J. R. Statist. Soc. B (2004)
66, *Part* 3, *pp.* 575–589

Bayesian variable selection and regularization for time–frequency surface estimation

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

University of Cambridge, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on 'Statistical approaches to inverse problems' on Wednesday, December 10th, 2003*, Professor J. T. Kent *in the Chair*]

Summary. We describe novel Bayesian models for time–frequency inverse modelling of non-stationary signals. These models are based on the idea of a *Gabor regression*, in which a time series is represented as a superposition of translated, modulated versions of a window function exhibiting good time–frequency concentration. As a necessary consequence, the resultant set of potential predictors is in general overcomplete—constituting a frame rather than a basis—and hence the resultant models require careful regularization through appropriate choices of variable selection schemes and prior distributions. We introduce prior specifications that are tailored to representative time series, and we develop effective Markov chain Monte Carlo methods for inference. To highlight the potential applications of such methods, we provide examples using two of the most distinctive time–frequency surfaces—speech and music signals—as well as standard test functions from the wavelet regression literature.

Keywords: Bayesian inference; Gabor frames; Model selection; Regression; Regularization

1. Introduction

Here we describe a novel approach for the modelling of time–frequency surfaces, which we term a *Gabor regression* (Ng, 2000; Wolfe, 2003). This approach consists of a Bayesian regularization scheme in which prior distributions over the time–frequency coefficients are constructed to favour both smoothness of the estimated function and sparseness of the coefficient representation. The methodology that we propose may be related to that of wavelet regression and variable selection (see, for example, Chipman *et al.* (1997), Clyde *et al.* (1998), Vidakovic (1998) and Abramovich *et al*. (1998)), but a crucial difference lies in the dependences that are introduced between the coefficients in the model, both through the inherent non-orthogonality and overcompleteness of the representation and through the prior structures that are imposed on these coefficients. Such structures borrow Markov random-field ideas from image processing (see, for example, Besag (1974), Geman and Geman (1984) and Ripley (1988)) to model the persistences through time and frequency which are expected to be present in many naturally generated signals of interest.

Although it is common practice to apply a Fourier series representation for nonparametric analysis of time series (see, for example, Priestley (1981), Wahba (1983), Gallant and Monahan (1985) and Lenk (1999)), such a representation is no longer appropriate in situations where spectral content may vary with time. One possible replacement for the Fourier representation in such cases is a time–scale representation using wavelets (Müller and Vidakovic, 1999; Vidakovic,

Address for correspondence: Patrick J. Wolfe, Signal Processing Group, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK.
E-mail: p.wolfe@ieee.org

 1369–7412/04/66575

1999); for an interpretation of such an approach in terms of non-stationary processes, see Nason *et al*. (2000). Indeed, researchers in many fields have been drawn in particular to wavelets as one class of compactly supported functions forming an orthonormal basis (in the critically decimated case). However, Gabor analysis represents a similarly suitable and technically convenient choice, as we demonstrate. In fact, just as the Fourier transform diagonalizes stationary random processes, Gabor representations can be viewed as approximately diagonalizing a class of slowly time-varying systems—i.e. locally or quasi-stationary random processes (Priestley, 1981)—known as underspread operators (Kozek, 1998).

Gabor analysis begins with the idea of an energy density distributed over a *lattice* of points in the time–frequency plane (the spectrogram or squared modulus of the Gabor transform being a well-known example), corresponding to a representation of the signal in terms of well-localized *time–frequency atoms*. This concept is illustrated in Fig. 1, where the lattice is given by $a\mathbb{Z} \times b\mathbb{Z}$ for *time–frequency shift parameters* $a, b > 0$, and the atom whose time–frequency coverage is centred at the origin is denoted g, such that an atom $g_{m,n}$ is a time-shifted (translated by na) and frequency-shifted (modulated by mb) version thereof. *Gabor frames* formalize the notion of the short time Fourier transform and the concept of valid *tilings* of the time–frequency plane (see, for example, Gröchenig (2001)). Indeed, a frame in a (separable) Hilbert space $\mathcal{H}$ is a sequence $\{\Psi_k : k \in \mathcal{K}\}$ having the additional property that there are constants $A, B > 0$ (frame bounds) such that

$$A\|f\|^2 \leqslant \sum_{k\in\mathcal{K}} |\langle f, \Psi_k\rangle|^2 \leqslant B\|f\|^2 \qquad \forall f \in \mathcal{H}. \tag{1}$$

In particular, a Gabor frame $(g_{m,n})$ is a special case of inequality (1) in which the set $\{\Psi_k\}$ comprises translated and modulated versions (indexed by n and m respectively) of a single, basic window function g.

The notion of a frame incorporates bases as well as certain redundant representations; for example, an orthonormal basis is a *tight* frame ($A = B$) with $A = B = 1$, and the union of two orthonormal bases yields a tight frame with frame bounds $A = B = 2$. A key result in time–frequency analysis—the *Balian–Low theorem*—implies that 'redundancy is a necessary

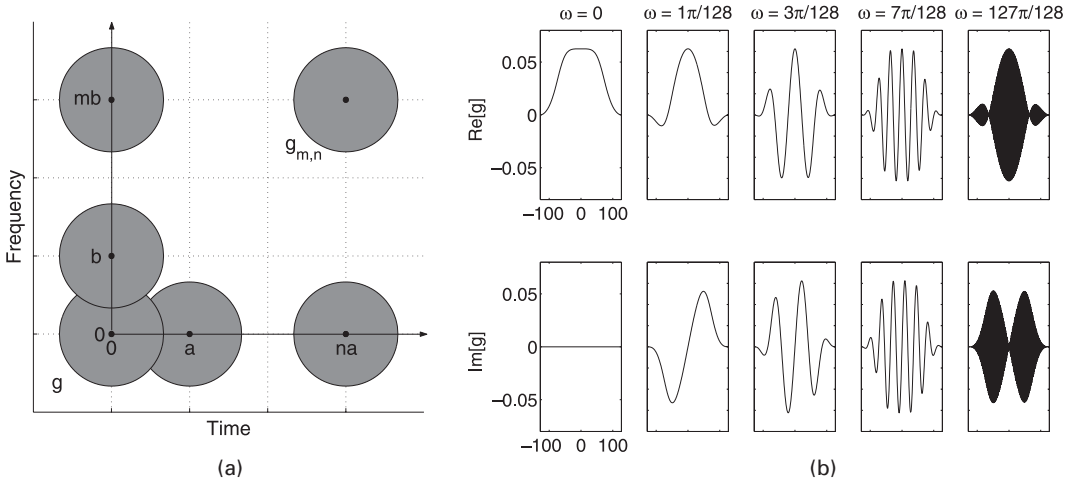

Fig. 1. (a) Tiling of the time–frequency plane via copies of an atom g, translated in time and frequency (after Feichtinger and Strohmer (1998)) and (b) the real and imaginary parts of some typical Gabor atoms (see equation (3)), with modulating angular frequency $\omega = 2\pi m/M$

consequence of good time–frequency localization' (Gröchenig, 2001), and hence that there is no well-concentrated 'Gabor basis' in the critically sampled case. (An exception to the Balian–Low theorem for real-valued functions is provided by the idea of *Wilson bases* (Wilson, 1987), including certain local trigonometric bases (Coifman and Meyer, 1991) and lapped transforms (Malvar, 1990), examples of which we consider briefly in Section 5.1.)

