J. R. Statist. Soc. B (2004)
66, *Part* 3, *pp.* 627–652

Dicussion on the meeting on 'Statistical approaches to inverse problems'

Guy Nason (*University of Bristol*)

I congratulate Johnstone and his colleagues and Wolfe and his colleagues on a stimulating and fascinating pair of papers. I shall discuss each paper in turn.

Johnstone, Kerkyacharian, Picard and Raimondo

Johnstone, Kerkyacharian, Picard and Raimondo are interested in the inverse problem of estimating f where f has been convolved with g and then contaminated with white noise. This popular problem has been tackled by a wide variety of procedures and wavelet methods have recently generated considerable interest. Donoho's (1995) seminal wavelet–vaguelette paper introduced the notion that wavelets would be a good choice for the representation of f since real life objects, such as images, are more likely to be efficiently represented using wavelets when compared with, for example, Fourier representations.

Johnstone and his colleagues have moved the field on significantly. In particular, their procedure is more direct than wavelet–vaguelette or Abramovich and Silverman's (1998) vaguelette–wavelet method; it can handle boxcar blur theoretically and practically, they have rates of convergence for $p \neq 2$ (p defines the type of loss) and the paper innovates through use of the new maxiset approach. For me, the most appealing of these innovations is that of enabling the treatment of boxcar blur which is one of the most common types of inverse problem. However, is it really, really, the case that for rational a nothing can be done? Formula (4) compels us to say no, nothing can, but naïvely it still feels wrong.

Formula (19) is the popular 'signal-plus-noise' model but here it is a little different from what normally appears in the literature because the quantities are complex-valued random variables. More specifically, the z_l are zero-mean Gaussian variables which are complex valued and satisfy $E(z_l \bar{z}_k) = \delta_{lk}$. One question is why threshold the $\hat{\beta}_k$ and not the y_l directly? The covariance of the $\hat{\beta}_k$ is given by

$$\operatorname{cov}(\hat{\beta}_k, \hat{\beta}_l) = n^{-1} \sum_m \frac{\bar{\Psi}^k_m \Psi^l_m}{g^2_m}, \tag{1}$$

which would seem to be non-zero for $k \neq l$. Hence, the $\hat{\beta}_k$ are correlated whereas the y_l in formula (19) are (complex valued) independent.

For the Fourier case there is probably not much more that can be said about the transformed noise. Recent work by Barber and Nason (2003) with complex Daubechies wavelets exploits correlations between the real and imaginary parts of transformed noise. In the Fourier case that type of intracoefficient correlation does not exist. It is not clear how we could use the symmetries that do exist.

However, I am not sure that I would advocate direct shrinkage on y_l. This is because h_l is not necessarily a sparse sequence because neither f_l nor g_l is. One might think about performing *complex-valued* wavelet shrinkage on y_l, using, for example, Sardy (2000). The wavelet transform of h_l might possibly be sparse. Doing this would not destroy the computational advantages of the method of Johstone and his colleagues as the wavelet denoising step is $O(n)$.

Some further questions: what can we do when g is unknown (blind deconvolution)?: what about using Bayesian methods in this context? And there are the same old questions: what about non-Gaussian and correlated errors; what about non-uniform designs?

Wolfe, Godsill and Ng

In the enjoyable paper by Wolfe, Godsill and Ng it is reassuring to revisit the 'sparse prior' model

$$p(c_k|\sigma_{ck}, \gamma_k) = (1 - \gamma_k)\, \delta_0(c_k) + \gamma_k\, \mathcal{N}(c_k|0, \sigma^2_{ck}).$$

For me, what is most novel in this paper is the prior specification of γ_k which is capable of modelling several kinds of behaviour: no pattern (unstructured), persistence through time (local stationarity) and persistence through frequency and time. It seems to me, however, that the formulation seems to induce persistence in coefficient *magnitude* rather than *value*. For example, consider a one-stage Markov-in-time prior. If $c_{k-1} = 0$ then does this mean that, with high probability, c_k is also 0? If so, then that is fine.

 1369–7412/04/66627

However, if c_{k-1} is not 0, then, with high probability, c_k is also not 0 but it seems that it need not have the same sign as c_{k-1} and thus could easily be radically different from its neighbour. This seems to be a little different from other models of local stationarity where local control is applied to values (see, for example, Dahlhaus (1997) or Nason *et al*. (2000)).

Mostly, I have some more specific practical questions. What can be said about modelling processes that oscillate too fast for the methodology? Can Markov chain Monte Carlo sampling keep up? How well can it keep up? Can we impose conditions on the speed of processes that can be modelled in this methodology? Furthermore, Wolfe and his colleagues use their methodology on several standard test functions corresponding to examples in Marron *et al.* (1998). It would have been nice to see comparisons with other wavelet shrinkage methods both in terms of quantities such as the mean integrated squared error but also computational performance.

It is with great pleasure that I would like to propose the vote of thanks to both sets of authors for an extremely interesting pair of papers.

Eric Moulines (*Ecole Nationale Supérieure des Télécommunications, Paris*)

Johnstone, Kerkyacharian, Picard and Raimondo

The paper by Johnstone, Kerkyacharian, Picard and Raimondo addresses the noisy deconvolution problem. The problem is formulated in continuous time

$$Y_n(\mathrm{d}t) = f * g(t)\,\mathrm{d}t + \sigma n^{-1/2}\,\mathrm{d}W(t) \tag{2}$$

where f is an unknown function and g is the blurring function assumed to be known and W is Gaussian white noise.

The technique proposed is, compared with the previous method developed in this field, both original and simple, relying on the wavelet shrinkage method. Because, contrary to the functional reconstruction problem, the noisy version of the wavelet coefficients of the unknown function f are not readily available, the subtle trick that is played in this paper consists of using a three-stage estimation technique. In the first stage, the Fourier coefficients f_l of the unknown function f are estimated from the Fourier coefficient of the observation y_l by direct inversion of a blurring function: $\hat{f}_l = y_l/g_l$, where g_l are the Fourier coefficients of the blurring function. This step is at the heart of the Fourier domain deconvolution technique. However, there is a main difference here: in Fourier domain deconvolution, shrinkage is applied directly on the Fourier coefficient, and the deblurred function is directly obtained from the shrinkage estimator of the Fourier series coefficients by inverse Fourier transform. In this implementation, there is no shrinkage at this stage. In the second stage, estimates of the wavelet coefficients of the function f are estimated. This amounts to estimating the scalar product of the unknown signal with the wavelet coefficients, or, thanks to the Parseval identity, as the scalar product of $\hat{f}_l$ and the Fourier coefficients of the wavelet Ψ_k,

$$\hat{\beta}_k = \sum_l \frac{y_l}{g_l}\Psi_l^k.$$

In the paper, the wavelet is assumed to be strictly band limited which restricts the choice of wavelets: this should, however, be regarded as a technicality to avoid truncating sums and introducing an additional source of bias. In the final step, the unknown function f is approximated by standard wavelet shrinkage. Conceptually, this is equivalent to first reconstructing a noisy version of the signal by plain Fourier deconvolution where plain means that the deconvolution is carried out *without any kind* of regularization (e.g. generally implemented by means of shrinkage of the Fourier coefficients) and then to denoise the deblurred signal by the standard wavelet shrinkage estimate with a threshold rule adapted to account for the heteroscedasticity in the noise level affecting the wavelet coefficients introduced by the inverse of the blurring filter. However, playing the trick of computing the wavelet coefficient by using the Parseval identity eliminates the need for reconstructing the 'noisy' deblurred signal and then to compute the wavelet of transform of these coefficients.

The algorithm proposed, which is very interesting and beautifully simple, has, however, some potential practical pitfalls which are not fully addressed. First, the authors assume that the unknown function f is periodic, which is a rather restrictive and most often unnatural assumption in the applications. Second, the authors do not address the potential problems that are associated with sampling. If the sampling issue can presumably be handled without too much trouble the correction of edge effects might raise more serious difficulties. Edge correction is a common problem in all deconvolution methods and in particular in all methods which use a Fourier transform in an intermediate step. If f is not periodic, the Fourier transform

of the convolution $f*g$ no longer is the product of the Fourier coefficients of f and g. In image restoration, edge effects are considered by many researchers to be one of the main sources of distortions: in certain conditions, the artefacts that are associated with the image boundary truncations can dominate the image reconstruction. The problem is perhaps less severe than it might seem because there are several practical methods to address this issue, based on data tapering to smooth the effect of truncation combined with zero padding to the length of the restoration filter (see Aghdasi and Ward (1996) among others). This type of time–spatial domain preprocessing before the computation of the Fourier coefficients might, however, have an effect on the accuracy of the reconstruction (in particular, near the end point) which cannot be predicted without further analysis.

The algorithm proposed shares many similarities with the ForWaRD algorithm of Neelamani *et al.* (2004), which might be thought of as a (mild) generalization of the technique proposed. Neelamani *et al.* (2004) advocate the use of a two-stage shrinkage, which amounts to replacing the 'plain' estimate of the Fourier coefficient $\hat{f}_l = y_l/g_l$ by

$$\hat{f}_l = \frac{y_l}{g_l}\frac{|g_l|^2}{|g_l|^2+\tau}$$

where τ is a data-driven regularization parameter. The intuition behind this estimator is that a certain amount of Fourier domain shrinkage may be beneficial in situations where the function f has a 'reasonably' compact representation in the Fourier domain. The improvements that are brought by this 'two-stage' shrinkage are sometimes quite impressive as illustrated in several experiments that were carried out by Neelamani *et al.* (2004) in image restoration. Of course, the theory to support this type of estimate (which could also have been advocated in the direct observation case) is far less well understood than the theory that is presented here. There is presumably a way to understand the potential advantages of the shrinkage in the Fourier domain before wavelet shrinkage but the results that are provided here show that this might only occur outside Besov spaces!

To conclude, I warmly second the vote of thanks for this paper which presents significant advances both from the practical and the theoretical standpoints!

Wolfe, Godsill and Ng

The paper by Wolfe and his colleagues presents a novel denoising method for audio signals based on a Bayesian regularization scheme applied on a time–frequency representation. The idea is to construct priors for the coefficients of the time–frequency representation which on the one hand favour the smoothness of the reconstructed functions and sparseness of the representation and on the other hand capture the dependences of the coefficients (where the dependence is due in part to the lack of orthogonality of the time–frequency atoms and to the intrinsic time–frequency structure of the signals). The design of prior distribution to guarantee smoothness and sparseness of representation is now well understood and has been applied in many different settings. The use of priors that can capture the dependence between the coefficients of the representation is a more delicate problem which involves *expert a priori knowledge*.

The use of a dependent prior is not a novel idea in image processing, but the idea has not been fully worked out for audio signals. Dependent priors are at the heart of many image segmentation, restoration and retrieval techniques that have been developed (see Geman (1988) among many others). More recently, the use of a dependent prior has also been investigated as a means to improve signal reconstruction in multiscale (wavelet) restoration techniques. Many empirical studies have concluded that the wavelet coefficients (even if the transformation is maximally decimated) of natural images are strongly dependent. The use of appropriately chosen dependent priors led to significant improvements over an 'independent prior' (see for example Wainwright *et al.* (2001), Choi and Baraniuk (2001) and Portilla *et al.* (2003)), strongly supporting the potential advantages that can be obtained by exploiting the dependence structure—and, thus, the redundancy—of the coefficients.

For natural images, the appropriate prior distributions have generally been obtained from the analysis of (large) sets of images (see Wainwright *et al.* (2001)). It would have been of interest to carry out such an analysis for the coefficients of time–frequency representations of audio signals. Whereas the use of a scale mixture of Gaussian distributions to model the marginal distribution is sound and is presumably adequate to account for the large dynamic of these coefficients, an appropriate model for the dependence between the coefficients of the representation is more difficult to guess.

The prior models proposed are sensible but still raise some questions. The authors suggest the use of a Markov random field based on first-order neighbourhood structure. This choice is obviously good to restore noisy black-and-white images: in these applications, we are willing to obtain homogeneous regions

of black and white pixels. This is perhaps not a very sensible prior to restore audio signals because it will have the tendency, in periodic segments, to spread the harmonic energy, thus smearing out the harmonic structure. This effect is noticeable in Fig. 3(e), where the harmonic structure which can clearly be seen even in the time–frequency representation of the noisy signal has almost disappeared in the denoising process. The Markov chain prior is appropriate for restoring horizontal lines and the example shown (Fig. 3(d)) in the paper proves that this prior has a great potential to restore clean speech in the situations where the pitch (fundamental frequency) does not vary significantly compared with the spectral resolution of the window of analysis. When the pitch varies makes less sense.

Another interesting direction of research would be to consider alternative time–frequency representations. There is of course a strong interaction between the choice of the representation and the choice of the prior, but it is likely that the choice of a more complex representation may sometimes help the design of an appropriate prior distribution. For example, given the strong harmonic content of most audio signals, it has been advocated by some to use a dictionary of *harmonic atoms* (see for example Gribonval and Bacry (2003)):

$$h(t) := \sum_{k=1}^{K} c_k \, g_{u,\xi_k}(t), \qquad \xi_k = k\xi_0,$$

where ξ_0 is the pitch frequency and $g_{u,\xi}$ is the Gabor atom (at time u and frequency ξ). Harmonic atoms are designed to preserve harmonicity by forcing the presence of atoms at multiples of the fundamental frequencies. Therefore, when using harmonic atoms, it is no longer necessary to try to shape the prior to restore harmonics, because the harmonics are, with this representation, 'mechanically' reproduced.

As emphasized in these comments, the paper opens many novel directions of research in statistical speech processing. I warmly second the vote of thanks for this paper.

The vote of thanks was passed by acclamation.

