feature that is close to the point spread function and so the semi-blind method we use here is actually implemented, though perhaps more automatically.

III. A HIERARCHICAL MODEL FOR LINEAR INVERSE PROBLEMS

A hierarchical stochastic model that occurs in many settings [6] is

$$\begin{aligned} y|x,\theta &\sim N\left(Ax, \Sigma\left(\theta\right)\right) && (1)\\ x|\theta &\sim N\left(\mu, Q^{-1}\left(\theta\right)\right) && (2)\\ \theta &\sim \pi\left(\theta\right) && (3)\end{aligned}$$

Here y is the *observed data*, x is a *latent field,* and θ is a vector of *hyperparameters* that model uncertainty in the two covariance (inverse of precision) matrices Σ and Q^{-1} of the normal (Gaussian) distributions, $N\left(\mu, Q^{-1}\right)$ denotes a normal (aka Gaussian) distribution with mean μ and precision matrix Q or covariance Q^{-1}, while π denotes a general probability distribution.

For example, consider the linear inverse problem where data y is a blurred and noisy version of a true unknown image x. When x and y are $n \times n$ gray-scale pixel images, we represent x and y vectors of pixel values, in this case having length n^2. When the blurring may be modeled by convolution with a (fixed) point-spread function h and the noise is additive zero mean Gaussian, we write

$$\begin{aligned} y &= h * x + n \\ &= Ax + n \end{aligned} \quad (4)$$

where A is the linear operator representing convolution by the point spread function, and n is a noise vector from a zero mean Gaussian with some covariance Σ, i.e., $n \sim N\left(0, \Sigma\right)$. Often the noise covariance is unknown up to some scale δ. For example, we may know that the noise of measurements is independent and identically distributed (iid) with average power λ^{-1} so that $\Sigma = \frac{1}{\lambda} I$, where I is the identity matrix. Since n is a random variable then so is y, with a transformation of variables giving Eqn. 1 showing dependence on λ that is one component of the vector θ. In the language of Bayesian analysis, this defines the *likelihood function* for unknown x and θ once data y is observed.

A common *low-level* model for the unknown image x is to allow pixel values to be arbitrary, though prefer smooth images in which each pixel takes similar values to its neighbours. Thus we need to define a neighbourhood structure, which for the usual pixel lattice may be taken to be the nearest neighbours. We write $j \sim i$ when pixel j is a neighbour of pixel i, and ∂_i for the set of neighbours of pixel i, i.e.

$$\partial_i = \left\{j \neq i | j \sim i\right\}.$$

For the usual pixel lattice ∂_i consists of the pixel locations that are above, below, to the left, and to the right of pixel location i. Denote by $|\partial_i|$ the number of neighbours of pixel i, which will equal 4 for pixels in the interior of the image, 3 for pixels within an edge of the image, and 2 for the corner pixels. Then a suitable stochastic model is given by the *locally linear* Gaussian Markov random field (GMRF) defined by the conditional distributions

$$x_i | x_{\partial_i} \sim N\left(\frac{1}{|\partial_i|}\Sigma_{j\in\partial_i} x_j, \frac{1}{\delta\,|\partial_i|}\right)$$

in which δ is an unknown lumping constant. This, each pixel value is modeled as a Gaussian random variable with mean equal to the average of neighbouring pixel values, and with variance that decreases with the number of neighbours and with increasing lumping constant. The joint density for x may be written as [RH] the Gaussian in Eqn. 2 where the precision matrix $Q = \delta L$ in which

$$L_{ij} = \begin{cases} (\partial_i| & i = j \\ -1 & j \in \partial_i \\ 0 & \text{otherwise} \end{cases}$$

and δ is a component of the vector θ. This is the *prior distribution* in a Bayesian analysis.

Students of electronics will recognize L as the admittance matrix for a square lattice of 1Ω resistors, while graph theorists will recognize L as the graph Laplacian on the graph of pixel neighbours; in both cases the pixels lie at nodes of the graph

while the neighbourhood structure defines which nodes are connected. L also appears in Laplacian regularized inversion, in similar mathematical expressions to the stochastic model.

The hierarchical model is completed by specifying *hyper-priors* over the hyperparameters θ, that we have denoted $\pi(\theta)$ in Eqn. 3.

IV. Options for sample-based inference

A. MCMC sampling from the posterior distribution

In order to perform inference on the system 123 it is typical to utilize Bayes' rule to form the *posterior distribution* over unknowns x and θ conditioned on measured y,

$$\pi(x,\theta|y) = \frac{\pi(y|x,\theta)\,\pi(x,\theta)}{\pi(y)},$$

and then perform MCMC. The numerator is simply the product of the distributions in the stochastic model, i.e. $\pi(y|x,\theta)\,\pi(x,\theta) = \pi(y|x,\theta)\,\pi(x|\theta)\,\pi(\theta)$, and since each has a density that may be evaluated, the posterior may be evaluated up to the normalizing constant, which is all that is required for implementing the Metropolis-Hastings MCMC algorithm.