Allowing for overlap between adjacent atoms (via a denser sampling of the time–frequency plane) permits the use of window functions with better time–frequency concentration properties, albeit at the price of overcompleteness. Hence, the inference of a time–frequency surface for data measured in additive noise is an inherently ill-posed inverse problem, requiring a high degree of regularization to avoid overfitting and modelling of noise. A time–frequency representation for the data will often have a natural physical interpretation in terms of a generative model; for example, in the sound signals that we consider later, such a representation can be directly related to the physical sound production mechanism. As we show, the overcompleteness of a Gabor representation can actually be an advantage in the context of time–frequency surface estimation. Moreover, although such an analysis task arises in many areas of science and engineering, the associated applications of regression and compression are equally important—and in fact these issues are fundamentally related to the mathematics of harmonic analysis which underlie Gabor systems (Donoho *et al.*, 1998).

The remainder of this paper is organized as follows. In Section 2 we specify the regression model and introduce the concept of Gabor analysis which underlies it. In Section 3 we discuss our formulation of prior dependence structures in the time–frequency plane, and in Section 4 we describe the algorithmic implementation of our model by means of a Markov chain Monte Carlo (MCMC) sampling scheme. In Section 5 we present and interpret simulation results for three classes of the Gabor regression scheme: overcomplete representations using diffuse priors to induce sparsity, model-averaged representations using unstructured priors to allow for the transitions which naturally occur in non-stationary data and model-averaged representations using conditionally Markov priors to favour persistence of meaningful signal traits and trends in the time–frequency plane. We conclude in Section 6 with a discussion of related issues and future extensions to the framework proposed. Sampling scheme derivations are provided in Appendix A.

2. The Gabor regression model

2.1. Gabor analysis over finite cyclic groups

The *Gabor expansion* of an L-periodic sequence $f \in l^2(\mathbb{Z})$, where $l^2(\mathbb{Z})$ denotes the space of square summable sequences, is given by

$$f = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} c_{m,n} \tilde{g}_{m,n}, \tag{2}$$

where the *Gabor synthesis atoms* $\tilde{g}_{m,n}$ are derived from a (typically smooth) window function $\tilde{g}(t)$ as

$$\tilde{g}_{m,n}(t) \triangleq \tilde{g}\Big(t - \frac{n}{N}L\Big) \exp\Big(2\pi j \frac{m}{M} t\Big), \qquad t \in \{0, 1, \ldots, L-1\}, \tag{3}$$

and $c_{m,n} =^{\triangle} \langle f, g_{m,n} \rangle$ are the *Gabor expansion coefficients*—i.e. the inner products of f and translated, modulated versions of the dual (analysis) window function $g(t)$—corresponding to the Gabor transform of f, which we note may be efficiently computed in practice via a sequence of fast Fourier transforms applied after windowing f with translates of $g(t)$.

Here we are in fact implicitly considering periodic sequences as members of a finite cyclic group—i.e. in terms of the circular extension of finite length, complex-valued sequences $f \in \mathbb{C}^L$ on the ring $\mathbb{Z}_L =^{\triangle} \mathbb{Z} \bmod(L)$. Using vector matrix notation, we may denote the Gabor expansion as $f = \tilde{G}c$, where the sequence f is expressed in the form of a column vector: $f = (f_0\ f_1 \dots f_{L-1})^{\mathrm{T}}$, $\tilde{G}$ denotes the $L \times MN$ *Gabor synthesis matrix* having $\tilde{g}_{m,n}$ as its $(m+nM)$th column and the Gabor expansion coefficients $\{c_{m,n}\}$ are written in the form of a 'stacked' column vector c of length MN. The explicit definitions that are employed in our derivation follow those of Strohmer (1998) and are detailed in Appendix A.

2.2. Bayesian model

Let $x \in \mathbb{C}^L$ in general denote a complex-valued time series, the observation of which has been corrupted by additive noise. From the completeness property of frames, it follows that x may be represented as a linear combination of frame elements. Hence we consider regression using a collection of Gabor synthesis atoms $\{\tilde{g}_{m,n}\}$, under the assumption—without loss of generality—that the corresponding vector of sampled observations y has been extended to length L in a proper way at its boundary before being periodically extended on the ring $\mathbb{Z}_L$. Moreover, in the applications that we consider here, the data are constrained to be real valued, and hence by Hermitian symmetry properties we have that $c_{m,n} = c^*_{M-m,n}$ for all $m \in \{1, 2, \dots, M/2\}$. The models and sampling algorithms that we describe explicitly incorporate this constraint by including only modulations in the set $\{0, 1, \dots, M/2\}$ (where M is assumed to be even), and thus from equation (2) we have the following linear time series model in which all variables are *real*:

$$y_t = \sum_{m=0}^{M/2} \sum_{n=0}^{N-1} c_{m,n}\, \tilde{g}_{m,n}(t) + d_t, \qquad t \in \{0, 1, \dots, L-1\}.$$

The quantity d_t is a disturbance or noise term, modelled here as an independent, identically distributed Gaussian random variable with distribution $\mathcal{N}(0, \sigma^2)$—although this assumption is readily relaxed within the numerical Bayesian setting that we describe below. The term $c_{m,n}$ is interpreted as a two-dimensional row vector containing the real and imaginary parts of the complex coefficient at time–frequency lattice point (m, n), and $\tilde{g}_{m,n}(t)$ as a column vector containing the real and imaginary parts of the corresponding Gabor synthesis atom at time index t. For clarity of exposition, we shall use $k \in \{0, 1, \dots, K-1\}$ to index complex coefficients, bearing in mind that $K = (M/2+1)N$ and $c_k = c_{m+nM}$ corresponds to $c_{m,n}$.

3. Frames, sparsity and prior dependence structures

Frame theory guarantees minimal norm expansion coefficients in an l^2-sense via the Gabor transform $\{\langle f, g_{m,n}\rangle\}$; however, such a representation is not guaranteed to be maximally sparse and hence may not be desirable for all applications. Indeed, many time–frequency representations (including the short time Fourier transform) may be viewed as a convolution of the Wigner distribution of a function with a smoothing kernel (see, for example, Gröchenig (2001)). Our approach to time–frequency surface estimation is intended to overcome some of the limitations of such transform methods, both by providing an 'unsmoothing' effect in comparison with the Gabor transform as well as the potential for a sparse coefficient representation. We note that some researchers have pursued the variational approach to sparsity directly: for example, basis pursuit (Chen *et al.*, 2001) provides an algorithm for minimal l^1-norm decomposition, and 'greedy' approaches to function approximation such as the matching pursuit algorithm of Mallat and Zhang (1993) have also been proposed. (For recent results on the uniqueness of sparse

decompositions, as well as conditions for the equivalence of minimal l^1- and l^0-representations, see Donoho and Huo (1999).)