Christian P. Robert (*Université Dauphine and Centre de Recherche en Economie et Statistique, Paris*)
Both papers by Cornford and his colleagues and Haario and his colleagues are impressive pieces of work that are perfect representatives of the issues pertaining to inverse problems: huge amounts of (satellite) data, intractable likelihood functions, large dimensions and speed processing requirements. They also illustrate the necessity for multiple levels of approximations that arises in such problems and the correlated difficulty in assessing the effects of such approximations. In addition, they provide sophisticated additions to the theory and practice of Markov chain Monte Carlo (MCMC) algorithms.

Ultimately, the approximations that are used by the authors in both cases are Gaussian: in the work of Haario and his colleagues, the integral defining $T_{\lambda,l}^{\mathrm{abs}}$ is discretized into $T_l(N_l)$ and the base level observations y_l are normal $\mathcal{N}\{T_l(N_l), C_l\}$. In the work of Cornford and his colleagues a normal mixture is built for the inverse model,

$$\mathbf{v}_i | s_i^0 \sim \sum_{k=1}^{4} \alpha_k(s_i^0) \, \mathcal{N}_2\{\mu_k(s_i^0), \sigma_k^2(s_i^0) I_2\}, \tag{3}$$

with neural network estimation for the functions α_k, μ_k and σ_k, but the MCMC processing of the resulting model is found to be too slow and sequential Gaussian approximations are used instead for the dynamic case. (As an aside, note that the issue of additional Gaussian noise $\mathcal{N}_2(0, \tau^2 I_2)$ in the mixture model does not modify the mixture structure since it simply transforms the variance in $(\sigma_k^2 + \tau^2) I_2$.) Still, it seems that a model like

$$p(\mathbf{v}|\mathbf{S}^0) \propto \prod_i \frac{p(\mathbf{v}_i|s_i^0)}{p(\mathbf{v}_i)} \, p(\mathbf{V}),$$

where the $p(\mathbf{v}_i|s_i^0)$s are the mixtures of model (3), should be manageable for simulation as they stand, given that the parameters of the mixtures are known: the function can be computed analytically and either a missing data structure can be introduced for Gibbs sampling simulation (Diebolt and Robert, 1994) or a random-walk Metropolis algorithm can be implemented. The fact that the modes of $p(\mathbf{V}|\mathbf{S}^0)$ need to be known in advance is only an apparent challenge in that the bimodality is a consequence of the lack of identifiability of the direction of the wind.

Even though the papers work within the Bayesian paradigm, I find the prior input fairly vague and limited, as far as the description goes in both papers. For Haario and colleagues, although some further knowledge on the $\rho^{\mathrm{abs}}\{z(s)\}$s, other than their positivity, could be used (altitude and gas correlations

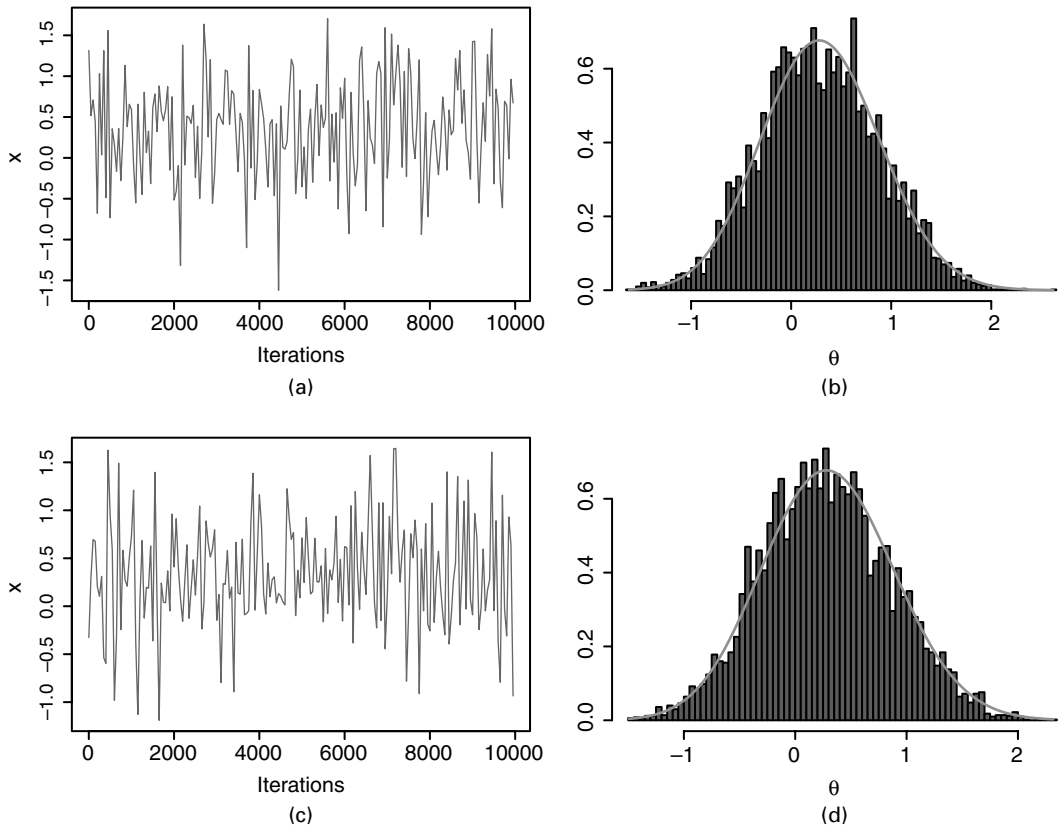

Fig. 1. Random-walk Metropolis simulation of the t posterior distribution $\Pi_{j=1}^{5}\{\nu+(x_j-\theta)^2\}^{-(\nu+1)/2}$ using (a), (b) a normal proposal with variance $\hat{\sigma}_t^2=(5/2)\,\Sigma_{i=1}^{t}(\theta^{(i)}-\hat{\mu}_t)^2$, where $\hat{\mu}_t$ is the average of the $\theta^{(i)}$s, and (c), (d) a ridge-type version $\tilde{\sigma}_t^2=(5/2)\,\Sigma_{i=1}^{t}(\theta^{(i)}-\hat{\mu}_t)^2+\varepsilon$ where $\varepsilon=0.1$: (a), (c) sequences of simulated values and (b), (d) histograms of these values against the true posterior distribution of θ; in both cases, the fit is satisfactory, despite the lack of theoretical guarantees for (a) and (b)

could also appear in the prior), it seems that the main prior modelling is related to the discretization or regularization choice $\gamma=\pm1,\pm2$:

$$x_i=\hat{x}_i\pm\varepsilon^{\mathrm{reg}}\sqrt{(\Delta z^{\gamma})}.$$

For the model of Cornford and colleagues, there also seems to be very little prior input, either in terms of spatial modelling or geophysical and historical knowledge. One reason for this limited use of the possibilities that are offered by the Bayesian approach is the restriction that is imposed by the Gaussian structure of the model, especially in the case of the scatterometer data. (The last sentence of section 4 is also fairly intriguing in that it seems to imply that prior distributions are simply stabilizing devices, rather than summaries of prior information.)

As stated above, an interesting feature of both papers is the devising of novel simulation methodologies to handle the complex posterior distributions that are found there. The adaptive MCMC algorithm of Haario and his colleagues has been introduced in Haario *et al.* (1999, 2001, 2003) to overcome the difficulty of scaling the random-walk Metropolis algorithm by an only adaptation

$$C_t=s_d\,\mathrm{cov}(X_1,\ldots,X_t)+s_d\varepsilon I_d$$

and these papers are forerunners of an emerging class of more efficient MCMC algorithms (Robert and Casella, 2004). Although ergodicity of the resulting process is necessary to establish the validity of the simulation method as an approximating technique, we may wonder about the practical relevance of such constraints. Figs 1 and 2 show that not all adaptive schemes are providing correct approximations to the

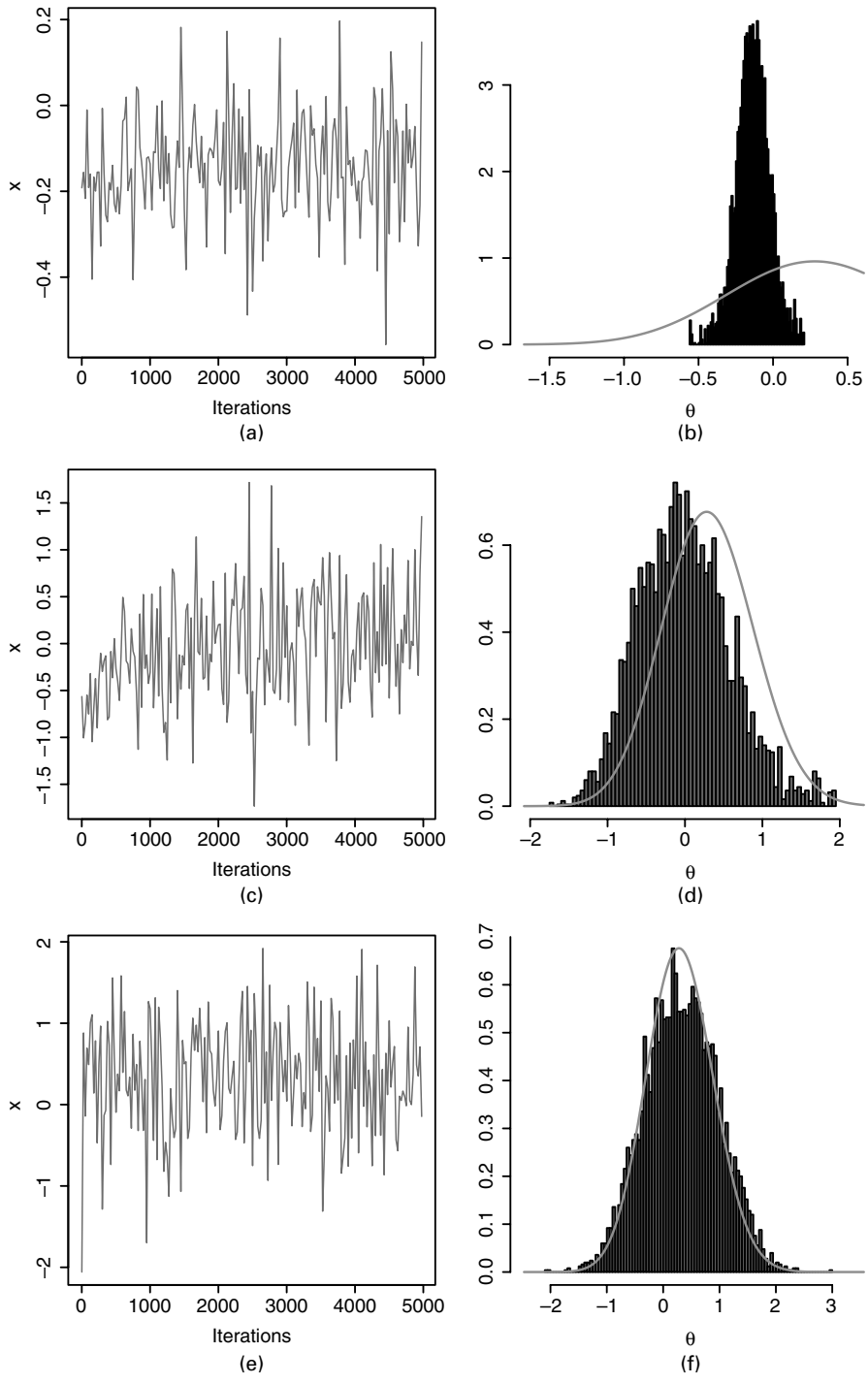

Fig. 2. Influence of the variance of the starting distribution in an adaptive MCMC algorithm with proposals $\mathcal{N}(\hat{\mu}_t, \hat{\sigma}_t^2)$, in the same setting as Fig. 1: the starting variances are (a), (b) 0.1, (c), (d) 1 and (e), (f) 5; even the largest variance fails to provide a convergent approximation to the stationary distribution, and (c), (d) exhibits a case of poor mixing

distribution of interest. As noted in Andrieu and Moulines (2003), other schemes could be used while preserving the stability of the proper distribution. For instance, the updating of the covariance matrix of the proposal could be embedded in a grand chain $(X^{(t)}, \Sigma^{(t)})$ by adding a performance component to the stationary distribution in a tempering mode, $\exp\{-\alpha\, H(\Sigma)\}$. It would also be of considerable interest to understand better why the single-site one-dimensional update SCAM algorithm can be so efficient in high dimensional models since the performances of the Gibbs algorithm usually deteriorate in higher dimensions.

In the case of Cornford and his colleagues, simulation technology is used at two levels: mode jumping MCMC sampling (section 4.1) and variational approximation (section 4.2). The specific algorithm of section 4.1 is not particularly appealing, given that it requires knowledge of the two modes of the posterior distribution. Local, Rao–Blackwellized and population Monte Carlo algorithms could be used as well (Cappé *et al.*, 2004; Robert and Casella, 2004). In particular, a particle filter (Doucet *et al.*, 2001) provides an efficient alternative to the two-stage approach of the authors that allows for simultaneous mode detection and simulation from the posterior. (This is even truer in a dynamic setting.) The variational Gaussian approximation in section 4.2 is presented as an alternative to the costly MCMC algorithm of section 4.1, but we must stress that the focus is also different; this technique provides acceptable approximations to only the first two moments of the posterior distribution and only when that distribution is unimodal.

Although the various approximations that are used by the authors are all acceptable as a result of the *reality constraint* that makes a truly Bayesian resolution impossible (?) to implement, more elaborate assessments that these approximations have limited consequences would be welcomed. Similarly, when iterative algorithms are used, the assessment (or non-assessment) that convergence is not a problem should somehow appear. To conclude, I want to congratulate both groups of authors for obtaining a none-the-less satisfactory inference in such complex but comprehensive inverse problems where statistics must play a role and I thus unreservedly propose the vote of thanks!

Christophe Andrieu (*University of Bristol*)

The two papers by Haario and his colleagues and Cornford and his colleagues that I have been invited to discuss apply standard Bayesian methodology to practical problems. It is quite striking to me that a large proportion of each of the papers is mainly dedicated to the computational, rather than the statistical, aspects of the inversion problem itself. This is reflected in my discussion.