Commonly a random-walk proposal is employed for components of the image x and hyperparameters θ, which can result in very slow algorithms due to high correlations in the distribution over images, and between the image and hyperparameters. Fortunately there are other sampling strategies that take advantage of the special structure of the stochastic model, that we now discuss.

B. Gibbs sampling with blocking of the image

A distinct speed up can be achieved by recognizing that the conditional distribution over the image x, given everything else, is Gaussian and that samples may be drawn from that distribution using efficient methods from linear algebra. This is often referred to as *blocking* the latent field x, since all components of (the vector) x are updated in a single step.

The joint distribution over images, given hyperparameters θ, can be easily determined by manipulating the Gaussian distributions [6]

$$\begin{pmatrix} x \\ y \end{pmatrix} | \theta \sim N\left(\begin{pmatrix} \mu \\ A\mu \end{pmatrix}, Q_{xy}^{-1} \right) \tag{5}$$

with the joint precision matrix

$$Q_{xy} = \begin{pmatrix} Q + A^T\Sigma A & -A^T\Sigma \\ -\Sigma A & \Sigma \end{pmatrix}.$$

We have omitted the dependence of matrices on θ for brevity. Since conditional Gaussian distributions are easy to write once the precision matrix is known, we see that$\pi(x|y,\theta)$ (the distribution over x conditioned on everything else) is easily determined as the Gaussian distribution

$$x|y,\theta \sim N\left(\mu_{x|y}, Q_{x|y}^{-1}\right) \tag{6}$$

where

$$\begin{aligned} \mu_{x|y,\theta} &= \mu + \left(Q + A^T\Sigma A\right)^{-1} A^T\Sigma\,(y - A\mu) \\ Q_{x|y,\theta} &= Q + A^T\Sigma A. \end{aligned}$$

Bardsley [7]utilized this structure, along with *conjugate* prior distributions over the hyperameters δ and λ that allow the conditional distributions over these parameters to be formed, and hence evaluated. Those distributions are:

$$\lambda|x,\delta,y \sim \Gamma\left(n/2 + \alpha_\lambda, \frac{1}{2}\|Ax - y\|^2 + \beta_{y\lambda}\right) \tag{7}$$

$$\delta|x,\lambda,y \sim \Gamma\left(n/2 + \alpha_\delta, \frac{1}{2}\|Ax - y\|^2 + \beta_\delta\right) \tag{8}$$

where Γ denotes a Gamma distribution, and $\alpha_\lambda, \beta_\lambda, \alpha_\delta, \beta_\delta$ are constants chosen to make the hyperprior distributions relatively uninformative [7].

A block Gibbs sampler may then be implemented by cycling through drawing samples form the conditional distributions in Eqns 6, 7 and 8. The computational cost is contained in the draw from the large Gaussian latent field, though this may be performed efficiently using methods from numerical linear algebra that exploit sparsity, or even more rapidly by Fourier techniques that utilize the circulant structure of the various matrices.

The assignment question asked the students to implement this sampling strategy, and is the method used to generate the example solution, presented later. However one may further block this sampling structure to gain further speed ups, and also remove the need for conjugate prior distributions.

C. One block algorithm

The Gibbs sampler just presented is typically limited in efficiency by the strong correlations between hyperparameters that parameterize the covariance matrices, and the image fields, particularly the latent field x. This means that many (small) steps in each of the θ and x directions are required to explore the joint posterior distribution. Hence greater statistical efficiency may be achieved by devising schemes that can update both θ and x in an effective manner within a single step.

The one block algorithm is so called because the hyperparameter θ and latent field x are blocked together within a single Metropolis-Hastings accept/reject step. In this scheme a proposal hyperparameter θ' is drawn according to a random-walk distribution, then Eqn. 6 used to draw x' conditioned on θ' and y, with the composite proposal (x',θ') accepted with probability given by the usual Metropolis-Hastings rule. A simple calculation shows that the resulting chain in the hyperparameter θ is then ergodic with respect to the *marginal* distribution for $\theta|y$ as if we have been able to integrate out the (nuisance) latent field.

V. Mean deblurred image and discussion

Figure 2 shows the posterior mean deblurred image calculated using the Gibbs sampling scheme in section IV-B. For reasons of brevity we have not shown measures of posterior uncertainty, such as posterior pixel-wise variance; however, it should be appreciated that, just as for the mean statistic, it is possible to calculate any other (predictive) property of the reconstruction along with quantified uncertainties.

The parameters in the hyperpriors supplied in [7] turned out to work well, i.e. were sufficiently non-informative so that