Here, however, we consider regularization and sparsity within the Bayesian framework. In this vein, periodogram estimators have been interpreted as minimizers of regularized least squares criteria (Giovannelli and Idier, 2001). Bayesian shrinkage (soft thresholding) and variable selection (hard thresholding) have been used to induce sparsity in images (Olshausen and Millman, 2000) and other signals (Crouse *et al.*, 1998), as well as more generally in the statistical community (see, for example, Mitchell and Beauchamp (1988), George and McCulloch (1993) and Brown *et al.* (1998)); some recently proposed Bayesian variable selection schemes specifically address the overcomplete case (Brown *et al.*, 2002; West, 2003). Even without the explicit concentration of probability near zero *a priori*, overcompleteness coupled with heavy-tailed prior distributions may be used to induce sparsity. Such approaches have recently attracted much attention in the machine learning community (see, for example, Poggio and Girosi (1998), Lewicki and Sejnowski (2000), Figueiredo (2001) and Tipping (2001)). We model sparseness in the time–frequency coefficients explicitly via variable selection, achieved through the introduction of $\gamma_k \in \{0, 1\}$ such that

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

Although such a formulation is by now familiar through the many applications of MCMC methods to variable selection in the literature (see, for example, George and McCulloch (1997) and Godsill (2001)), we extend it by modelling dependence across the time–frequency lattice by using the set of indicator variables $\{\gamma_k\}$.

Conditionally on $\gamma_k = 1$, the prior distribution of the coefficients is a heavy-tailed Student t scale mixture of normals with inverted gamma mixing distribution $p(\sigma^2_{c_k}|\gamma_k = 1, \kappa, \nu_k) = \mathcal{IG}(\sigma^2_{c_k}|\kappa, \nu_k)$; this specification is designed to capture the wide range of coefficient values expected in a non-stationary process, with the shape parameter κ determining how heavy tailed this mixture will be. In comparison with a prior assumption of independent, normally distributed coefficients sharing a *common* variance, the regularization that is induced by such a model prevents oversmoothing of the time–frequency coefficients and can lead to a sparser representation in terms of coefficient energy concentration on the time–frequency lattice. The form of the mixing distribution is chosen for ease of implementation; we note that other mixing distributions can be readily incorporated into the sampling scheme when more domain-specific information is available (see, for example, Godsill and Rayner (1998) and Godsill (1999)). The scale parameter ν_k is assigned its own prior; specifically, we assume that $\nu_k = f(k)\nu$, with $\nu \sim \mathcal{G}(\alpha_\nu, \beta_\nu)$ being gamma distributed. Here $f(k)$ is a fixed weighting function that can be used to express a desired degree of smoothness in the reconstructed signal, quantifiable in terms of the decay of its Fourier transform via Bessel potential spaces (Gröchenig, 2001). In implementations to date, we have used a weighting function $f(k) = 1/m(k)^a$, where $m(k)$ is the frequency modulation number for coefficient k and typically $a = 1$. For an example of similar considerations in a Fourier series setting, see Carter and Kohn (1997) and Lenk (1999).

An important component of the model is the prior distribution for the indicator process $\gamma = \{\gamma_k\}$. Here dependence between coefficients in time and frequency can readily be incorporated. We may probably expect *a priori* some continuity through time and frequency for the coefficients; in particular, for natural signals we expect certain regions of the time–frequency lattice to be heavily populated, whereas others (where there is little signal activity) will be rather sparse. To achieve this 'persistence' of energy in the time–frequency plane, we specify the distribution of γ_k conditionally through $p(\gamma_k|\gamma_{-k}, \phi)$, where γ_{-k} denotes all the indicator variables

except γ_k, and ϕ are the parameters of the distribution. (For a related discussion concerning wavelet models for images, see Crouse *et al.* (1998).)

For this we consider several possibilities for $p(\gamma_k)$, all of which are assumed conditionally Markov such that $p(\gamma_k|\gamma_{-k}, \phi) = p(\gamma_k|\gamma_{\mathcal{N}(k)}, \phi)$, where $\mathcal{N}(k)$ is a local neighbourhood of the time–frequency lattice point k corresponding to atom c_k. We have experimented with various arrangements: $\mathcal{N}(k) = \emptyset$ (i.e. a Bernoulli prior on γ_k), leading to very sparse but potentially unstructured coefficient representations which may be most appropriate for compression; Markov chain priors favouring persistence across time (or equally frequency) on the lattice, which are potentially useful for signals exhibiting slowly time-varying oscillations; Markov random fields based on first-order neighbourhood structures, these being well suited to signals whose time–frequency activity occurs in 'patches', in which case we wish to avoid including spurious isolated components in the model. These latter two models can lead to a more structured and interpretable time–frequency surface, although at the same time they may possibly induce a less sparse representation. The precise form of the prior that is chosen will imply a valid joint distribution $p(\gamma|\phi)$, although the MCMC algorithms that are presented below require only the conditional prior distribution $p(\gamma_k|\gamma_{-k}, \phi)$. We note that such models are very flexible and may readily be constructed in a manner which exploits any available prior information that is relevant to a particular application.

The remaining prior hyperparameters $\alpha, \beta, \alpha_\nu$ and β_ν, along with the priors on the indicator parameters ϕ, are generally set to give very diffuse prior distributions, unless specific knowledge is available. Defining $y = \{y_0, y_1, \ldots y_{L-1}\}$, $c = \{c_k\}$, $\sigma_c = \{\sigma_{c_k}\}$ and $\gamma = \{\gamma_k\}$, the joint distribution of all parameters and data may hence be specified as

$$p(c, \sigma_c, \sigma, \nu, \gamma, \phi, y) = p(y|c, \sigma)\; p(c|\sigma_c, \gamma)\; p(\sigma_c|\nu)\; p(\gamma|\phi)\; p(\phi)\; p(\nu); \tag{4}$$

$$p(y|c, \sigma) = \prod_{t=0}^{L-1} \mathcal{N}\left\{ \sum_{m=0}^{M/2} \sum_{n=0}^{N-1} c_{m,n}\, \tilde{g}_{m,n}(t), \sigma^2 \right\},$$

$$p(c|\sigma_c, \gamma) = \prod_{k=0}^{K-1} p(c_k|\sigma_{c_k}, \gamma_k),$$

$$p(\sigma_c|\nu) = \prod_{k=0}^{K-1} \mathcal{IG}(\sigma_{c_k}^2|\kappa, \nu_k).$$

4. Markov chain Monte Carlo inference

The posterior distribution of the parameters is explored by using an MCMC procedure (Gilks *et al.*, 1996; Robert and Casella, 1999), the aim of which is to draw a large number of random realizations from the joint posterior density $p(c, \sigma_c, \sigma, \nu, \gamma, \phi|y)$ as defined in equation (4), from which any desired posterior inference can be computed as a Monte Carlo integral. In particular, in this application we shall be concerned with the time–frequency *surfaces* that are represented by $p(c|y)$ and the corresponding indicator distribution $p(\gamma|y)$.