The SCAM algorithm that is proposed by Haario and his colleagues is a natural extension of the authors' adaptive Metropolis algorithm. It is not clear in the current description of the algorithm whether the coordinates are systematically or randomly scanned. I wonder why the authors mainly focus on an adaptive algorithm based on the random-walk Metropolis update. I find this slightly restrictive and would suggest the use of a mixture of transition probabilities of various types. For example one of the kernels could be an adaptive independent sampler with a normal distribution as proposal: the mean and variance would be estimated as they currently are in the AM algorithm. Finally I would like to point out as a complement to theorem 1 that it is possible to prove interesting bounds on the error of adaptive Monte Carlo algorithms. Assume that the algorithm depends on some parameter θ, which is updated along the iterations and assumed to satisfy the two following conditions: for $k \geqslant 1, k, |\theta_k - \theta_{k-1}| \leqslant \varepsilon_k$ for some deterministic series $\{\varepsilon_k\}$ and $\theta_k \in \mathcal{K}$ for some bounded set $\mathcal{K}$. Under fairly general conditions on the transition probability of the chain, it can be proved (Andrieu and Moulines, 2002) that there are constants $A(\varepsilon, \mathcal{K})$ and $B(\varepsilon, \mathcal{K})$ such that, for any $n \geqslant 1$,

$$\sqrt{E|S_n|^2} \leqslant \frac{A(\varepsilon, \mathcal{K})}{\sqrt{n}} + B(\varepsilon, \mathcal{K}) \frac{\sum_{k=1}^{n} \varepsilon_k}{n},$$

where $S_n = n^{-1} \sum_{i=1}^{n} f(X_i)$ and it is assumed that $E_\pi(f) = 0$. The *first term* corresponds to the Monte Carlo fluctuations whereas the *second term* is the price to pay for adaptation. Assume that $\varepsilon_i = i^{-\gamma}$ for $\gamma \in (0, 1)$; then

$$\frac{\sum_{k=1}^{n} \varepsilon_k}{n} \sim \frac{1}{1-\gamma} n^{-\gamma},$$

i.e. the classical Monte Carlo rate is preserved for $\gamma \in [\frac{1}{2}, 1)$. Note that the ratio $A(\varepsilon, \mathcal{K})/B(\varepsilon, \mathcal{K})$ can be shown to depend on some form of continuity of the transition probabilities as a function of θ.

The paper by Cornford and his colleagues is obviously the accumulation of several years of work which were difficult to fit into fewer than 20 pages. It should be pointed out that the problem is difficult in many respects (size and non-linearity), and that the authors should be praised for their efforts to tackle the problem with computationally realistic algorithms. My first comment is related to the lack of estimated parameters. In particular it seems to be assumed that the variance of the observation noise is known and therefore never estimated. How do the authors handle the problem in practice? My second comment is related to the prior distribution and the covariance function that are found by the authors in section 3. It clearly suggests a continuous spatial autoregressive process. It is surprising that the authors did not point this out, since an important body of literature on the topic exists. See among others Jones and Vecchia (1993) and the references therein; one could suggest various adaptations of their method (which computes a maximum likelihood estimate for normal observations and not a posterior distribution for a mixture of normal distributions) to the present context. My third comment is related to the Monte Carlo algorithm that is used. Given the structure of the posterior distribution, which is a mixture, I was wondering why the authors did not consider or discuss the use of latent variables to ease the simulations. Why restrict Monte Carlo methods to the random-walk Metropolis algorithm? Why not use, for example, an independent sampler, e.g. two normal distributions, since the two modes of the posterior distribution are well identified? In the same vein, why did the authors not use their approximation of the posterior distribution in section 4.2 to feed a Markov chain Monte Carlo sampler? This would lead to a fairer comparison of the techniques as they would use the same *a priori* information as is used to design the approximation and could to a certain extent provide them with some feed-back concerning their analytical approximation. These last two points are very often overlooked by many researchers when comparing analytical approximations and Monte Carlo methods.

The vote of thanks was passed by acclamation.

Ad Stoffelen (*Royal Netherlands Meteorological Institute, De Bilt*)
Scatterometer wind retrieval is a non-linear inverse problem and the presentation by Cornford and his colleagues excellently explores a Baysian framework to solve this problem in an effective manner. However, in this contribution the links to state of the art scatterometer processing algorithms are further elaborated. Given the context of the meeting emphasis is put on statistical aspects.

Lorenc (1986) presented variational data assimilation starting from a Baysian framework. In this context, the information that is contained in local observations is exploited by combining it with information in a numerical weather prediction (NWP) model state. As such, Stoffelen and Anderson (1997) characterized the local information content of scatterometer measurements by an ambiguous function of the wind components, based on the results of a local inversion of the geophysical model function. Consequently, this allows effective ambiguity removal in a two-, three- or four-dimensional variational data assimilation system, which is current operational practice. Moreover, and I believe that the authors also show this, not much prior information is needed to perform removal of ambiguity. For this reason, ambiguity removal algorithms have a success rate over 99%, thereby making the dependence on NWP prior information minimal.

Improvements in speed in current scatterometer data assimilation may thus be mainly obtained by more effective local inversion, although the look-up table approach in operational practice is quite effective.

These arguments provide a practical view of the methodology that is presented by Cornford and his colleagues and it is not obvious how they can improve operational scatterometry, if one wanted to. A few possible avenues for exploitation are given below.

(a) In NWP data assimilation formidable effort is put into statistics of observations minus NWP model background, to estimate the observation error covariance structure. The wind vector ambiguity affects such statistics, thus invalidating random-noise assessments. This effect may be quantified.
(b) Besides rain contamination and lower signal-to-noise ratios, Nasa scatterometers measure in a more challenging configuration. The challenge lies in those parts of the swath where the wind vector is quite undetermined, and more sophisticated processing may be useful.
(c) In a quest for higher resolution scatterometer wind products, an intelligent trade-off between resolution and noise must be made to optimize the information content.

Debashis Paul (*Stanford University*)
As an alternative to the thresholding method used in the paper by Johnstone and his colleagues, we propose an estimator derived from a complexity penalized least squares approach. This applies, for example, to linear operators having a wavelet–vaguelette decomposition.

In keeping with the notation that is used in the paper, we denote by $(\mathcal{U}, \mathcal{V}) = (\{u_{jk}\}, \{v_{jk}\})$ the biorthogonal vaguelette pairs associated with the convolution kernel g and the Meyer scaling and wavelet functions. Then by applying a vaguelette transform in the system $\mathcal{U}$ to the noisy data, given by *equation (1)* of the paper, we obtain the sequence model

$$\begin{aligned} \hat{\alpha}_{j_0k} &= \alpha_{j_0k} + \varepsilon z_{j_0k}, & k &= 0, \ldots, 2^{j_0} - 1, \\ \hat{\beta}_{jk} &= \beta_{jk} + \varepsilon z_{jk}, & k &= 0, 1, \ldots, 2^j - 1, \quad j \geqslant j_0. \end{aligned}$$

Here the quantities $\hat{\alpha}_{jk}$ and $\hat{\beta}_{jk}$ have the same definition as in the paper and the noise vectors $z_j = (z_{jk} : k = 0, \ldots, 2^j - 1)$ are Gaussian with non-singular covariance matrices (for details see Donoho (1995)). Under the current set-up $\varepsilon = \sigma n^{-1/2}$.

Let κ_j denote the pseudosingular values of the system defined as in *section 3.3* of the paper. Assuming $2^{-\nu j} A_1 \leqslant \kappa_j \leqslant 2^{-\nu j} A_2$ for constants $A_2 \geqslant A_1 > 0$, we define our penalized estimator for the vectors $\beta_j = (\beta_{jk} : k = 0, 1, \ldots, 2^j - 1)$ as

$$\tilde{\beta}_j = \arg\min_{\mu \in \mathbb{R}^{2^j}} \left\{ \sum_{k=0}^{2^j - 1} (\hat{\beta}_{jk} - \mu_k)^2 + \varepsilon^2 \kappa_j^{-2} \, P_j(\mu) \right\},$$

where $P_j(\mu)$ is the penalty function. Let C_{up} be the upper frame bound of the vaguelette basis $\mathcal{U}$. This implies that $C_{\mathrm{up}} \geqslant 1$ and satisfies $\mathcal{U}^* \mathcal{U} \leqslant C_{\mathrm{up}} I$. Using this we set the penalty function as $P_j(\mu) = N(\mu) \lambda^2_{j, N(\mu)}$ where $N(\mu)$ denotes the number of non-zero entries of μ and

$$\lambda_{j,l} = \zeta C_{\mathrm{up}} [1 + \sqrt{\{2(2\nu + 1) \log(\eta^{-1} l^{-1} 2^j)\}}], \qquad l = 1, \ldots, 2^j,$$

where $\zeta > \sqrt{2}$ and $0 < \eta < \frac{1}{2}$ are constants. The final estimate of f is

$$\hat{f}_n(t) = \sum_{k=0}^{2^{j_0}} \hat{\alpha}_{j_0k} \, \phi_{j_0k}(t) + \sum_{j=j_0}^{J} \sum_{k=0}^{2^j - 1} \tilde{\beta}_{jk} \, \psi_{jk}(t), \qquad \text{where } J = \log_2(n).$$

Using a concentration inequality for Gaussian random variables it can be shown that this levelwise penalized least squares scheme has the optimal risk bound over a wide range of Besov function classes, when measured in squared error loss. The details can be found in Johnstone and Paul (2004).

The following contributions were received in writing after the meeting.

Felix Abramovich (*Tel Aviv University*)

I congratulate Johnstone and his colleagues on their interesting paper. As they point out, the resulting estimator (17) essentially can be viewed as the wavelet–vaguelette decomposition (WVD) estimator for the deconvolution problem. The implementation is indeed different since the vaguelette coefficients $\hat{\beta}_k$ are not constructed explicitly but are evaluated instead in the Fourier domain. A natural alternative to the WVD is the *vaguelette–wavelet* decomposition (VWD) (Abramovich and Silverman, 1998). Whereas the WVD is based on the wavelet expansion of the unknown f that implies the corresponding vaguelette expansion of the observed data, in the VWD the latter is expanded explicitly in wavelet series.

The VWD estimator for the deconvolution problem can be described briefly as follows. Consider model (1) of the paper discussed:

$$Y_n(\mathrm{d}t) = f * g(t) \, \mathrm{d}t + \sigma n^{-1/2} \, W(\mathrm{d}t), \qquad t \in T = [0, 1].$$

Let K_g be the convolution operator with the kernel g, $h = K_g f = f * g$. Following the VWD approach we expand h (instead of f in the WVD) in the periodized Meyer wavelet series:

$$h = \sum_{k \in I_0} \alpha_k \Phi_k + \sum_{k \in I_1} \beta_k \Psi_k,$$

where $\alpha_k = \langle h, \Phi_k \rangle$ and $\beta_k = \langle h, \Psi_k \rangle$. Although for convenience I keep the same notation α_k and β_k for the scaling and wavelet coefficients as in the paper, the coefficients now are obviously different. The same will be true for the vaguelettes that are introduced below. Let $V_k = K^{-1} \Phi_k$ and $U_k = K_g^{-1} \Psi_k$. Similarly to the authors' arguments one can show that $||V_k||^2 \asymp 2^{2j_0 \nu}$ and $||U_k||^2 \asymp \tau_j^2 \asymp 2^{2j\nu}$. For smooth convolution kernels it can be verified that the system of normalized functions $v_k = 2^{-j_0 \nu} V_k$, $k \in I_0$, and $u_k = 2^{-j\nu} U_k$, $k \in I_1$, has vaguelette properties and generates a Riesz basis. The function f is then recovered by expanding in

vaguelette series as

$$f = \sum_{k \in I_0} 2^{j_0 \nu} \alpha_k v_k + \sum_{k \in I_1} 2^{j\nu} \beta_k u_k = \sum_{k \in I_0} \alpha_k V_k + \sum_{k \in I_1} \beta_k U_k.$$

The functions $U_k = K_g^{-1} \Phi_k$ and $V_k = K_g^{-1} \Psi_k$ can be evaluated in the Fourier domain where obviously $U_l^k = \Phi_l^k / g_l$, $V_l^k = \Psi_l^k / gl$ and, similarly to equation (47),

$$\Psi_l^k = 2^{-j/2} \, \hat{\psi}(2^{-j} \times 2\pi l) \, \exp(-2\pi \mathrm{i} l k \times 2^{-j}).$$

Given noisy data, the observed y is expanded in wavelet series with coefficients $\hat{\alpha}_k = \alpha_k + \sigma n^{-1/2} z_k$ and $\hat{\beta}_k = \beta_k + \sigma n^{-1/2} z_k$. Estimation of α_k and β_k from $\hat{\alpha}_k$ and $\hat{\beta}_k$ is done by a standard wavelet thresholding procedure (e.g. hard thresholding) with a properly chosen threshold λ and yields the VWD estimator $\hat{f}$ of the form

$$\hat{f} = \sum_{k \in I_0} \hat{\alpha}_k \mathbf{1}_{\{|\hat{\alpha}_k| \geqslant \lambda\}} V_k + \sum_{k \in I_1} \hat{\beta}_k \mathbf{1}_{\{|\hat{\beta}_k| \geqslant \lambda\}} U_k.$$

Following the arguments of Abramovich and Silverman (1998) and the paper, the optimal threshold should be $\lambda = \hat{\sigma} \sqrt{\{2(2\nu + 1) \log(n)/n\}}$.