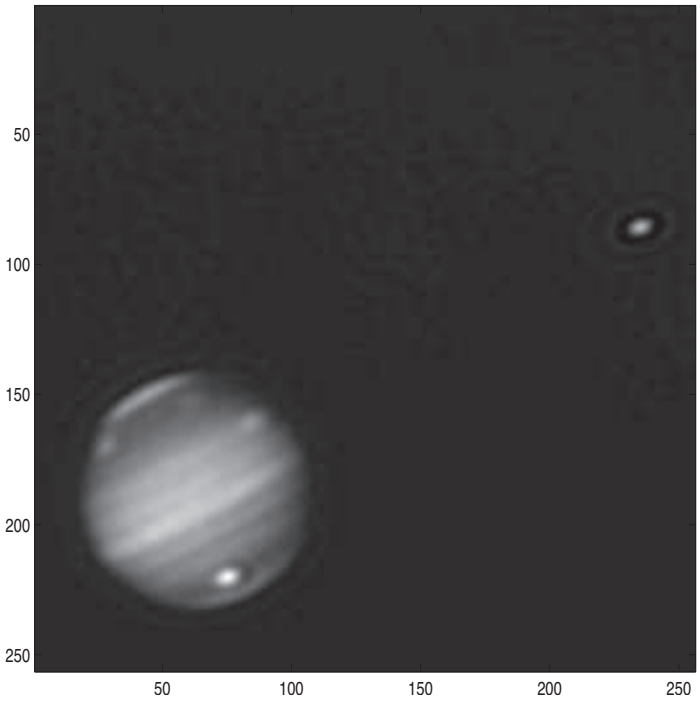

Fig. 2. Posterior mean over deblurred image from a run of 1000 samples with the first 500 removed as burn-in.

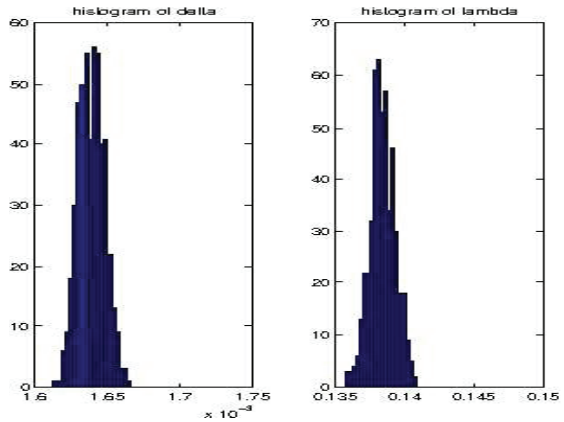

Fig. 3. Posterior histograms over prior lumping hyperparameter δ (left) and measurement noise hyperparameter λ (right).

results do depend on the choice of these parameters. This meant that students were able to use suggested values without modification, which makes the assignment quite manageable.

Marginal posterior histograms over hyperparameters δ and λ are shown in figure 3. The posterior histogram over effective regularizing parameter δ/λ is shown in 4. Note that the regularizing parameter is never explicitly used within the algorithm, but the effective value can be calculated on the basis of inferred lumping constant and measurement noise level. It is very reassuring for students to see that the marginal posterior distribution is exactly centered on values that one would select by eye, or that are selected (for this problem) by the 'L-curve' method. This demonstrates that the inference framework includes automatic selection of the regularizing parameter, so does not need the *ad hoc* procedure that they had to implement

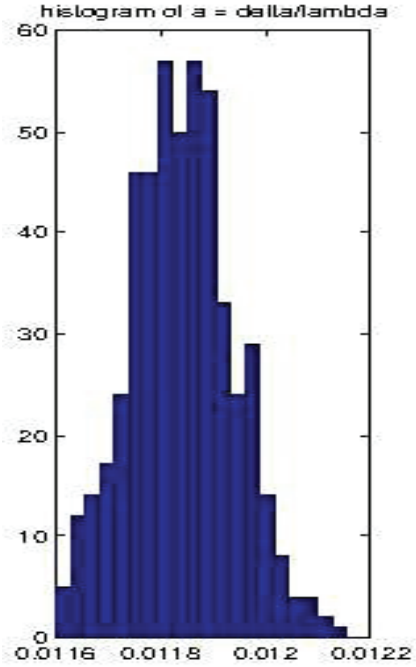

Fig. 4. Posterior histogram over effective regularizing parameter δ/λ.

for regularization.

We have not shown details of the computing required for direct sampling over the marginal posterior distributions for hyperparameters presented in section IV-C. That calculation requires evaluation of determinants that is feasible and cheap in this example, though somewhat outside the scope of this paper.

As it turned out, the main obstacle to the students completing this assignment on time had nothing to do with computational or model complexity, but everything to do with ENZCon2014; all students were fully occupied writing their ENZCon papers!

References

[1] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I, Detection, Estimation, and Linear Modulation Theory*, John Wiley & Sons, New York, 1968, reprinted 2001.

[2] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory, Part II, Nonlinear Modulation Theory*, John Wiley & Sons, New York, 1971, reprinted 2001.

[3] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory, Part III, Radar-Sonar Signal Processing and Gaussian Signals in Noise*, John Wiley & Sons, New York, 1971, reprinted 2002.

[4] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*, John Wiley & Sons, New York, 1971, reprinted 2002.

[5] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Monogr. Statist. Appl. Probab. 104, Chapman and Hall/CRC, Boca Raton, FL, 2005.

[6] D. Simpson, F. Lindgren and H. Rue *Think continuous: Markovian Gaussian models in spatial statistics*, Spatial Statistics Volume 1, May 2012, Pages 16-29.

[7] J. M. Bardsley, *MCMC-Based Image Reconstruction with Uncertainty Quantification*, SIAM J. Sci. Comput., 34(3), A1316–A1332, 2012.