The MCMC procedures involve elements of model mixing through variable selection (sampling γ) (George and McCulloch, 1993, 1997; Godsill, 2001) in conjunction with more standard Gibbs sampling moves for the other parameters. For such a complex model there are clearly many possible blocking strategies; indeed, the reduced conditionals that are available for certain parameter subsets facilitate Rao–Blackwellized sampling (Robert and Casella, 1999). The object of this paper, however, is not to compare the various sampling strategies and their individual merits—but rather to present a new modelling methodology for time-varying systems. Hence we

adopt a simple and efficient sampling scheme that has been observed to converge very quickly in the cases (using both synthetic and natural data) that we have studied.

In the implementation that we consider here, the shape parameter κ and the frequency decay parameter a are fixed, although we note that these could be updated in a relatively straightforward manner. Hence each sweep of the sampling scheme proceeds as follows.

(a) *Update* σ_c: a standard Gibbs step is used to update the full conditional for σ_c—

$$p(\sigma_{c_k}^2 | c, \sigma_{-c}, \sigma, \nu, \gamma, \phi, y) = \mathcal{IG}\left(\gamma_k + \kappa, \gamma_k \frac{\|c_k\|^2}{2} + \nu_k\right).$$

The implication here is that, with $\gamma_k = 0$, $\sigma_{c_k}^2$ is drawn from the prior mixing distribution, since the coefficients and the data are conditionally independent of this parameter. In fact other choices are possible for this distribution, which is arbitrary provided that it is proper and leads to an ergodic Markov chain overall, an issue which is explored in detail in Godsill (2001). Although other choices, which may depend on other parameters in the current MCMC state or indeed the data themselves, could lead to improved convergence of the sampler, we have found that this choice is perfectly adequate.

(b) *Update* c *and* γ: the full conditional distribution for the Gabor coefficient vector $p(c|\sigma_c, \sigma, \nu, \gamma, \phi, y) = p(c|\sigma_c, \sigma, \gamma, y)$ is multivariate Gaussian (see Appendix A.2), and hence blocking strategies and Rao–Blackwellization schemes can readily be incorporated. In particular, we have implemented full block draws from $p(c|\sigma_c, \sigma, \gamma, y)$ as well as successive draws from conditional subblocks of c. In the results that are presented here, the basic sampling step is a conditional Gibbs draw from $p(c_k, \gamma_k | c_{-k}, \gamma_{-k}, \sigma_{-c}, \sigma, \nu, \phi, y)$, which is found to ensure rapid exploration of the coefficient and indicator space. Moreover, this and other schemes that are described in Appendix A.2 have the advantage that the full design matrix need never be constructed.

The block conditional is summarized as follows (see, for example, Barnett *et al*. (1996) and Godsill and Rayner (1998) for derivations in a related indicator modelling framework):

$$p(c_k, \gamma_k = 1 | c_{-k}, \gamma_{-k}, \sigma_{c-k}, \sigma, \nu, \phi, y) = \frac{\tau_k}{1+\tau_k} \mathcal{N}(c_k | \mu_k, \sigma^2 \Sigma_k),$$

$$p(\gamma_k = 0 | c_{-k}, \gamma_{-k}, \sigma_{c-k}, \sigma, \nu, \phi, y) = \frac{1}{1+\tau_k},$$

where

$$\Sigma_k \triangleq \left(\tilde{G}_k^{\mathrm{T}} \tilde{G}_k + \frac{\sigma^2}{\sigma_{c_k}^2} I_2\right)^{-1},$$

$$\mu_k \triangleq \Sigma_k \tilde{G}_k^{\mathrm{T}} (y - \tilde{G}_{-k} c_{-k})$$

and

$$\tau_k \triangleq \frac{p(\gamma_k = 1 | \gamma_{-k})}{p(\gamma_k = 0 | \gamma_{-k})} \frac{\sigma^2}{\sigma_{c_k}^2} |\Sigma_k|^{1/2} \exp\left(\frac{\mu_k^{\mathrm{T}} \Sigma_k^{-1} \mu_k}{2\sigma^2}\right).$$

We emphasize that this result is presented for the case in which the Gabor representation is constrained to be real valued—i.e. $\tilde{G}_k \in \mathbb{R}^{L \times 2}$, $\Sigma_k \in \mathbb{R}^{2 \times 2}$ and $\mu_k, c_k \in \mathbb{R}^2$.

(c) *Update* ϕ: the parameter subset ϕ relates to the joint indicator distribution $p(\gamma|\phi)$, for which purpose we present three possibilities—

(i) *an independent Bernoulli prior* $p(\gamma_k|\phi)$, where ϕ is assigned a beta prior; in this case ϕ is marginalized directly as specified in Appendix A.3, leading to a marginal joint distribution $p(\gamma_k)$;

(ii) *a Markov chain prior in time* $p(\gamma_{m,n}|\gamma_{m,n-1},\phi)=\phi^m_{\gamma_{m,n-1},\gamma_{m,n}}$, where ϕ^m is the transition matrix of the Markov chain at the mth modulation index; the terms $\phi^m_{0,0}$ and $\phi^m_{1,1}$ are assigned independent beta priors, and the initial distribution of each chain is taken to be its stationary distribution; in this case the transition probabilities may be sampled by using a Metropolis–Hastings step, for which we use as a proposal distribution the respective full conditional for the parameter in question in the case of a fixed, uniformly distributed initial state (see Appendix A.4);

(iii) *a first-order Markov random field*, in which case Ising-type priors with fixed parameters are adopted (see, for example, Geman and Geman (1984)).

(d) *Update* ν: the common scale parameter ν can be drawn from its full conditional distribution $p(\nu|\sigma_c,\kappa)$; however, the dependence of ν on *all* components of σ_c can lead to very slow convergence if a large proportion of elements in γ are 0. Hence we propose a draw from the reduced conditional

$$p(\nu|\{\sigma_{c_k}:\gamma_k=1\},\kappa)=\mathcal{G}\left\{\kappa|\gamma|+\alpha_\nu,\ \sum_{k:\gamma_k=1}\frac{f(k)}{\sigma^2_{c_k}}+\beta_\nu\right\},$$

where $|\gamma|$ is the cardinality of the set $\{\gamma_k:\gamma_k=1\}$, following which the remaining terms $\{\sigma_{c_k}:\gamma_k=0\}$ are reimputed from their conditional given the new value of ν. This may be considered to be a block draw from $p(\nu,\{\sigma_{c_k}:\gamma_k=0\}|\{\sigma_{c_k}:\gamma_k=1\},\kappa)$.