The VWD estimator is essentially a *plug-in* estimator, where we first find a wavelet-based estimator of $K_g f$ and then apply the inverse operator K_g^{-1} to estimate f itself. In fact, using the previous remarks, $\hat{f}$ can be simply derived in the Fourier domain via its Fourier coefficients $\hat{f}_l$:

$$\hat{f}_l = (1/g_l) \sum_{k \in I_0} \hat{\alpha}_k \mathbf{1}_{\{|\hat{\alpha}_k| \geqslant \lambda\}} \Phi_l^k + (1/g_l) \sum_{k \in I_1} \hat{\beta}_k \mathbf{1}_{\{|\hat{\beta}_k| \geqslant \lambda\}} \Psi_l^k.$$

I believe that similarly to homogeneous operators it can be shown that both WVD and VWD estimators achieve the minimax convergence rates over the Besov classes that are considered in the paper. Nevertheless, it would be definitely interesting to understand when the VWD or the WVD estimator would be preferred in finite sample problems.

Robert G. Aykroyd, Robert M. West and Sha Meng (*University of Leeds*)

A major source of challenging inverse problems is tomography. Recently we have been working on electrical tomography in collaboration with members of the Virtual Centre for Industrial Process Tomography (`http://www.vcipt.org.uk`). Although the use of Bayesian modelling is now widely accepted in the statistical community there is substantial resistance among engineers working on inverse problems. So there is a great need for statisticians, in close collaboration with engineers, to develop simple, yet realistic and convincing, examples to highlight the flexibility of modern statistical approaches. Below we describe the mixing of liquids in a tank (West *et al*., 2003) with another example, movement of heart and lungs, appearing in West *et al*. (2004).

In electrical tomography, electrodes are attached to the boundary of an object and, while currents are applied, voltages are recorded. In contrast with many imaging applications the relationship between data and parameters is non-linear. This means that the calculation of voltages from a specified conductivity distribution, the forward problem, is numerically demanding. When coupled with Markov chain Monte Carlo methods computational efficiency is a key issue, and this may influence parameterization and modelling as well as the design of the algorithm. To deliver a practical solution of conductivity from voltages, the inverse problem, it is necessary to make considerable use of prior information. Although a pixel-based formulation provides a generic approach, it is not always the most useful. In particular, knowledge-based formulations allow parameter reduction and direct estimation of key quantities without the need for *ad hoc* post-processing.

A laboratory experiment was performed to investigate the mixing of liquids. At different stages prior knowledge regarding the spatial and temporal smoothness changes. During the first few frames the tank contained a homogeneous solution (tap-water) and the conductivity should be spatially and temporally very smooth. In the second stage a high conductivity saline solution was injected into the tank. In the spatial distribution we expect a conductivity discontinuity at the interface, with high smoothness elsewhere. As the saline injection begins to disperse there is moderate temporal and increasing spatial smoothness. As the process evolves the state is expected to change slowly and smoothly, which is modelled by using temporal priors. This example well illustrates the use of relevant prior information. In circumstances where

mixing is believed to be poor, such as the initialization of mixing, less reliance can be made on spatially smoothing priors. As mixing progresses the balance between spatial and temporal components is shifted.

Cristina Butucea (*Université Paris X*)

My comment on the paper by Johnstone and his colleagues addresses some developments related to deconvolution problems in a periodic setting. Their approach via wavelets is sparing in the case of the boxcar blurring operator $g_a(x) = k(x/a)$, where k is the uniform density over $[-1,1]$. Indeed, a deconvolution kernel K_h is strongly related to the boxcar blurring function g_a, via their Fourier transforms K_h^* and g^*, by the expression $K_h^*(u) = k(u)/g^*(u/h)$, where $h > 0$ is a tuning bandwidth. A strong condition for kernel deconvolution is that $g^* \neq 0$, whereas this is obviously not the case here.

Thus, wavelet deconvolution allows an extension of other existing results to this setting. To gain some rate order, we may ask only the question whether the underlying signal is some given function or, more generally, whether it belongs to some parametric subfamily of smoothness classes considered. The alternative can be defined in this case in the L_2-norm. The first step in the testing problem is to estimate the L_2-norm of a signal from noisy observations. Following the main lines of Butucea (2004), we consider for example estimation of $||f||_2^2$ ($p = 2$) via wavelet coefficients, in the dense case (condition (34)). We expect a bias of order 2^{-4js} and a variance of order $c_1/n + 2^{j+4j\nu}c_2/n^2$, where $\nu = 3/2$ for boxcar noise g_a. Thus we can attain parametric rates of convergence over sufficiently regular signals, i.e. $s > \nu + 1/4$ or $s > 7/4$.

Moreover, we can extend this easily to adaptive estimation of $||f||_2^2$, with quadratic rate

$$\{\sqrt{\log(n)/n}\}^{8s/\{4(s+\nu)+1\}}.$$

Turning to testing in the L_2-norm we expect that it is possible to establish intermediate results as in Butucea and Tribouley (2003) and Butucea (2004) and attain the minimax rate of testing from noisy data of $n^{-4s/\{4(s+\nu)+1\}}$. Moreover, the techniques in this paper allow generalization to a smoothness-adaptive procedure with noisy data attaining the rate

$$(\sqrt{[\log\{\log(n)\}]/n})^{4s/\{4(s+\nu)+1\}},$$

where the loss $\sqrt{\log\{\log(n)\}}$ is known to be unavoidable.

My question to the authors concerns estimation of general L_p-norms over Besov classes $B^s_{\pi,r}(T)$ compared with the direct observation scheme.

Laurent Cavalier (*Université Aix-Marseille 1, Marseille*)

The paper by Johnstone and his colleagues contains many interesting results and comments, e.g. the use of maxisets for inverse problems. However, I shall only discuss the specific case of boxcar deconvolution, which is a difficult model.

It appears that, owing to the term $\sin(\pi la)$ which appears in its Fourier series, some Fourier coefficients of the boxcar g can be 0. In this case, clearly, the corresponding coefficients of f cannot be recovered. To avoid this problem, 'badly approximable' irrational a are considered. This notion is of great mathematical interest, but from a statistical point of view the difference between a rational and an irrational number is perhaps less clear.

My next remark is linked to this irrational–rational phenomenon. In the paper, as very often in the inverse problems framework, the filter g is supposed to be known. However, it is clear that the filter can never be known, at least completely. There are references (see Efromovich and Koltchinskii (2001) and Cavalier and Hengartner (2003)) showing how to deal with such a problem. If we can suppose that the operator is observed with some noise, e.g. by using a training data set, then there is no real price to pay for the incomplete knowledge of the operator. Nevertheless, in this framework, the problem is perhaps different. Indeed, if we have a noisy boxcar, then is the difference between rational and irrational still fine to use? Otherwise, this would be a problem, which would mean that the framework is not very robust. The irrational–rational number set-up would then be less convenient in a statistical framework.

There is another approach to the problem without badly approximable numbers, which is to deal with it by changing the goal. Instead of trying to recover the whole function f, another idea is to reconstruct only the part of f which is not in the kernel of the operator, i.e. not the frequencies corresponding to null eigenvalues. This is a usual method in inverse problems, and it is linked to the notion of the Moore–Penrose (generalized) inverse (see Groetsh (1977)). In this case, it seems possible to obtain results, even if the problem is written differently.

Noel Cressie (*Ohio State University, Columbus*)
Haario and his colleagues start with an authoritative account of current numerical analytic methods for solving a high dimensional inverse problem for the 'Global ozone monitoring of occultation of stars' instrument on board the satellite Envisat. They go on to propose a statistical approach to the same problem based on Markov chain Monte Carlo (MCMC) sampling and an adaptive algorithm for the Metropolis–Hastings step in the MCMC algorithm. For their approach to work on massive data sets, certain assumptions appear to be crucial. Everything is Gaussian, everything is independent at each wavelength and height, and cross-sections do not depend on altitude. The first and second assumptions are probably not true for real data; for example, if the Gaussian distribution is actually an approximation to a Poisson distribution, should not the Gaussian variance be a function of the mean? The third assumption is a separability assumption that allows the problem to be split up into spectral inversion and vertical inversion. Without this assumption, the massiveness of the data overwhelms the MCMC algorithm. The paper assesses the methodology from simulated data, which I assume were simulated according to these three assumptions. How robust is the methodology when real data are involved? The answer to this question could be determined by simulating from a model that drops one or more of the three basic assumptions in favour of the physics underlying the instrument. The authors could then analyse these simulated data by using their methodology and compare their results with the truth. This would be a more convincing assessment of their inversion method.

Manuel Davy (*UMR, Nantes*)
I would like to congratulate Dr Wolfe, Dr Godsill and Mr Ng for their paper, which addresses a major issue in time–frequency analysis. There are indeed many possible applications of this work. Firstly, Gabor-style analysis is indeed especially efficient for an analysis of music, and the results that are presented in this paper are quite encouraging. In my own work, which is mainly aimed at the estimation of pitch, the frequency location of Gabor atoms is adjusted to the music signal that is analysed whereas in this work it is predetermined by the lattice settings. It is possible to position Gabor atoms on a very fine frequency grid, but, from the Heisenberg–Gabor inequality, this implies a weak time resolution. How is this issue dealt with in the current approach? Is it adapted to the estimation of pitch? Secondly, I find this approach quite useful when seeking time–frequency components for, for example, signal classification. Many references address this issue; however, this approach is one of the most convincing, together with that of Crouse *et al.* (1998) and that of Huang *et al.* (1998). There are many possible ways to define a 'time–frequency component', and this approach is especially powerful in that a given prior distribution over the indicator variables γ_k leads to a different characterization of time–frequency components. In particular, the Markov random-field prior implements a convincing mix between image processing techniques and time–frequency representations. Finally, I would like to underline the excellent adequacy of the selected Markov chain Monte Carlo sampling scheme with the posterior to be sampled from. Block sampling is clever, and my own results in Bayesian harmonic analysis confirm that it indeed enables rapid exploration of space.

Daniela De Canditiis (*Istituto per le Applicazioni del Calcolo "M. Picone", Rome*) **and Marianna Pensky** (*University of Central Florida, Orlando*)
First, we congratulate Johnstone, Kerkyacharian, Picard and Raimondo on the development of a truly non-linear fully adaptive deconvolution algorithm which delivers better precision than deconvolutions by Fourier or wavelet regularizations. The remarkable feature of the method is that it also works with kernels whose Fourier transforms vanish on the real line, e.g. the boxcar kernels which have the degree of ill-posedness $\nu = \frac{3}{2}$. The method is based on the fact that Fourier coefficients of the boxcar kernel do not vanish at frequencies $\pi k a$ when a is a badly approximated number.

The performance of the method can be further and significantly improved by the so-called multichannel convolution system approach which was pioneered by Casey and Walnut (1994) and adapted for statistical use by Pensky and Zayed (2002). We assume that the signal is convolved with functions $g_i(t)$, $i = 1, 2$, separately yielding two convolutions each of length n

$$Y_{ni}(\mathrm{d}t) = f * g_i(t)\,\mathrm{d}t + \sigma n^{-1/2}\, W_i(\mathrm{d}t), \qquad t \in T = [0, 1], \quad i = 1, 2.$$

The problem of solving the system for f is well posed if the Fourier transforms $\hat{g}_i(\omega)$ are entire functions on the complex ω-plane and

$$\sqrt{\{|\hat{g}_1(\omega)|^2 + |\hat{g}_2(\omega)|^2\}} \geqslant a\, \exp\{-2\pi b|\mathrm{Im}(\omega)|\}(1 + |\omega|)^{-N}$$

for some $a, b > 0$ and $N \geqslant 1$. Then, there are entire functions $\hat{v}_i(\omega), i = 1, 2,$ such that

$$\hat{g}_1(\omega)\,\hat{\nu}_1(\omega)+\hat{g}_2(\omega)\,\hat{\nu}_2(\omega)=1.$$

Hence,

$$(g_1*\nu_1)(t)+(g_2*\nu_2)(t)=\delta(t),$$

where $\delta(t)$ is the delta function. It follows that

$$f=(f*\delta)=f*\{(g_1*\nu_1)+(g_2*\nu_2)\}=(f*g_1)*\nu_1+(f*g_2)*\nu_2. \tag{4}$$

Finding the deconvolvers ν_i explicitly is a very difficult problem, yet, as a first approximation for $\hat{\nu}_i$, we can take the functions

$$\hat{\nu}_i(\omega)=\frac{\bar{\hat{g}}_i(\omega)}{|\hat{g}_1(\omega)|^2+|\hat{g}_2(\omega)|^2}.$$

To apply this technique in the set-up and notation of Johnstone and his colleagues let $g_{li}=\langle g_i,e_l\rangle$, $y_{li}=\langle Y_{ni},e_l\rangle$ and

$$\nu_{li}=\frac{\bar{g}_{li}}{|g_{l1}|^2+|g_{l2}|^2},\qquad i=1,2.$$

Using equation (4) we estimate wavelet coefficients β_k by

$$\hat{\beta}_k=\sum_l(y_{l1}\nu_{l1}+y_{l2}\nu_{l2})\bar{\Psi}_l^k=\sum_l\frac{y_{l1}\bar{g}_{l1}+y_{l2}\bar{g}_{l2}}{|g_{l1}|^2+|g_{l2}|^2}\bar{\Psi}_l^k.$$

This technique is especially advantageous when

$$g_i(t)=\frac{1}{2a_i}\,I(|t|<a_i),\qquad i=1,2,$$

are two boxcar functions with a_1 and a_2 being relatively prime badly approximated numbers.

Figs 3 and 4 show reconstructions based on one or two convolutions with boxcar functions. Although the number of observations is doubled in the case of one convolution, Figs 3 and 4 demonstrate a signifi-

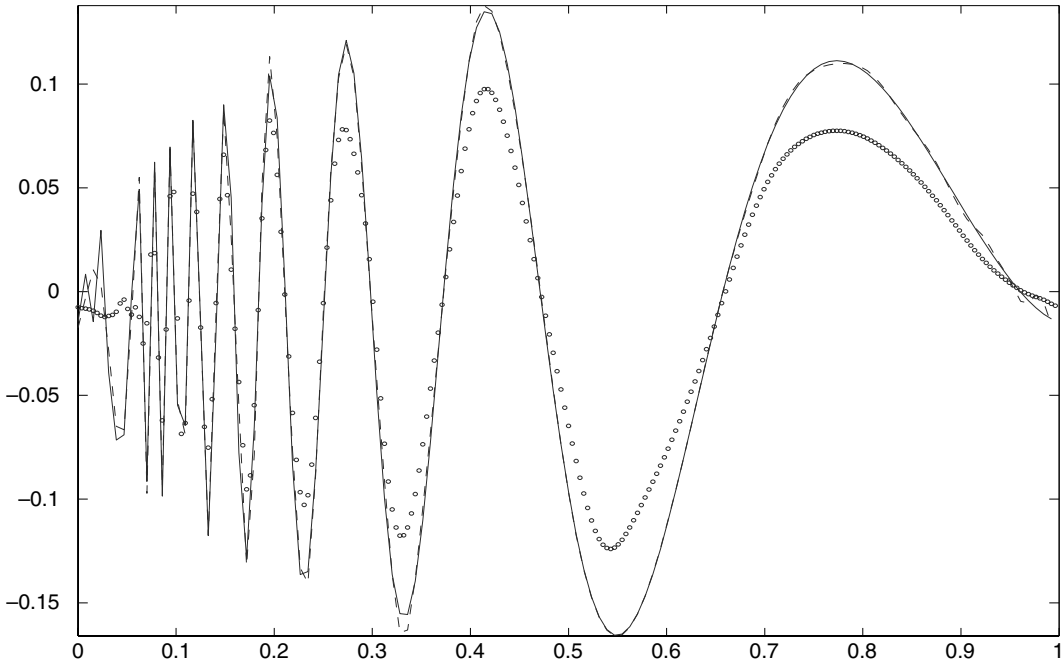

Fig. 3. Reconstruction of the Doppler signal based on one convolution ($a=1/\sqrt{37}$and $n=256$) and two convolutions ($a_1=1/\sqrt{37}$ and $a_2=1/\sqrt{23}$ and $n=128$): in both cases the signal-to-noise ratio is 1; the wavelet that is used is the periodized Meyer 3 wavelet (———, true; – – –, two channel; ∘, one channel)

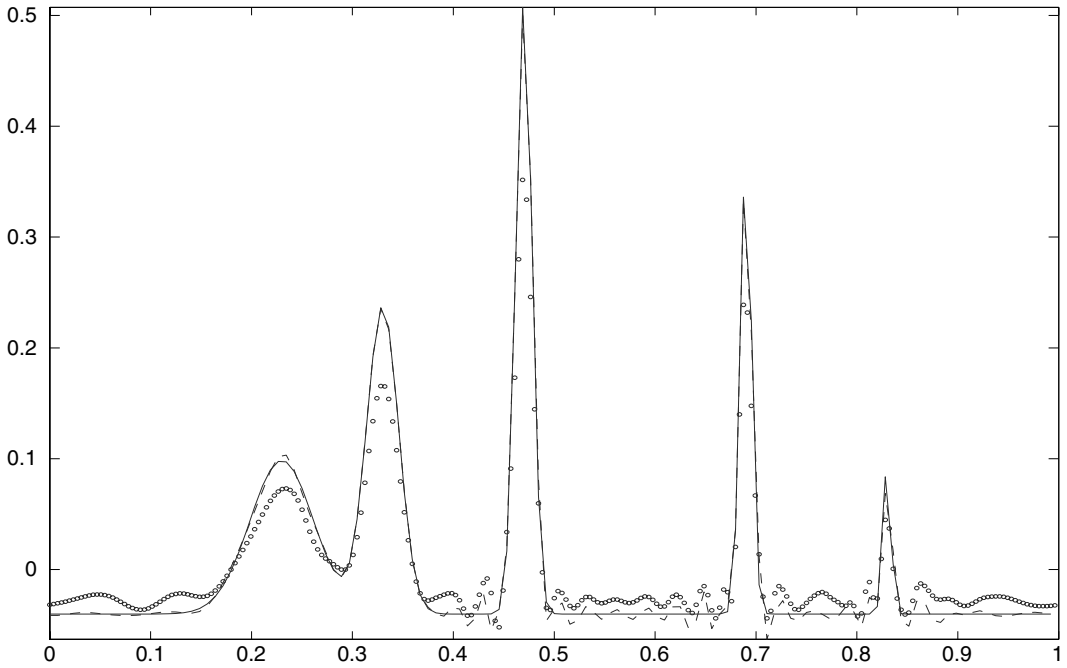

Fig. 4. Reconstruction of the 'Spikes' function based on one convolution ($a = 1/\sqrt{37}$ and $n = 256$) and two convolutions ($a_1 = 1/\sqrt{37}$, $a_2 = 1/\sqrt{23}$ and $n = 128$): in both cases the signal-to-noise ratio is 1; the wavelet that is used is the periodized Meyer 3 wavelet (———, true; – – –, two channel; ∘, one channel)

cant improvement over the performance of the multichannel technique. The one-channel reconstructions differ significantly from the true functions, whereas the two-channel reconstructions almost coincide with them.

Yu. Golubev (*Université Aix-Marseille 1, Marseille*)
The paper by Johnstone and his colleagues is a nice paper, which I enjoyed reading. It contains interesting new theoretical and practical results for solving inverse problems by a wavelet method. The goal of this short contribution is to draw attention to two statistical aspects of wavelet deconvolution that are related to large blur. To simplify some technical details, I shall consider an equivalent model of noisy blurred data in which we want to recover an unknown function f based on the noisy data

$$Z_n(t) = f(t) + \zeta_n(t),$$

where $\zeta_n(t)$ is coloured stationary Gaussian noise with spectral density

$$\lambda(\omega) = \frac{\sigma^2}{\eta\, G^2(\omega)}, \qquad G(\omega) = \int \exp(2\pi \mathrm{i}\omega t)\, g(t)\, \mathrm{d}t.$$

Wavelet basis
When a statistician decides to use wavelets, he has in mind that functional singularities like jumps or spikes may exist and they provide essential information about the function of interest. Therefore a good method of recovering a signal should preserve such singularities. In some sense this contradicts the fact that in deconvolution problems the spectral density $\lambda(\omega)$ may tend to ∞ very fast. It means that the noise at high resolution levels may be very large, and these levels cannot be used to recover f. It seems that this desperate situation could be slightly improved by the proper choice of a wavelet basis that meets two contradictory requirements:

(a) provide a reasonable representation of irregular functions by using a small number of decomposition levels;
(b) provide relatively small variances of the empirical wavelet coefficients.

The Meyer basis meets the second requirement perfectly since his scaling and wavelet functions are band limited. Undoubtedly this basis works well in situations when we want to recover smooth functions or when the blur is small. However, its properties in the time domain prevent good recovery of singularities. So at a glance we need a compromise solution depending on $\lambda(\omega)$. For instance, a properly chosen basis within the family of the Daubechies (1995) wavelets could do the work better in situations when the blur is statistically essential.

Shrinkage method

The fact that in wavelet deconvolution only a relatively small number of decomposition levels can be used requires more sophisticated methods of thresholding. The method of conservative thresholding that is used in the paper has evident advantages simplifying some computations. However, it leads evidently to oversmoothing which hides functional singularities. This effect is related to recovering sparse vectors when the number of non-negligible wavelet coefficients is not small. It seems to me that in this situation a good adaptive algorithm for choosing the threshold (see for instance Abramovich *et al.* (2000)) could improve the performance of the wavelet deconvolution.

Ross N. Hoffman (*Atmospheric and Environmental Research, Lexington*)

Extracting useful information from satellite scatterometer data is an important and intriguing problem. Although the direct use of the backscatter observations (usually denoted σ^0) in a global numerical weather prediction data assimilation system has been demonstrated by Thépaut *et al.* (1993), most practical uses of scatterometer data use a two-step procedure. In the first step, σ^0-data are used to retrieve multiple wind solutions (called ambiguities) at each cell. In the second step, spatial consistency and/or prior information are used to resolve the ambiguity by using methods that range from *ad hoc* median filters (Schultz, 1990; Shaffer *et al.*, 1991) to methods based on thin plate splines (Hoffman, 1984) or Karhunen–Loève models of the wind field (Draper and Long, 2002) to Bayesian-based variational methods in two (de Vries and Stoffelen, 2000) to four dimensions (Leidner *et al.*, 2003). In addition to computational efficiency, the two-step procedure allows limiting consideration in the second step to only the two leading (most likely) ambiguities. This limitation is important in practice since otherwise solutions corresponding to the third or fourth ambiguity are obtained too often. This occurs when an inaccurate background (i.e. an inaccurate prior from a previous forecast) is not close to either of the two leading ambiguities. In a recent extension of the method of Hoffman (1984), Hoffman *et al.* (2003) included the information from the third and fourth ambiguity in a second pass through the data. The results from the first pass provide an initial estimate for the second pass (however, the background is unchanged) and either ambiguous winds or σ^0-data are used. Henderson *et al.* (2003) have applied this new method to the nine months of data from the Nasa scatterometer mission.

Eliot Khabie-Zeitoune (*Computer Contract Consultants, London*)

I congratulate Haario and his colleagues for tackling high dimensionality. The simplifying assumption of a lack of correlation at wavelength and height may be argued, for reasons including scintillation and diffraction; see Robinson (1967) for a geophysical analogy.

Repetitive calculation of the likelihood of the line density vectors requires inversion of covariance matrices, with on-line recursive update. Should a more refined analysis use correlation, the question arises 'Why not update directly the inverse covariance matrix on line, as indicated below for instance?'.

Further, on-line calculation of the likelihood also involves two determinants. The first, that of the covariance matrix, is dealt with in Khabie-Zeitoune (1982). The second, the Jacobian of the transformation from line densities to transmissions, is derived from formula (3) of the paper under discussion.

An n-dimensional vector observation Y_t of zero expectation arises at inhomogeneous time occurrence t owing to a change in wavelength or line density.

K recent observations enter the computation of the sample covariance matrix S_t:

$$S_t = K^{-1} \sum_{k \in \{0, K-1\}} Y_{t-k} Y'_{t-k},$$

where K is either K or $K-1$. Assuming that S_t^{-1} has been computed, to update it for incoming Y_{t+1}, various rules are postulated.

(a) Only K observations are used at timet t: K is K or $K-1$. At $t+1, Y_{t+1}$ is added to, and Y_{t-K} is subtracted from, the sample,

$$S_{t+1} = S_t + K^{-1} Y_{t+1} Y'_{t+1} - K^{-1} Y_{t-K} Y'_{t-K}.$$

(b) All observations are used at time t: let t be t or $t-1$. At $t+1, Y_{t+1}$ is added to the sample,

$$S_{t+1} = (1+1/t) S_t + (t+1)^{-1} Y_{t+1} Y'_{t+1}.$$

Other rules are possible within the framework of adding an expression UV' to obtain S' from $S = S_t$, U and V being vectors such as

$$S' = S + UV'.$$

A formula attributed to Sherman-Morrison, and a related formula to Woodbury (see Householder (1964)), provides the inverse update:

$$S'^{-1} = S^{-1} - (1 + US^{-1}V')^{-1} (S^{-1}U)(S^{-1}V)'.$$

Define

$$\begin{aligned} \bar{S}_{t+1} &= S_t + K^{-1} Y_{t+1} Y'_{t+1}, \\ \bar{Z}_{t+1} &= S_t^{-1} Y_{t+1}, \\ Z_{t+1} &= \bar{S}_{t+1}^{-1} Y_{t+1}. \end{aligned}$$

Then we have the rule (a) inverse update

$$\begin{aligned} \bar{S}_{t+1}^{-1} &= S_t^{-1} - (K + X_{t+1} \bar{Z}'_{t+1})^{-1} \bar{Z}_{t+1} \bar{Z}'_{t+1}, \\ S_{t+1}^{-1} &= \bar{S}_{t+1}^{-1} + (K - X_{t+1} \bar{Z}'_{t+1})^{-1} \bar{Z}_{t+1} \bar{Z}'_{t+1} \end{aligned}$$

and the rule (b) inverse update

$$S_{t+1}^{-1} = S_t^{-1} - (t/(t+1) + Z_{t+1} \bar{Z}'_{t+1}) \bar{Z}_{t+1} \bar{Z}'_{t+1}.$$

The covariance matrix update for rule (a) and rule (b) requires $2(n^2+2)$ and n^2+4 multiplications respectively. The inverse update for rule (a) and rule (b) requires $2(n^2+n+3)$ and n^2+n+4 multiplications respectively.

A. Munk (*Göttingen University*) **and F. H. Ruymgaart** (*Texas Tech University, Lubbock*)

The beautiful paper by Johnstone and his colleagues appears to contain all the ingredients for exciting mathematics: an interesting practical example, an unexpected mathematical tool, a blend of statistics and approximation theory, and an optimal solution. In this comment we shall focus on convolution with the boxcar, but some remarks may remain true for other convolutions as well. Boxcar convolution has some peculiarities (Donoho, 1995) and standard spectral cut-off inversion does not seem to yield optimal convergence rates for the mean integrated squared error.

If the convolution is considered on the entire real line (i.e. in $L^2(\mathbb{R})$), rather than on the circle (in $L^2(\mathbb{T})$, say), the Fourier transform of the kernel is essentially the *sinc* function with infinitely many equally spaced 0s and in addition a 0 at ∞. A similar situation arises in $L^2(\mathbb{T})$, the case considered by the authors, in particular for rational end points. There is a remarkable difference, however, because in the latter case the set of 0s has infinite counting measure (which is the reason that the input cannot be fully recovered), whereas in the former the set of 0s has Lebesgue measure 0 (so that the convolution operator is still injective). To understand which of all these 0s is the actual source of ill-posedness we may solve the simple boxcar equation directly in the time domain. Under certain conditions on the input, the solution turns out to be a sum of shifted derivatives of the output, where apparently the derivative (corresponding to the 0 at ∞) causes the ill-posedness.