(e) *Update* σ: the noise variance parameter σ^2 is drawn in a Gibbs step from its full conditional, which is of the form

$$\mathcal{IG}\left(\frac{L+\alpha}{2},\frac{\|y-\tilde{G}c\|^2+\beta}{2}\right).$$

5. Examples of time–frequency surface estimation

Here we demonstrate the application of the Gabor regression scheme that was detailed in Section 2 to a selection of time series. In the experiments that we describe, a window corresponding to that shown in Fig. 1(b) was employed as the prototype Gabor synthesis function, and a regular time–frequency lattice was constructed to yield a redundancy of 2 (corresponding to the common practice in applications of a 50% 'window overlap' in time).

5.1. Coefficient shrinkage in the overcomplete case

We first consider the special case of our model obtained when $\gamma_k=1$ for all k, so that the resultant time–frequency surfaces are overcomplete. In such a case, we have consistently observed that, as a result of the heavy-tailed coefficient prior $p(c)$, the estimated time–frequency surfaces have much smaller l^1-norms than the corresponding Gabor transform representations of minimal l^2-norm, and indeed appear 'sharpened' in comparison.

It is reasonable to ask what is to be gained by such overcompleteness; in particular Daubechies *et al.* (1991) showed how a Wilson basis may be obtained by eliminating members of such an overcomplete Gabor frame. We hence consider, by way of comparison, a lapped transform basis obtained by removing members of the collection $\{\tilde{g}_{m,0}\}$ to yield a locally orthogonal set which was replicated and then translated in such a manner as to emulate the 50% window overlap that

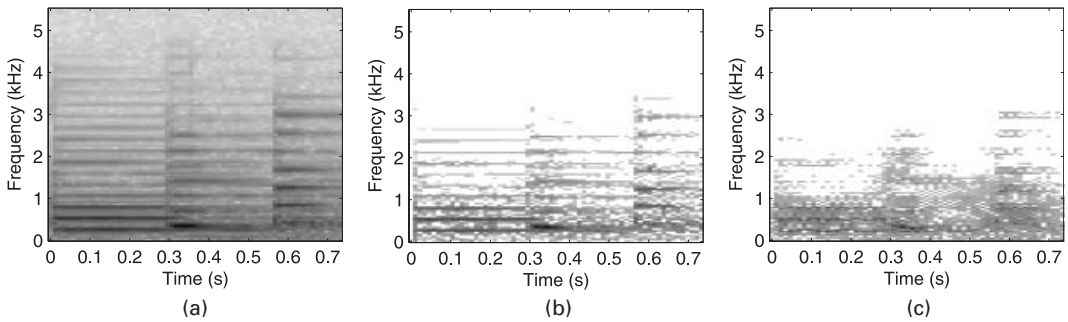

Fig. 2. (a) Gabor transform of a piano signal and (b) the 25% largest estimated coefficients as modelled by an overcomplete frame and (c) a basis derived from the same window function (the grey scale is shown in proportion to log-magnitude)

was employed in our usual redundancy 2 Gabor frame. Experiments indicate that such a scheme proves effective for regression (see Wolfe and Godsill (2003) for further quantitative results in this direction), in which case the time series reconstruction is the final consideration. However, we note that properties provided by an overcomplete representation (such as translation invariance) may be preferable in applications where the time–frequency coefficients themselves are of ultimate interest.

To demonstrate these differences in a modelling context, Fig. 2 shows a section of a piano signal spectrogram, along with the largest 25% of coefficients taken from the posterior mean estimates (averaged over 5000 iterations after 2000 iterations of 'burn-in') of time–frequency surfaces corresponding to the coefficient vector c resulting from the overcomplete Gabor regression and the basis regression using the same model. It may be seen that the overcomplete estimation scheme captures time–frequency ridges more effectively than does the estimated basis representation. The manual thresholding also illuminates the effect of the heavy-tailed prior in inducing a sparse representation with lower l^1-norm, which would hence be more appropriate for compression via thresholding.

5.2. Variable selection using structured and unstructured models

As detailed in Section 3, the more general formulation of the Gabor regression framework includes latent indicator variables which switch individual time–frequency atoms in and out of the model, thus allowing us to perform model selection for compression or Bayesian model averaging for regression. For this, we now consider variable selection in the context of the Gabor scheme, using both structured and unstructured models. Although we do not address compression explicity in this case, we note that results that are similar to those presented in Section 5.1 have been obtained, and in general the average number of selected regressors tends towards a low percentage of the total, indicating that such a scheme is likely to be an effective methodology for thresholding.

To test the ability of the Gabor regression model to estimate time–frequency surfaces in the presence of noise, a short speech utterance was artificially degraded with white Gaussian noise to yield a signal-to-noise ratio of 15 dB. Fig. 3 shows a comparison of the estimated time–frequency surfaces in terms of the posterior mean indicator estimate $\hat{\gamma}|y$, obtained by using each of the three prior dependence configurations that were described in Section 3 and step (c) of Section 4. Note especially the spectral ridges that are captured by the independent Bernoulli prior and the Markov chain prior in time (particularly at the beginning of the utterance) in contrast with the large patches of signal activity modelled by the local Markov random-field prior.

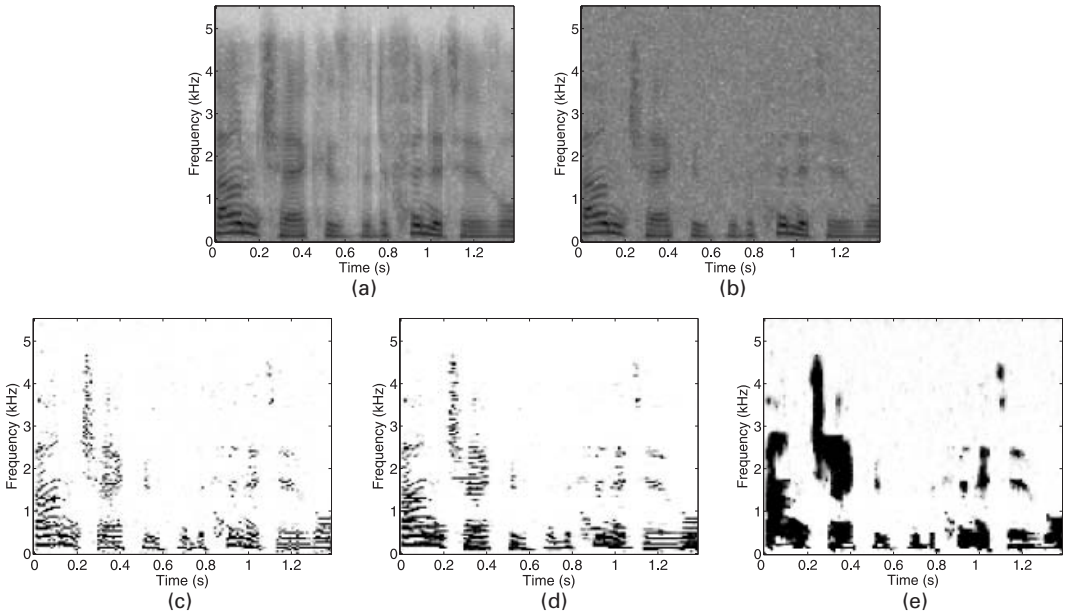

Fig. 3. (a) Gabor transforms of a speech signal and (b) a degraded version thereof, and the corresponding indicator estimates by using (c) a Bernoulli prior, (d) a Markov chain prior in time and (e) a first-order Markov random-field prior

We also present the results of a typical realization of our Gabor regression model on several standard test functions from the wavelet regression literature, each consisting of a 1024-point time series degraded with additive white Gaussian noise corresponding to the 'high noise' case as specified in Marron *et al.* (1998). Table 1 summarizes the results of these regression experiments, including the error norms of the degraded and reconstructed signals, as well as the measured noise variance in comparison with that estimated by the Gabor regression scheme. In these experiments, Bayesian model averaging was employed in conjunction with an Ising prior and a tight frame of redundancy 2, based on a 32-sample version of the Gabor window function that was used in the experiments described above.