In fact, for many convolutions the inverses can be computed in the time domain (Zemanian (1987), chapter 8). This raises the question whether application of the Fourier transform could be avoided altogether in $L^2(\mathbb{T})$ as it was in $L^2(\mathbb{R})$ (Hall *et al.*, 2001), even though identifiability might cause problems and require restrictions on the input. More precisely, if K is the convolution operator and $e_1, e_2, \ldots$ any orthonormal basis of sufficiently smooth functions, we have $f = \Sigma_k \langle f, e_k \rangle e_k = \Sigma_k \langle K^{-1} g, e_k \rangle e_k = \Sigma_k \langle g, (K^{-1})^* e_k \rangle e_k$, where

the unbounded $(K^{-1})^*$ can be computed in the time domain and applied to the smooth basis elements. In Hall *et al.* (2001) this procedure was carried out for a wavelet basis and smooth f. For smooth f, wavelets are not needed, but the procedure would extend, for instance, to functions with discontinuities of the first kind as in Hall *et al.* (2003) for the Abel equation.

S. C. Olhede (*Imperial College London*)

I thank Dr Wolfe and his co-authors for their very interesting contribution to time–frequency analysis.

Note that the plotting of time–frequency concentration has inherent thresholding if contours are used, making the choice of contours important. The time–frequency plots of the signal and its noisy version are represented in terms of their Gabor coefficients with an inherent magnitude, whereas the time–frequency representation based on indicator estimates has no direct magnitude attached to it. Nor can numerical comparisons be made between the signal's noise-free representation and the indicator estimates based on the noisy signal. I would also like to note that a Bayesian formulation of the denoising problem for images and the posterior probability of inclusion have been treated by Malfait and Roose (1997), for wavelet-based frames. Another point which perhaps merits further discussion is the inherent bandwidth of the analysis. Certainly it is known that the bandwidth will have a large effect on the analysis.

On page 578 is a nice suggestion for improving the spectrogram by using ideas of sparse representations and unsmoothing the blurring: followed by the actual time–frequency modelling, that should contain any pre-set notions of time–frequency behaviour. Wolfe and his colleagues suggest

(a) a Bernoulli prior,
(b) a Markov chain prior in time and
(c) a Markov random-field prior

Suggestion (a) seems to need additional continuity constraints, (b) smears the time–frequency contours in time, chopping up the curved structure in frequency, and (c) smears the pictures in all directions, rather than unsmooths. These comments are borne out by the reconstructions in Fig. 3 of the paper. Perhaps consideration could be taken of the fact that time and frequency operators do not commute, rather than treating the concentration as a more arbitrary two-dimensional density, without limiting the range of application.

Finally, with regard to the denoising, much theory has been developed in this area, based on time–frequency ideas and frames ('cycle spinning'). Some comparison in the discussion with standard method performance with more information on when the method is expected to perform better would have aided me. Brevity permitting, plots of signal reconstructions would also be of further interest.

I would also like to reiterate my thanks to the authors.

A. Tsybakov (*Université Paris 6*)

The paper by Johnstone and his colleagues contains many interesting and novel ideas from both theoretical and applied points of view. Several theoretical aspects of the paper are innovative: the approach based on combining Fourier and wavelet bases, simultaneous analysis of the behaviour of the estimators in all L_p-norms with $1 < p < \infty$, the maxiset paradigm applied to deconvolution, etc. The results are based on elegant techniques developed in appendix B. In particular, the Temlyakov property for wavelet bases is a beautiful tool that presents independent interest. How can this be applied to more general inverse problems than deconvolution and how crucial is the periodicity constraint and the particular choice of mixed Fourier–wavelet bases? My intuition is that an extension to other problems might be fruitful, whereas I am not convinced that for the periodic setting the use of a dyadic wavelet structure is the best way to obtain optimal estimators. A disadvantage is that dyadic blocks are well tailored for Besov spaces and not so well adapted to other spaces, e.g. to Sobolev spaces. For the L_2-setting of periodic deconvolution, another approach consists in using smaller and more flexible blocks of logarithmic size or weakly geometrical blocks and to threshold over blocks, since this leads not only to optimal rates but also to asymptotically optimal constants (Cavalier and Tsybakov, 2001, 2002). Small blocks work well for both Besov and Sobolev scales. The general L_p-setting is different because there is no hope of obtaining constants, but giving more flexibility to blocks might be good.

A problem with the L_p-results ($p \neq 2$) is a gap between theory and applications. The theoretical thresholds depend on p for $p > 2$: thus they can be effectively applied only in the range $1 < p \leqslant 2$. But this also seems to be bad: the simulations advocate the *ad hoc* threshold constant $\eta = \sqrt{2}$ which is considerably smaller than the constant that is given by theory. In fact, the simulations are done for $p = 2$ in which case smaller thresholds than in proposition 1 would work, so it remains unclear how we can benefit from

general L_p-results. Another point is near optimality of the rates. To my knowledge, there are no lower bounds showing that the rates of proposition 1 are optimal or near optimal: although the conjecture is folkloric, it would be useful to prove it formally.

Grace Wahba (*University of Wisconsin, Madison*)
We thank Johnstone and his colleagues for some very interesting modern results in deconvolution theory with wavelets.

There was in the late 1980s an argument going on about whether a predictive mean-square error criterion such as generalized cross-validation for choosing the regularization parameter was good for minimizing the solution mean-square error, in smooth deconvolution problems solved by Tihonov regularization.

Wahba and Wang (1990) answered that question as 'sometimes yes and sometimes no, depending on conditions on four parameters: the rate of decay of the Fourier coefficients of the solution, of the convolution kernel, of the coefficients in the penalty functional, and in the loss function'. Convergence rates are also given (available via the 'TRLIST/golden oldies' link at `http://www.stat.wisc.edu/~wahba`).

The **authors** replied later, in writing, as follows.

Iain M. Johnstone, Gérard Kerkyacharian, Dominique Picard and Marc Raimondo
We thank the discussants for their very interesting comments and suggestions. Since many issues raised in the discussion are related, we shall address them by topic.

The periodic setting and Meyer wavelets
The WaveD estimator exploits the natural representation of the convolution operator in the Fourier domain as well as the typical characterization of Besov classes in the wavelet domain. In practice, we may have to deal with edge effects for non-periodic signals. As suggested in the discussion by E. Moulines, there are practical methods (time or space domain preprocessing) to do so; see for example Aghdasi and Ward (1996). An alternative approach would be to choose a compactly supported wavelet family whose Fourier transforms vanish rapidly (e.g. Daubechies wavelets). This, of course, would introduce a bias in the WaveD paradigm (22) but would allow us to relax the periodicity assumption. This approach further extends to the discussion on bias–variance trade-off suggested by Golubev. However, in a recent study on translation invariant deconvolution, Donoho and Raimondo (2004) show that the WaveD estimator (with band-limited wavelets) outperformed the ForWaRD method (with compactly supported wavelets) over a wide range of test functions (larger margins were observed for smooth signals). These numerical results suggest that the bias that is introduced in the Fourier domain by using compactly supported wavelets may compromise the gain in signal representation in the time domain. Of course, further analysis should be made to assess the 'edge effect' fully for non-periodic signals.

Refinements and extensions of the WaveD method
In real applications the convolution kernel g may not be known completely, which renders the estimation of f even more delicate. This is the so-called *blind deconvolution* problem. In many practical situations, however, the investigator has some prior knowledge about the convolution operator which may be integrated in the model following a Bayesian method; see for example Hopgood and Rayner (1999). Recent approaches to blind deconvolution include Efroimovich and Koltchinskii (2001) and Cavalier and Hengartner (2003). It would be of great interest to derive a blind deconvolution (WaveD) algorithm adapting some of the previously cited works. This project is currently under investigation by us. Another exciting extension of the WaveD estimator is, of course, the development of a two-dimensional WaveD algorithm. The two-dimensional setting is particularly important for the convolution operator since there are many applications to digital image processing. For example, the boxcar kernel plays a central role in modelling motion blur (Bertero and Boccacci, 1998). Given the good performance of WaveD in the one-dimensional setting we expect some interesting results in two dimensions, even while bearing in mind some fundamental limitations to wavelet representations of images (Candès and Donoho, 2004).

Finally, the WaveD estimator is relatively simple, being based on hard thresholding, with a deterministic (as opposed to data-adaptive) choice of thresholds at each level. Although it enjoys good numerical properties, in particular when using the more recent translation invariant algorithm of Donoho and Raimondo (2004), more sophisticated thresholding algorithms may improve the performance of the WaveD estimator. This point emerges nicely from several discussion contributions. De Canditiis and Pensky consider a model in which more data are available, through several convolutions, and obtain promising results by using a multichannel approach (Pensky and Zayed, 2002).

Numerical comparison of the WaveD method with the vaguelette–wavelet decomposition variant suggested by Abramovich would be of interest—the distinction between choosing thresholds for Abramowich's 'image domain' coefficents of $f * g$ *versus* for our 'object domain' coefficients of f is reminiscent of the issues in regularization parameter selection for spline-like estimators that are helpfully recalled by Wahba. Our *a priori* expectation is that there is advantage to exploiting sparsity of representation, which is usually more pronounced in the object domain.

Golubev proposes the use of more adaptive or data-dependent choices of thresholds, such as would be suggested by the FDR principle. We are enthusiastic about such a possibility—the contribution of Paul is in this spirit (though a little easier to treat theoretically) and leads to exactly optimal rates of convergence in cases where the method of our paper has to give up logarithmic terms.

Other proposals with interesting possible implications for WaveD include the use of block thresholding (Cavalier and Tsybakov, 2001, 2002) and complex-valued wavelet shrinkage (Barber and Nason, 2004). Further investigation of properties and numerical performance of all these ideas are good topics for future work.

Near optimality and other rate issues

As noted in section 3.3's discussion of the connection with the wavelet–vaguelette decomposition, lower bounds can be derived over Besov classes following the method of Donoho (1995) for convolution operators which satisfy condition (27) in the paper; this includes smooth convolutions but not the boxcar kernel. To our knowledge, the only lower bounds that are available for this kernel are those in Johnstone and Raimondo (2004) derived for ellipsoids and hyper-rectangles.

Many contributors (Cavalier, Moulines, Munk and Ruymgaart, Tsybakov, ...) question our choice of a wavelet basis instead of other orthonormal bases and especially the Fourier basis with its fundamental property of diagonalizing the convolution operator. The advantage of wavelet bases lies in their ability to yield a sparse representation of the unknown coefficients whether the function has isolated singularities or is smooth. In more mathematical terms, wavelet bases fit well with L_p-norms, Besov constraints as well as Sobolev spaces (except perhaps for the theoretically very challenging problem of constructing minimax procedures up to constants as suggested by Tsybakov).

Obviously, as observed by Nason, this leads to the first difficulty that the estimated wavelet coefficients are correlated. But, fortunately, because of the concentration of wavelet bases, this has relatively mild consequences for the thresholding procedures. Thresholding of the (independent) Fourier coefficients before inversion was suggested by several contributors. This natural idea may, however, introduce greater bias, as the Fourier space representation of $g*f$ may not be well concentrated. The idea, however, seems more successful if applied to blocks of coefficients, as in Cavalier and Tsybakov (2001, 2002).

We are glad to note Butucea's suggestion that the method presented here could be exploited to construct optimal smoothness adaptive procedures for testing in L_2-norm in the deconvolution setting. As for Butucea's question about estimation of L_p-norms in this setting, although there may be some hints, we have no complete solution. It is a very challenging question. We are also optimistic that our procedure could perform quite well in the interesting framework suggested by Cavalier, and referring to the case where the operator is known with noise (or observed in training data).

We agree with Tsybakov that simulations comparisons using L_∞- and more general L_p-norms should be performed. In this case, there is an obvious general 'trend' towards L_2-comparisons. Since many procedures are theoretically constructed for L_2-norms, researchers generally do perform their simulation by using this norm. Hence, if we want to present a quite fair comparison, we need to obey the same rules. However, it is definitely an advantage of our method to behave quite well practically with L_2-comparisons, *in addition to* having optimal theoretical properties for a variety of other norms.

Issues for boxcar blur

Some contributors are (perhaps understandably) puzzled that practical implications of the WaveD method should or even could be affected by the irrationality properties of the boxcar width a. Part of the difficulty arises because simple models are inevitably idealizations and may excessively sharpen distinctions—such as whether a number is 'badly approximable' (BA)—that are blurred in real data settings. If we 'recomplicate' the model, such distinctions can disappear. Thus, for example, as mentioned in section 2.2, in the finite sample implementation of the boxcar blur certain kernels with rational width are permitted. Examples include boxcars whose width α are convergents $\alpha = p_k/q_k$ of a BA number a, where k is sufficiently large that there are no 0s in the Fourier coefficients of g up to indices equalling the sample size. In addition whether the (known) boxcar width is under the investigator's control or simply determined by nature need not in certain cases be an issue as detailed in the discussion section of Johnstone and Raimondo (2004).

For statistical applications, an important issue is whether the WaveD estimator is robust against departure from the BA assumption. Although the set of BA numbers contains quadratic irrationals like $\sqrt{5}$, it has Lebesgue measure 0. Following the arguments of section 2.2 the robustness problem for WaveD may be stated as follows: how do 'most' numbers behave with respect to rational or irrational approximations? The answer may be found in a theorem of Khinchin (1997) (chapter 2): for each $\delta > 0$, there is a set A_δ of full Lebesgue measure such that, for all $a \in A_\delta$, $q_{n+1} \leqslant q_n \log(n)^{1+\delta}$. In words, for *almost all* numbers, the geometric growth of the convergent denominators is only a log-term faster than for BA numbers. In fact, for ellipsoid function classes, Johnstone and Raimondo (2004) have shown that, as long as log-terms are ignored, for almost all numbers the boxcar blur has the same degree of ill-posedness, namely $\frac{3}{2}$. Adaptive performance of the WaveD estimator over Besov classes outside the BA case is studied in Kerkyacharian *et al*. (2004), in which the maxiset approach of Kerkyacharian and Picard (2000) is used to show that for almost all numbers the WaveD estimator achieves near optimal rates (with a degree of ill-posedness of $\frac{3}{2}$) on slightly different Besov scales than that presented in the paper.

Cavalier and Munk and Ruymgaart observe that, if we change the observation model, it is possible and perhaps desirable to deal with boxcar convolution without the introduction of BA numbers. For example, if in the density model f has compact support, Hall *et al*. (2001) have given a reconstruction formula in the time domain which does not impose any condition on the scale of the boxcar. The compact support assumption is key: identifiability issues would emerge in extending this approach to f with non-compact support or to the periodic setting. Further, it is not yet clear whether such an approach extends to adaptive estimators though this would be an interesting problem to investigate. In contrast, our number theoretical approach to the boxcar blur yields a fast adaptive algorithm and extends the wavelet–vaguelette decomposition paradigm of Donoho (1995), at least for boxcar blur with BA-like scale.

Patrick J. Wolfe, Simon J. Godsill and Wee-Jing Ng

We thank the discussants for raising several pertinent points regarding our contribution. Some we had considered initially but sacrificed for brevity; all we feel point towards exciting future directions for this line of study. Our intention was to strike a balance between introducing general methodology with the potential for widespread application in the context of overcomplete representations and outlining a specific example of how one might exploit prior knowledge in the time–frequency domain to formulate generative statistical models for audio time series.

Regarding the former of these aims, we note that the choice of (frame) representation that is employed will necessarily depend on the class of signals under study. Hence it would not be prudent to recommend the use of Gabor frames *per se* for all applications, but rather only for those which are well characterized by the idea of underspread operators as described in the paper. Our intention was to provide a principled statistical approach to choosing from among an overcomplete 'dictionary' of frame elements, be it for the task of signal enhancement, compression or modification. The design of such a dictionary is itself a separate issue; different frames will cover the time–frequency plane in different manners and with varying amounts of overlap.

With regard to our latter aim, we have also attempted to open novel avenues for audio signal processing, as Gabor analysis provides appropriate methodology for investigating (relatively) slowly time-varying phenomena such as speech and music. In fact, it is common engineering practice to implement an overlap–add method of Gabor analysis (corresponding to a diagonalized frame operator) in which the phase factors resulting from the non-commutativity of the translation and modulation operators are 'absorbed' into the phase of the Gabor coefficients (Dörfler, 2001). The resultant mathematical structure enables an efficient generation of the Gabor system and makes the implementation of Markov chain Monte Carlo methods feasible in cases where the data rate often exceeds 10 000 samples per second. Although computational efficiency remains an issue, we believe that such work may readily lead to faster suboptimal algorithms based on a combination of stochastic and deterministic methods.

Prior dependence in the time–frequency plane

A vital component of our modelling strategy is the prior structure on the Gabor regression coefficients $\{c_k\}$. In the models that we consider, these coefficients are modelled as conditionally mutually independent, zero-mean, complex Gaussian random variables. Such a structure allows for heavy-tailed behaviour of the coefficient process through the introduction of associated unknown and varying variance parameters $\{\sigma^2_{c_k}\}$. This type of heavy-tailed model is considered highly desirable for capturing the characteristics of many naturally occurring processes, including the audio examples that we describe. The priors that are evaluated here model sparseness directly by allowing non-zero probability mass at zero, expressed in a

standard way through the use of indicator variables $\{\gamma_k\}$. Prior dependence is then introduced between neighbouring indicator variables via Markov random-field models.

This choice of priors provides a potentially powerful approach, as it directly encapsulates the idea that non-zero coefficients tend to cluster in certain regions of the time–frequency plane, often surrounded by regions of 'inactivity' where there is little or no energy. Such a property has been observed in many naturally occurring signals, including audio time series and images. Here we have introduced a flexible and general means of modelling this property, using either Markov-in-time, Markov-in-frequency or Markov-in-time–frequency models. In the case of audio and speech data, time dependence is expected during stationary 'pitched' sounds such as musical notes or spoken vowels, whereas dependence in frequency is expected at note onsets ('attacks'), consonants or fricatives. In this paper we have explored only a few possibilities from this very general class of priors; no doubt some combination of all three types of dependence would provide the most realistic model.

In terms of the signal extraction performance for audio noise reduction or coding purposes, these priors have been observed to reduce the so-called 'musical noise' artefacts that are often associated with the suppression of noise. Such artefacts can arise through erroneous fitting of time–frequency components to noise; by favouring connected regions of non-zero coefficients, it becomes much less probable that an isolated peak in the noise spectrum will mistakenly be modelled as a signal component. As with more heuristic noise suppression schemes, a trade-off between oversmoothing of the extracted signal and the production of musical noise artefacts is observed (Wolfe and Godsill, 2003). One may think of our approach as providing a very sensitive and accurate segmentation of the time–frequency space into 'active' and 'inactive' regions—a task which is itself of interest in speech and audio processing.

Evolution of time–frequency coefficients

The discussants have rightly commented that our prior structures do not directly model the dependence of scale that is expected within 'active' regions of the time–frequency plane (i.e. those regions containing non-zero coefficients); nor do they directly model the amplitude dependence of the coefficients themselves. Such dependences undoubtedly exist; however, for general signal modelling applications it may be difficult to specify how these priors should operate. We regard the procedures that are reported here as a 'base-line' prior that captures a salient feature of many data sets, without being overly specific to a particular application. For the case of audio signals—or in other specific applications—it may be possible to construct more effective prior structures that directly model dependences of the variance components $\{\sigma_{c_k}^2\}$, or even of the coefficient values themselves.

Indeed, in earlier work (Ng, 2000) we have experimented with a second layer of Markov random-field modelling, this time applied to the continuous-valued variance parameters $\{\sigma_{c_k}^2\}$. The results of this approach are promising but require further development before publication. Clearly the specification of the structure and parameters of such a prior becomes a much more complex issue than in the relatively simple case of the indicator variable field. The potential benefits of such an approach would be further reduction in musical noise artefacts for audio, and greater signal fidelity in reconstructions.

Structure of time–frequency atoms

Another important point that is raised by the discussants is the use of more structured time–frequency 'building-blocks' for the specific case of sound signal modelling. Many speech and music waveforms are 'quasi-periodic' over a short timescale, having (approximately) regularly spaced harmonic components at integer multiples of a fundamental frequency. In this case, why not construct time–frequency atoms that model this effect directly? Such ideas could certainly be adopted to improve the modelling capabilities of our generic approach, with a necessary loss of generality in the modelling. This idea has in fact been extensively explored within the field of polyphonic musical pitch modelling (see, for example, Davy and Godsill (2003) and references therein); similar methods may also be applied to speech.

The approach that was taken by Davy and Godsill (2003) is to model musical notes as atoms with an unknown number of harmonics at fixed multiples of unknown fundamental frequencies, in contrast with the less computationally intensive approach that was described here in which a fixed sampling of the time–frequency plane is adopted. Bayesian hierarchical models and associated variable dimension Markov chain Monte Carlo methods are developed for the estimation of the highly complex models that arise in this setting; these approaches can be compared with the projection-based methods of Gribonval and Bacry (2003). However, as the more restrictive models of Davy and Godsill (2003) are aimed at extracting high level features such as the pitch and timbre of musical notes, their noise reduction performance is not of quality comparable with that described in our contribution. Nevertheless, they certainly reinforce

the discussants' (and our own) view that harmonic time–frequency components hold further potential for improvement in audio signal processing applications.

H. Haario, M. Laine, M. Lehtinen, E. Saksman and J. Tamminen

We thank all the discussants for their contributions and valuable remarks on the topic.

The paper discusses the inversion of satellite measurements, but the emphasis is more on the methodological side. The validation of the assumptions and approximations that are employed against emerging real data is subject to on-going research. Robert found the prior input to be vague. The lack of more detailed prior input should be seen against the background that such geophysicl information simply does not exist at this point. Actually, the 'Global ozone monitoring of occultation of stars' (GOMOS) instrument is among the first instruments to yield global measurements of the atmospheric gas profiles that strongly depend on time and location. We do, indeed, present an altitude-dependent way for regularization (see expression (12)).

Cressie questions the validity of the independent Gaussian approximation for the measurement noise. This topic is at least partially dealt with in section 2.2 of our paper. The variance depends on the wavelength and the line of sight, and indeed takes into account the Poisson-type character of the measurements. A repetitive inversion of the noise covariance matrix would be needed if the noise structure is to be estimated together with the model parameters. In that case formulae like those presented by Khabie-Zeitoune should be employed. Here the noise is supposed to be known, as it can be estimated on line for each star from the data. So no repetitive inversion of the covariance matrix is done.

Contrary to what Cressie suspects, the Markov chain Monte Carlo (MCMC) runs also can be performed without the separability assumptions; they only are required by the fast parallel MCMC runs. In fact, this is what we demonstrate with the SCAM algorithm for the 'one-step' version of the inversion. It is true that the examples in the paper were done with simulated data. But the methods described have already been successfully applied with real data to the GOMOS spectral inversion. As might be expected, certain non-idealities have turned up, and the more general one-step inversion is a potential approach for more detailed off-line processing. The first results for validation with real data will be described elsewhere (Tamminen, 2004).

The use (or non-use) of convergence criteria is questioned by Robert. Here we only tested the high dimensional SCAM runs by comparisons with the results from the parallel runs, thoroughly validated in earlier studies. They did, indeed, virtually coincide. Naturally one could employ several other convergence criteria.

Robert also raises the question about the practical relevance of theoretical ergodicity results. Obviously we should use methods that are known to produce correct results. This entails two things: firstly, we should know theoretically that the method converges. Equally important are exhaustive tests to measure the robustness of the method and the practical implementations. What comes from the examples of Robert is that it appears to us that his Fig. 2 tells more about the well-known sensitivity of the independence sampler on the proposal distribution than about adaptivity (note that this example corresponds to an adaptive independence sampler, not to the AM algorithm). Even a fairly reasonable non-adaptive, theoretically correct independence sampler can yield results that are comparable with those in Fig. 2. One should keep in mind that theoretically correct but ill-tuned algorithms may yield incorrect runs, for both adaptive and non-adaptive MCMC sampling. As for the example in Robert's Fig. 1, it also demonstrates the robustness of the AM algorithm with respect to the parameter ε.

Both Robert and Andrieu ask why we used just the AM algorithm or its variant, the SCAM algorithm, and why not adapt in a more refined or complicated manner? A partial answer is simple: the algorithms that we used were quite satisfactory for our needs here.

However, there is nothing to prevent us from using other schemes. We refer to the interesting work of Andrieu and Robert (2001) that generalizes the non-Markovian adaptation that is introduced in Haario *et al.* (2001) to a general framework which allows adaptation of various selected parameters. We also mention Haario *et al.* (2003), where we apply mixtures of transition probabilities by combining the delayed rejection and AM algorithms. However, intricate adaptation schemes often are less automatic in practical implementation and may possibly slow down the convergence. Andrieu presents the important observation that it is possible to estimate the effect of adaptation on the convergence. Especially, in the AM algorithm, the error that is caused by adaptation theoretically decays quicker than the unavoidable Monte Carlo fluctuations (see Andrieu and Moulines (2002)).

Finally, we comment on the SCAM algorithm and, more generally, on high dimensional MCMC methods. We first note that SCAM provides no miracles as it works no better than a well-tuned single-compo-

nent Metropolis algorithm in the corresponding situation. In many cases, the difference in the performance between the SCAM and AM algorithms is not so large, taking into account the total number of target function evaluations that are needed. The advantages that are provided by the SCAM algorithm seem to stem from the simplicity of the adaptation scheme. The adaptation of the one-dimensional variances is also computationally cheap if compared with sampling from high dimensional Gaussian proposals as is done in the AM algorithm. Here the co-ordinates were systematically scanned. Our experience so far indicates that the more complicated covariance update of the AM algorithm may increase the side-effects of adaptation in high dimensions. This needs more testing, however. In general, our opinion is that the overall knowledge of performance of MCMC sampling in high dimensions, whether adaptive or not, is rather limited at the moment. Since many applications, various inverse problems as typical examples, lead to high dimensional problems, we feel that real understanding of high dimensions poses an interesting challenge to computational statistics.

Dan Cornford, Lehel Csató, David J. Evans and Manfred Opper

We start with an expression of thanks to the organizers and the proposer and seconder of the paper, who have done an excellent job on what is rather a difficult paper to read. Before answering the individual points we would like to reiterate the aim of our work: to provide a principled, yet operationally feasible, probabilistic method to solve the inverse problem of scatterometer wind retrieval. We note that almost all previous work on the scatterometer inverse problem, such as cited by Ad Stoffelen and Ross Hoffman, are either *ad hoc* or seek a maximum-likelihood-like solution, the exception that we are aware of being Royle *et al.* (1998) which constructs a hierarchical Bayesian model for wind fields but does not treat the inverse problem itself. Our aim is not to compare the Markov chain Monte Carlo (MCMC) approach with our variational approximation, but rather to contrast the applicability of the two methods and to *work towards* an operational, statistically rigorous retrieval. The effect of our work on practical scatterometry will only be apparent once data assimilation methods become truly probabilistic, and seek to estimate the *distribution* of the state of the atmosphere, not just the most probable value.

Christian Robert makes the observation that, for the mixture models that we employ, the posterior can be computed analytically—this is true, but there is an exponential growth in complexity as the number of observations increases, making this impractical beyond around 50 observations. We were a little surprised by the comment about the absence of prior information; our priors, both the 'climatological' zero-mean prior and the numerical weather-prediction-derived 'dynamical' prior, use a considerable amount of historical data and meteorological expertise to define them. Their role is important in the retrieval: the imposition of an approximately non-divergent flow, with energy at certain scales, provides extra information to the posterior distribution. This is particularly true when applying the dynamical prior (we refer to this as 'data assimilation' in the paper) where the (non-zero) mean in the prior imparts a different, approximately unimodal, structure to the posterior distribution. Referring to section 4.1 we agree that population Monte Carlo methods (Cappé *et al.*, 2004) look very promising for this problem and intend to follow up this useful suggestion. Although the 'global' variational approximation to the posterior in section 4.2 does indeed provide only the first two moments, we would like to stress that 'local' marginal posteriors can be computed with non-trivial (e.g. multimodal) structure using our framework. We agree that the assessment of the quality of the approximation could have been more thorough: it is on our to-do list. As for convergence, although we cannot guarantee convergence we can prove the existence of a cost function with local minima agreeing with the fixed points of the learning dynamics (Csató *et al.*, 2002).