Results were averaged over 1000 iterations following 1000 iterations of burn-in, and no attempt was made to optimize the values of fixed parameters. (As noted below, these and other experiments described herein can be reproduced with the aid of a MATLAB toolbox that has been developed by the authors.) It may be seen from Table 1 that the Gabor regression scheme has accurately estimated the noise variance in each case, as well as having reduced the error norm by over 50% in comparison with the noisy versions of these test functions.

6. Discussion

In this paper we have presented examples of nonparametric time–frequency surface estimation, for which Bayesian models of the Gabor coefficients have been shown to provide appropriate methodology. This framework can be naturally extended to include (overcomplete) multiresolution dictionaries of time–frequency atoms, hence allowing local adaptation to the characteristics and degree of non-stationarity of the data. Indeed, the multiresolution wavelet-like schemes that were explored in Ng (2000), Wolfe *et al.* (2001) and Wolfe (2003) represent one of many possibilities in this direction and constitute a continuing area of research.

Table 1. Typical Gabor regression results on several standard test functions (Marron *et al.*, 1998)

Test function	*Measured variance*	*Estimated variance*	*Error norm, original*	*Error norm, reconstruction*
Step	0.0107	0.0113	3.3079	1.4491
Wave	0.0107	0.0107	3.3079	1.5624
Blip	0.0107	0.0105	3.3079	1.5107
Blocks	0.0107	0.0109	3.3079	1.5979
Bumps	0.0107	0.0098	3.3079	1.5305
HeaviSine	0.0107	0.0106	3.3079	1.5560
Doppler	0.0107	0.0107	3.3079	1.5219
Angles	0.0107	0.0102	3.3079	1.4390
Parabolas	0.0107	0.0107	3.3079	1.4607
TSh Sine	0.0107	0.0105	3.3079	1.5031

Although we have limited our treatment here to the case of overcomplete dictionaries comprising time–frequency shifts (Gabor frames) rather than time–scale shifts (wavelet frames), we note that the underlying framework is sufficiently general to handle either case and indeed can be extended to frames constructed as unions of these families as well as other types of functions. We have also developed this methodology bearing in mind the large data sets that are typical in applications (over 10000 samples per second in speech processing, for example), and a further advantage of our approach is that memory requirements for the inference algorithms are minimal.

Other areas warranting investigation include the extension of our methodology to the sequential case by using particle filter methods (Doucet *et al.*, 2001), as well as to the case of non-Gaussian and non-stationary noise models, and missing or irregularly sampled data. A thorough exploration of more general Markov random-field dependence structures is another area of current work, and automatic selection between different classes of Markov random-field models is of much interest as a longer-term goal. As a final note, both the data described herein and the corresponding MATLAB code are available on line at `http://www-sigproc.eng.cam.ac.uk/~pjw47`. Code updates and further results of interest will also be posted as they become available.

Acknowledgements

Simon Godsill acknowledges partial support in research from the European Union Framework 5 project 'Models for unified multimedia information retrieval'. Patrick Wolfe and Simon Godsill also acknowledge the support of the UK Engineering and Physical Sciences Research Council 'Realising our potential award' project 'High-level modelling and inference for audio signals using Bayesian atomic decomposition'. The authors also thank the reviewers for their helpful suggestions in revising this work.

Appendix A. Sampler derivations for the Gabor regression model

A.1. Real-valued implementation of the sampling scheme

As our interest here is in finite time series $x \in \mathbb{R}^L$ and real-valued noise processes, when formulating the Gabor synthesis matrix $\tilde{G}$ we need only to consider (assuming that the frequency lattice constant M is even) modulations $m \in \{0, 1, \ldots, M/2\}$ and the corresponding real part of the reconstruction of x from

its complex Gabor coefficients c. In this case we may specify an implementation which exploits conjugate symmetry and uses only real numbers.

We may formulate such an implementation by taking advantage of the fact that both the Gabor coefficient vector $c \in \mathbb{C}^K$ and the Gabor synthesis matrix $\tilde{G} \in \mathbb{C}^{L \times K}$ may be rewritten in terms of their real and imaginary parts respectively. Hence, for the sake of implementation, we may use properties of complex multiplication to redefine $\tilde{G}$ and c as follows:

$$\tilde{G} \triangleq \begin{pmatrix} \mathrm{Re}\{\tilde{g}_{0,0}(0)\} & \mathrm{Im}\{\tilde{g}_{0,0}(0)\} & \cdots & \mathrm{Re}\{\tilde{g}_{M/2,N-1}(0)\} & \mathrm{Im}\{\tilde{g}_{M/2,N-1}(0)\} \\ \mathrm{Re}\{\tilde{g}_{0,0}(1)\} & \mathrm{Im}\{\tilde{g}_{0,0}(1)\} & \cdots & \mathrm{Re}\{\tilde{g}_{M/2,N-1}(1)\} & \mathrm{Im}\{\tilde{g}_{M/2,N-1}(1)\} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{Re}\{\tilde{g}_{0,0}(L-1)\} & \mathrm{Im}\{\tilde{g}_{0,0}(L-1)\} & \cdots & \mathrm{Re}\{\tilde{g}_{M/2,N-1}(L-1)\} & \mathrm{Im}\{\tilde{g}_{M/2,N-1}(L-1)\} \end{pmatrix}$$

and

$$c \triangleq (\mathrm{Re}(c_{0,0}) \quad -\mathrm{Im}(c_{0,0}) \quad \mathrm{Re}(c_{1,0}) \quad -\mathrm{Im}(c_{1,0}) \quad \ldots \quad \mathrm{Re}(c_{M/2,N-1}) \quad -\mathrm{Im}(c_{M/2,N-1}))^{\mathrm{T}},$$

where we note that in this case the redefined coefficient vector c is trivially related to the 'true' vector of complex Gabor coefficients.