Christophe Andrieu starts with a very clear analysis of the issues (both with the paper and our inverse problem). We note that the scatterometer observation is composed of several repeated measurements of the back-scattered signal, which are averaged for the observation, and used to compute an estimate of the variance (signal-to-noise ratio). We do indeed treat these as fixed, since we believe that we have reasonable estimates from earlier work (Bullen *et al.*, 2003). Although we could set up a hierarchical model and jointly estimate the posterior distribution of the 'hyperparameters' and wind vectors, we would be concerned about the identifiability of the hyperparameters. On the Gaussian process side we were aware of the work of Jones and Vecchia (1993) which could provide an interesting framework but would require substantial work to implement. We agree that the Gaussian process prior that we are using is strongly related to the class of models they, and others before and after them, proposed. It would indeed be interesting to exploit the nearest neighbour approximations that were used in Jones and Vecchia (1993) to develop sparse approximations to the inverse covariance matrix in our equation (2). The suggestion to explore the possibility of alternative MCMC algorithms is relevant, but we feel that the most interesting suggestion is to combine the variational method with an MCMC sampler (e.g. as a proposal distribution

for a population Monte Carlo method, or importance sampler). There are several issues which might need to be addressed to implement such a scheme, such as the observation that is made by Christian Robert that, often, variational approximations might not cover the full support of the posterior distribution and thus might require some inflation before being used in a sampling context. We look forward to exploring this, together with the use of other sampling methods, to provide a more rigorous comparison with the variational approximation.

We conclude by thanking again the organizers and discussants—they have given us food for thought, and some interesting suggestions for extensions to the work.

References in the discussion

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000) Adapting to unknown sparsity by controlling false discovery rate. *Technical Report 2000-19*. Department of Statistics, Stanford University, Stanford.

Abramovich, F. and Silverman, B. (1998) Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, **85**, 115–129.

Aghdasi, F. and Ward, R. (1996) Reduction of boundary artifacts in image restoration. *IEEE Trans. Image Process.*, **5**, 611–618.

Andrieu, C. and Moulines, E. (2002) On the ergodicity properties of some adaptive MCMC algorithms. *Technical Report*. University of Bristol, Bristol.

Andrieu, C. and Robert, C. P. (2001) Controlled MCMC for optimal sampling. *Preprint*.

Barber, S. and Nason, G. (2004) Real nonparametric regression using complex wavelets. *J. R. Statist. Soc.* B, to be published.

Bertero, M. and Boccacci, P. (1998) *Introduction to Inverse Problems in Imaging*. Philadelphia: Institute of Physics.

Bullen, R. J., Cornford, D. and Nabney, I. T. (2003) Outlier detection in scatterometer data: neural network approaches. *Neur. Netwrks*, **16**, 419–426.

Butucea, C. (2004) Quadratic functional estimation and testing from noisy data. *Preprint*. Université Paris 6, Paris.

Butucea, C. and Tribouley, K. (2003) Nonparametric homogeneity tests. *Preprint PMA-871*. Université Paris 6, Paris.

Candès, E. J. and Donoho, D. L. (2004) New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Communs Pure Appl. Math.*, **57**, 219–266.

Cappé, O., Guillin, A., Marin, J. M. and Robert, C. P. (2004) Population Monte Carlo. *J. Comput. Graph. Statist.*, to be published.

Casey, S. D. and Walnut, D. F. (1994) Systems of convolution equations, deconvolution, Shannon sampling, and the wavelet and Gabor transforms. *SIAM Rev.*, **36**, 537–577.

Cavalier, L. and Hengartner, N. W. (2003) Adaptive estimation for inverse problems with noisy operators. *Manuscript*.

Cavalier, L. and Tsybakov, A. (2001) Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Meth. Statist.*, **10**, 247–282.

Cavalier, L. and Tsybakov, A. (2002) Sharp adaptation for inverse problems with random noise. *Probab. Theory Reltd Flds*, **123**, 323–354.

Choi, H. and Baraniuk, R. G. (2001) Multiscale image segmentation, using wavelet-domain hidden markov models. *IEEE Trans. Image Process.*, **10**, 1309–1321.

Crouse, M. S., Nowak, R. D. and Baraniuk, R. G. (1998) Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.*, **46**, 886–902.

Csató, L., Opper, M. and Winther, O. (2002) TAP Gibbs free energy, belief propagation and sparsity. In *Advances in Neural Information Processing Systems*, vol. 14 (eds T. G. Dietterich, S. Becker and Z. Ghahramani). Cambridge: MIT Press.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Ann. Statist.*, **25**, 1–37.

Daubechies, I. (1995) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Davy, M. and Godsill, S. J. (2003) Bayesian harmonic models for musical signal analysis (with discussion). In *Bayesian Statistics 7* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc.* B, **56**, 363–375.

Donoho, D. (1995) Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harm. Anal.*, **2**, 101–126.

Donoho, D. and Raimondo, M. (2004) Translation invariant deconvolution in a periodic setting. *Int. J. Wavlts Multiresoln Inform. Process.*, to be published.

Dörfler, M. (2001) Time-frequency analysis for music signals: a mathematical approach. *J. New Mus. Res.*, **30**, 3–12.

Doucet, A., de Freitas, N. and Gordon, N. (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer.
Draper, D. W. and Long, D. G. (2002) An assessment of seawinds on QuikSCAT wind retrieval. *J. Geophys. Res.*, **107**.
Efromovich, S. and Koltchinskii, V. (2001) On inverse problems with unknown operators. *IEEE Trans. Inform. Theory*, **47**, 2876–2893.
Geman, D. (1988) Random fields and inverse problems in imaging. *Lect. Notes Math.*, **1427**.
Gribonval, R. and Bacry, E. (2003) Harmonic decomposition of audio signals with matching pursuit. *IEEE Trans. Signal Process.*, **51**, 101–111.
Groetsch, C. W. (1977) *Generalized Inverses of Linear Operators: Representation and Approximation*. New York: Dekker.
Haario, H., Laine, M., Mira, A. and Saksman, E. (2003) DRAM: efficient adaptive MCMC. *Report 374*. University of Helsinki, Helsinki.
Haario, H., Saksman, E. and Tamminen, J. (1999) Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.*, **14**, 375–395.
Haario, H., Saksman, E. and Tamminen, J. (2001) An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.
Hall, P., Paige, R. and Ruymgaart, F. H. (2003) Using wavelet methods to solve noisy Abel-type equations with discontinuous inputs. *J. Multiv. Anal.*, **86**, 72–96.
Hall, P., Ruymgaart, F., van Gaans, O. and van Rooij, A. (2001) Inverting noisy integral equations using wavelet expansions: a class of irregular convolutions. In *State of the Art in Probability and Statistics* (eds M. de Gunst, Ch. Klaassen and A. van der Vaart), pp. 533–546. Beachwood: Institute of Mathematics and Statistics.
Henderson, J. M., Hoffman, R. N., Leidner, S. M., Ardizzone, J. V., Atlas, R. and Brin, E. (2003) A comparison of a two-dimensional variational analysis method and a median filter for NSCAT ambiguity removal. *J. Geophys. Res.*, **108**, 3176.
Hoffman, R. N. (1984) SASS wind ambiguity removal by direct minimization, II: Use of smoothness and dynamical constraints. *Mnthly Weath. Rev.*, **112**, 1829–1852.
Hoffman, R. N., Leidner, S. M., Henderson, J. M., Atlas, R., Ardizzone, J. V. and Bloom, S. C. (2003) A two-dimensional variational analysis method for NSCAT ambiguity removal: methodology, sensitivity, and tuning. *J. Atmos. Ocean. Technol.*, **20**, 585–605.
Hopegood, R. J. and Rayner, J. W. (1999) Bayesian single channel blind deconvolution using parametric signal and channel models. *IEEE Wrkshp Applications of Signal Processing, New York, Oct. 17th–20th*.
Householder, A. S. (1964) *The Theory of Matrices in Numerical Analysis*. New York: Dover Publications.
Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C. and Liu, H. H. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond.* A, **454**, 903–995.
Johnstone, I. M. and Paul, D. (2004) Rate adaptive estimation in linear inverse problems through penalized least squares procedure. *Technical Report*. Stanford University, Stanford.
Johnstone, I. M. and Raimondo, M. (2004) Periodic boxcar deconvolution and diophantine approximation. *Ann. Statist.*, **32**, no. 5, in the press.
Jones, R. H. and Vecchia, A. V. (1993) Fitting continuous ARMA models to unequally spaced spatial data. *J. Am. Statist. Ass.*, **88**, 947–954.
Kerkyacharian, G. and Picard, D. (2000) Thresholding algorithms and well-concentrated bases. *Test*, **9**, 283–344.
Kerkyacharian, G., Picard, D. and Raimondo, M. (2004) Adaptive boxcar deconvolution on full Lebesgue measure sets. *Manuscript*.
Khabie-Zeitoune, E. (1982) Prediction in continuous time. In *Time Series Analysis: Theory and Practice 2* (ed. O. D. Anderson), pp. 7–24. Amsterdam: North-Holland.
Khinchin, A. Y. (1997) *Continued Fractions*. New York: Dover Publications.
Leidner, S. M., Isaksen, L. and Hoffman, R. N. (2003) Impact of NSCAT winds on tropical cyclones in the ECMWF 4D-Var assimilation system. *Mnthly Weath. Rev.*, **131**, 3–26.
Lorenc, A. C. (1986) Analysis methods for NWP. *Q. J. R. Meteorol. Soc.*, **112**, 1177–1194.
Malfait, M. and Roose, D. (1997) Wavelet-based image denoising using a Markov random field a priori model. *IEEE Trans. Image Process.*, **6**, 549–565.
Marron, J. S., Adka, S., Johnstone, I. M., Neumann, M. H. and Patil, P. (1998) Exact risk analysis of wavelet regression. *J. Comput. Graph. Statist.*, **7**, 278–309.
Nason, G. P., von Sachs, R. and Kroisandt, G. (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Statist. Soc.* B, **62**, 271–292.
Neelamani, R., Choi, H. and Baraniuk, R. (2004) Fourier-wavelet regularized deconvolution for ill-conditioned systems. *IEEE Trans. Signal Process.*, **52**, 418–433.
Ng, W.-J. (2000) Noise reduction for audio signals using the Gabor expansion. *MPhil Thesis*. University of Cambridge, Cambridge.
Pensky, M. and Zayed, A. I. (2002) Density deconvolution of different conditional distributions. *Ann. Inst. Statist. Math.*, **54**, 701–712.
Portilla, J., Strela, V., Wainwright, M. and Simoncelli, E. (2003) Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process.*, **12**, 1338–1351.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.
Robinson, E. A. (1967) *Multichannel Time Series Analysis with Digital Computer Programs*. San Francisco: Holden-Day.
Royle, J., Berliner, L., Wikle, C. and Milliff, R. (1998) A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador sea. In *Case Studies in Bayesian Statistics IV*, pp. 367–382. New York: Springer.
Sardy, S. (2000) Minimax thresholds for denoising complex signals with waveshrink. *IEEE Trans. Signal Process.*, **48**, 1023–1028.
Schultz, H. (1990) A circular median filter approach for resolving directional ambiguities in wind fields retrieved from spaceborne scatterometer data. *J. Geophys. Res.*, **95**, 5291–5304; erratum, 9783.
Shaffer, S. J., Dunbar, R. S., Hsiao, S. V. and Long, D. G. (1991) A median-filter-based ambiguity removal algorithm for NSCAT. *IEEE Trans. Geosci. Remote Sens.*, **29**, 167–174.
Stoffelen, A. and Anderson, D. (1997) Ambiguity removal and assimilation of scatterometer data. *Q. J. R. Meteorol. Soc.*, **123**, 491–518.
Tamminen, J. (2004) Validation of nonlinear inverse algorithms with Markov chain Monte Carlo method. To be published.
Thépaut, J.-N., Hoffman, R. N. and Courtier, P. (1993) Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mnthly Weath. Rev.*, **121**, 3393–3414.
de Vries, J. C. W. and Stoffelen, A. C. M. (2000) 2D variational ambiguity removal. *Technical Report 226*. Royal Netherlands Meteorological Institute, De Bilt.
Wahba, G. and Wang, Y. (1990) When is the optimal regularization parameter insensitive to the choice of the loss function? *Communs Statist. Theory Meth.*, **19**, 1685–1700.
Wainwright, M., Simoncelli, E. and Willsky, A. (2001) Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Appl. Comput. Harm. Anal.*, **11**, 89–123.
West, R. M., Aykroyd, R. G., Meng, S. and Williams, R. A. (2004) MCMC techniques and spatial temporal modelling for medical EIT. *Physiol. Measmnts*, **25**, 181–194.
West, R. M., Meng, S., Aykroyd, R. G. and Williams, R. A. (2003) Spatial-temporal modelling for electrical impedance imaging of a mixing process. In *Proc. 3rd Wrld Congr. Industrial Process Tomography, Banff*, pp. 226–232.
Wolfe, P. J. and Godsill, S. J. (2003) Bayesian estimation of time-frequency coefficients for audio signal enhancement. In *Advances in Neural Information Processing Systems*, vol. 15 (eds S. Becker, S. Thrun and K. Obermayer). Cambridge: MIT Press.
Zemanian, A. H. (1987) *Generalized Integral Transformations*. New York: Dover Publications.