A.2. Full conditional and blocking schemes for sampling c

Considering the redefinitions described in Appendix A.1 such that $\tilde{G} \in \mathbb{R}^{L \times 2K}$ and $c \in \mathbb{R}^{2K}$, it is straightforward to verify that the distribution of the included Gabor coefficients $c_{\gamma_k=1} =^{\triangle} \{c_k : \gamma_k = 1\}$, conditional on the remaining model parameters and the data, is given for any of the 2^K possible configurations of γ by

$$p(c_{\gamma_k=1} | \sigma_c, \sigma, \gamma, y) \propto \exp\left\{ -\frac{1}{2\sigma^2} (c_{\gamma_k=1} - \mu_{\gamma_k=1})^{\mathrm{T}} \Sigma_{\gamma_k=1}^{-1} (c_{\gamma_k=1} - \mu_{\gamma_k=1}) \right\},$$

with

$$\Sigma_{\gamma_k=1} \triangleq (\tilde{G}_{\gamma_k=1}^{\mathrm{T}} \tilde{G}_{\gamma_k=1} + \sigma^2 D_{\gamma_k=1}^{-2})^{-1},$$

$$\mu_{\gamma_k=1} \triangleq \Sigma_{\gamma_k=1} \tilde{G}_{\gamma_k=1}^{\mathrm{T}} y,$$

and

$$D_{\gamma_k=1} \triangleq \mathrm{diag}\{(\sigma_{c_{k_0}} \sigma_{c_{k_0}} \sigma_{c_{k_1}} \sigma_{c_{k_1}} \ldots)\}, \qquad k_i \in \{k' : \gamma_{k'} = 1\},$$

where it is understood that $\sigma_{c_{\gamma_k=1}}$, $\tilde{G}_{\gamma_k=1}$, $\Sigma_{\gamma_k=1}$, $\mu_{\gamma_k=1}$ and $D_{\gamma_k=1}$ are implicitly defined in accordance with $c_{\gamma_k=1}$, i.e. in a manner corresponding to the terms in the model specified by $\{\gamma_k : \gamma_k = 1\}$. Thus, the vector of included Gabor coefficients $c_{\gamma_k=1}$ may be drawn from a multivariate normal distribution having mean $\mu_{\gamma_k=1}$ and covariance matrix $\sigma^2 \Sigma_{\gamma_k=1}$.

Such a scheme is of course preferable when possible, as it permits Rao–Blackwellization of the posterior mean coefficient estimate; however, the prohibitively large size of many data sets that are of interest often renders a multivariate draw of the full synthesis coefficient vector infeasible. In this case $\tilde{G}$ may be partitioned as detailed in Bernardo and Smith (1994), page 138, to implement conditional subblock draws. Indeed, a judicious choice of blocking scheme also eliminates the storage problems that are associated with $\tilde{G}$ and $\tilde{G}^{\mathrm{T}}\tilde{G}$. Moreover, in the tight frame implementation that is considered here (see Section 2.1), the subblocks of $\tilde{G}$ containing all modulations at a particular time sampling point n on the time–frequency lattice will differ only by a phase factor, modulo the redundancy of the chosen Gabor system (Dörfler, 2001). This means that only one Gabor submatrix needs to be stored, along with the necessary phase factors according to the system redundancy.

A.3. Marginalization of ϕ for the case of an independent Bernoulli prior $p(\gamma_k | \phi)$

To implement the joint sampling step for γ_k and c_k that is described in step (b) of Section 4, we require the ratio $p(\gamma_k = 1 | \gamma_{-k}) / p(\gamma_k = 0 | \gamma_{-k})$. For a Bernoulli prior on γ_k for all k, we have that $\Pr(\{\gamma_k = 1\} | \phi) = \phi$;

combined with the hyperprior specification $p(\phi) = \mathcal{B}(\alpha_\phi, \beta_\phi)$ it follows that

$$\begin{aligned}\frac{p(\gamma_k = 1|\gamma_{-k})}{p(\gamma_k = 0|\gamma_{-k})} &= \frac{\int p(\gamma_k = 1, \gamma_{-k}|\phi)p(\phi)\,\mathrm{d}\phi}{\int p(\gamma_k = 0, \gamma_{-k}|\phi)p(\phi)\,\mathrm{d}\phi} \\ &= \frac{\Gamma(|\gamma_{-k}| + 1 + \alpha_\phi)\,\Gamma(K - |\gamma_{-k}| - 1 + \beta_\phi)}{\Gamma(|\gamma_{-k}| + \alpha_\phi)\,\Gamma(K - |\gamma_{-k}| + \beta_\phi)} \\ &= \frac{|\gamma_{-k}| + \alpha_\phi}{K - |\gamma_{-k}| - 1 + \beta_\phi},\end{aligned}$$

where $|\gamma_{-k}|$ is the cardinality of the set $\{\gamma'_k : \gamma'_k \in \gamma_{-k}, \gamma'_k = 1\}$, and we note that it is possible to obtain a uniform hyperprior for the case $\alpha_\phi = \beta_\phi = 1$.

A.4. Sampling the transition probabilities $\phi^m_{0,0}$ and $\phi^m_{1,1}$

As summarized in step (c) of Section 4, the Markov chain transition probabilities $\phi^m_{0,0}$ and $\phi^m_{1,1}$ defined for each modulation index $m \in \{0, 1, \ldots, M/2\}$ are assigned independent beta priors such that $\phi^m_{0,0} \sim \mathcal{B}(\alpha_{\phi^m_{0,0}}, \beta_{\phi^m_{0,0}})$ and $\phi^m_{1,1} \sim \mathcal{B}(\alpha_{\phi^m_{1,1}}, \beta_{\phi^m_{1,1}})$; the initial distribution of each chain is taken to be its stationary distribution. In this case the transition probabilities may be sampled by using a Metropolis–Hastings step, for which we use as a proposal distribution the respective full conditional distribution for a fixed, uniformly distributed initial state:

$$\begin{aligned}q(\phi^{m*}_{0,0}|\phi^{m(i)}_{0,0}) &\triangleq \mathcal{B}\{|\phi^m_{0,0}(00)| + \alpha_{\phi^m_{0,0}}, |\phi^m_{0,0}(01)| + \beta_{\phi^m_{0,0}}\}, \\ q(\phi^{m*}_{1,1}|\phi^{m(i)}_{1,1}) &\triangleq \mathcal{B}\{|\phi^m_{1,1}(11)| + \alpha_{\phi^m_{1,1}}, |\phi^m_{1,1}(10)| + \beta_{\phi^m_{1,1}}\},\end{aligned}$$

where $|\phi^m_{i,i}(ij)|$ is defined as the cardinality of the set $\{\gamma_{m,n} = j | \gamma_{m,n-1} = i\}$.

The stationary distribution P_m of the mth chain is given by

$$P_m = \begin{cases} \dfrac{1 - \phi^m_{1,1}}{2 - \phi^m_{1,1} - \phi^m_{0,0}} & \text{if } \gamma_{m,0} = 0, \\ \dfrac{1 - \phi^m_{0,0}}{2 - \phi^m_{1,1} - \phi^m_{0,0}} & \text{otherwise,} \end{cases}$$

and hence at the $(i+1)$th sweep of the sampler we take

$$\phi^{m(i+1)} = \begin{cases} \phi^{m*} \text{ with probability } p(\phi^{m(i)}, \phi^{m*}), \\ \phi^{m(i)} \text{ with probability } 1 - p(\phi^{m(i)}, \phi^{m*}), \end{cases}$$

$$p(\phi^{m(i)}, \phi^{m*}) = \min\left(\frac{P^*_m}{P^{(i)}_m}, 1\right).$$

References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc.* B, **60**, 725–749.

Barnett, G., Kohn, R. and Sheather, S. (1996) Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *J. Econometr.*, **74**, 237–254.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Besag, J. E. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Brown, P. J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc*. B, **60**, 627–641.

Brown, P. J., Vannucci, M. and Fearn, T. (2002) Bayes model averaging with selection of regressors. *J. R. Statist. Soc*. B, **64**, 519–536.

Carter, C. K. and Kohn, R. (1997) Semiparametric Bayesian inference for time series with mixed spectra. *J. R. Statist. Soc*. B, **59**, 255–268.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001) Atomic decomposition by basis pursuit. *SIAM Rev.*, **43**, 129–159.

Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997) Adaptive bayesian wavelet shrinkage. *J. Am. Statist. Ass.*, **92**, 1413–1421.
Clyde, M., Parmigiani, G. and Vidakovic, B. (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–401.
Coifman, R. R. and Meyer, Y. (1991) Remarques sur l'analyse de Fourier à fenêtre. *C. R. Acad. Sci.*, **312**, 259–261.
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.
Daubechies, I., Jaffard, S. and Journé, J.-L. (1991) A simple Wilson orthonormal basis with exponential decay. *SIAM J. Math Anal.*, **22**, 554–573.
Donoho, D. L. and Huo, X. (1999) Uncertainty principles and ideal atomic decompositions. *Technical Report 1999-13*. Department of Statistics, Stanford University, Stanford.
Donoho, D. L., Vetterli, M., Daubechies, I. and DeVore, R. A. (1998) Data compression and harmonic analysis. *IEEE Trans. Inform. Theory*, **44**, 2435–2476.
Dörfler, M. (2001) Time-frequency analysis for music signals: a mathematical approach. *J. New Mus. Res.*, **30**, 3–12.
Doucet, A., de Freitas, J. F. G. and Gordon, N. J. (eds) (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer.
Feichtinger, H. G. and Strohmer, T. (eds) (1998) *Gabor Analysis and Algorithms: Theory and Applications*. Boston: Birkhäuser.
Figueiredo, M. A. T. (2001) Adaptive sparseness using Jeffreys prior. In *Advances in Neural Information Processing Systems*, vol. 14 (eds T. G. Dietterich, S. Becker and Z. Ghahramani). Cambridge: MIT Press.
Gallant, A. R. and Monahan, J. F. (1985) Explicitly infinite-dimensional Bayesian analysis of production technologies. *J. Econometr.*, **30**, 171–201.
Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.
George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.
George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.
Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
Giovannelli, J.-F. and Idier, J. (2001) Bayesian interpretation of periodograms. *IEEE Trans. Signal Process.*, **49**, 1388–1396.
Godsill, S. J. (1999) MCMC and EM-based methods for inference in heavy-tailed processes with alpha-stable innovations. In *Proc. Institute of Electrical and Electronic Engineers Signal Processing Wrkshp Higher-order Statistics, Caesarea*. New York: Institute of Electrical and Electronic Engineers.
Godsill, S. J. (2001) On the relationship between MCMC model uncertainty methods. *J. Comput. Graph. Statist.*, **10**, 230–248.
Godsill, S. J. and Rayner, P. J. W. (1998) Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. *IEEE Trans. Speech Audio Process.*, **6**, 352–372.
Gröchenig, K. (2001) *Foundations of Time–Frequency Analysis*. Boston: Birkhäuser.
Kozek, W. (1998) Adaptation of Weyl-Heisenberg frames to underspread environments. In *Gabor Analysis and Algorithms: Theory and Applications* (eds H. G. Feichtinger and T. Strohmer), ch. 10, pp. 323–352. Boston: Birkhäuser.
Lenk, P. J. (1999) Bayesian inference for semiparametric regression using a Fourier representation. *J. R. Statist. Soc.* B, **61**, 863–879.
Lewicki, M. S. and Sejnowski, T. J. (2000) Learning overcomplete representations. *Neur. Computn*, **12**, 337–365.
Mallat, S. G. and Zhang, Z. (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, **41**, 3397–3415.
Malvar, H. S. (1990) Lapped transforms for efficient transform/subband coding. *IEEE Trans. Speech Audio Process.*, **38**, 969–978.
Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H. and Patil, P. (1998) Exact risk analysis of wavelet regression. *J. Comput. Graph. Statist.*, **7**, 278–309.
Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *J. Am. Statist. Ass.*, **83**, 1023–1036.
Müller, P. and Vidakovic, B. (eds) (1999) Bayesian inference in wavelet-based models. *Lect. Notes Statist.*, **141**.
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.
Ng, W.-J. (2000) Noise reduction for audio signals using the Gabor expansion. *MPhil Thesis*. University of Cambridge, Cambridge.
Olshausen, B. A. and Millman, K. J. (2000) Learning sparse codes with a mixture-of-Gaussians prior. In *Advances in Neural Information Processing Systems*, vol. 12 (eds S. A. Solla, T. K. Leen and K.-R. Müller), pp. 841–847. Cambridge: MIT Press.
Poggio, T. and Girosi, F. (1998) A sparse representation for function approximation. *Neur. Computn*, **10**, 1445–1454.
Priestley, M. B. (1981) *Spectral Analysis and Time Series*. London: Academic Press.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.
Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer.
Strohmer, T. (1998) Numerical algorithms for discrete Gabor expansions. In *Gabor Analysis and Algorithms: Theory and Applications* (eds H. G. Feichtinger and T. Strohmer), ch. 8, pp. 267–294. Boston: Birkhäuser.
Tipping, M. E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learnng Res.*, **1**, 211–244.
Vidakovic, B. (1998) Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Am. Statist. Ass.*, **93**, 173–179.
Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. New York: Wiley.
Wahba, G. (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.
West, M. (2003) Bayesian factor regression models in the 'large p, small n' paradigm. 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), pp. 733–742. Oxford: Oxford University Press.
Wilson, K. G. (1987) Generalized Wannier functions. *Preprint*. Cornell University, Ithaca.
Wolfe, P. J. (2003) Perceptually motivated approaches to audio signal enhancement: broad-band noise reduction via Bayesian modelling of time-frequency coefficients. *PhD Thesis*. University of Cambridge, Cambridge.
Wolfe, P. J., Dörfler, M. and Godsill, S. J. (2001) Multi-Gabor dictionaries for audio time-frequency analysis. In *Proc. Institute of Electrical and Electronic Engineers Wrkshp Applications of Signal Processing to Audio and Acoustics*, pp. 43–46. New York: Institute of Electrical and Electronic Engineers.
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.