J. R. Statist. Soc. B (2006)
68, *Part* 2, *pp.* 305–332

Functional clustering by Bayesian wavelet methods

Shubhankar Ray and Bani Mallick

Texas A&M University, College Station, USA

[Received April 2004. Final revision October 2005]

Summary. We propose a nonparametric Bayes wavelet model for clustering of functional data. The wavelet-based methodology is aimed at the resolution of generic global and local features during clustering and is suitable for clustering high dimensional data. Based on the Dirichlet process, the nonparametric Bayes model extends the scope of traditional Bayes wavelet methods to functional clustering and allows the elicitation of prior belief about the regularity of the functions and the number of clusters by suitably mixing the Dirichlet processes. Posterior inference is carried out by Gibbs sampling with conjugate priors, which makes the computation straightforward. We use simulated as well as real data sets to illustrate the suitability of the approach over other alternatives.

Keywords: Functional clustering; Mixture of Dirichlet processes; Nonparametric Bayesian model; Wavelets

1. Introduction

Functional data arise in a wide variety of applications and are often clustered to reveal differences in the sources or to provide a concise picture of the data. For instance, clustered gene expression profiles from microarrays may point to underlying groups of functionally similar genes. Model-based clustering relies largely on finite mixture models to specify the cluster-specific parameters (Banfield and Raftery, 1993; Yeung *et al.*, 2001) assuming that the number of clusters is known in advance. This approach is unreasonable in practice, as it relies on one's ability to determine the correct number of clusters. Medvedovic and Sivaganesan (2002) used the Dirichlet-process- (DP) based infinite mixture model to overcome these deficiencies. None-the-less, all these approaches use multivariate normal distributions for modelling and disregard the functional form of the data.

This 'functional' approach was pursued only recently in a mixed effects spline model by James and Sugar (2003) and in the context of yeast cell cycle data analysis using periodic basis modelling by Wakefield *et al.* (2003). However, shifts in global and local characteristics in the functional data may not be detectable in these frameworks. As an example, the gene expression profiles of yeast cell cycles may occasionally depart from the usual cyclic behaviour and these shifts will be overlooked, in general, by the periodic basis model.

The Bayesian wavelet modelling that is used in this paper manages to overcome these limitations as wavelets have nice approximation properties over a large class of functional spaces (Daubechies, 1992) that can accommodate almost all the functional forms that are observed in real life applications. Indeed, this richness of the wavelet representation provides the back-bone for the popular frequentist wavelet shrinkage estimators of Donoho and Johnstone (1994, 1995),

Address for correspondence: Bani Mallick, Department of Statistics, TAMU 3143, Texas A&M University, College Station, TX 77843-3143, USA.
E-mail: bmallick@stat.tamu.edu

1369–7412/06/68305

which are the precursors of the more recent Bayesian wavelet estimation models (Abramovich *et al.*, 1998; Vidakovic, 1998; Clyde and George, 2000; Clyde *et al.*, 1998). Wavelet representations are also sparse and can be helpful in limiting the number of regressors. Dimension reduction is not inherent in other models; for example, this was done in James and Sugar (2003) by attaching an additional dimension reducing step on the spline coefficients. In Bayes wavelet modelling this is effortlessly achieved by a selection mixture prior to isolate a few significant coefficients for the collection of functions.

The nonparametric Bayes clustering model that is presented here is based on the mixture of DPs (Ferguson, 1973; Antoniak, 1974). The DP provides a rich two-parameter conjugate prior family and much of the posterior inference for a particular parametric conjugate family applies, when the same prior becomes the base measure for a DP prior, which is used instead. The base prior modelling of the wavelet coefficients can be motivated by traditional hierarchical wavelet models and allows the specification of the smoothness and regularity properties of the functional realizations from the DP. There are several advantages in the context of clustering. The computation is straightforward owing to the Gibbs sampling schemes that were proposed in the mid-1990s (Escobar and West, 1995) and the numbers of clusters are automatically determined in the process. In addition, the Bayesian methodology provides a direct way to predict any missing observations, extending the applicability of the model to incomplete data sets.

The paper is organized as follows. In Section 2, we overview the parametric Bayesian models for wavelet shrinkage. This is later extended to the DP-based nonparametric model in Section 3 and the posterior inference is detailed in Section 4. Some properties of the clustering model are discussed in Section 5. Finally, Section 6 addresses the simulations with a discussion of the model selection and the missing data case.

2. Hierarchical wavelet model

Consider a collection of unknown functions $\{f_i\}$, $i \in \{1, \ldots, n\}$, on the unit interval that are observed with white Gaussian noise at m equispaced time points as

$$y_{i,k} = f_i(k/m) + \varepsilon_{i,k}, \qquad \varepsilon_{i,k} \sim N(0, \sigma_i^2)$$

where $k \in \{1, \ldots, m\}$ and m is a power of 2. In a gene microarray, for example, the observed curve $y_{i,k}$ is the response profile at the kth time point for the ith gene. In nonparametric estimation, the functions are analysed in the sequence space of coefficients in an orthonormal wavelet basis for $L_2([0,1])$. Restriction of the functions to the unit interval introduces discontinuities at the edges. Boundary artefacts can be avoided by periodized bases when the functions are periodic (Daubechies, 1992); otherwise boundary folding or reflection extensions are used to improve the behaviour at the boundaries.

Wavelet representations are sparse for a wide variety of function spaces and their multiresolution nature enables us to combine results from different resolutions and to make conclusions for the estimation problem. In particular, the sparseness implies that, when the wavelet basis is orthogonal and compactly supported (Daubechies, 1992), the independent and identically distributed (IID) normal noise affects all the wavelet coefficients equally, whereas the signal information remains isolated in a few coefficients. In shrinkage estimation, these small coefficients, which are mostly noise, are discarded to retrieve an effective reconstruction of the function. The expansion of f_i in terms of periodized scaling and wavelet functions (φ, ψ) has the dyadic form

$$f_i(t) \approx \beta_{i00}\, \varphi_{00}(t) + \sum_{j=1}^{J} \sum_{k=0}^{2^{j-1}} \beta_{ijk}\, \psi_{jk}(t) \qquad (1)$$

where β_{i00} is the scaling coefficient, β_{ijk} are the detail coefficients and $J=\log_2(m)$ is the finest level of the wavelet decomposition. Wavelets also provide a direct way to study the functional smoothness in that the wavelet representation usually contains all the information that can tell whether f_i lies in a smoothness space.

2.1. Wavelet estimation models

Wavelet shrinkage estimation was popularized by Donoho and Johnstone (1994, 1995), who showed that thresholding rules on the empirical detail coefficients provide optimal estimators for a broad range of functional spaces. Bayesian wavelet shrinkage proceeds by eliciting mixture priors over the detail coefficients β_{ijk} $(j>0)$, with a degenerate mass at zero (Abramovich *et al.*, 1998; Clyde and George, 2000) for selection,

$$\beta_{ijk} \sim \pi_j\, N(0, g_j\sigma^2) + (1-\pi_j)\delta_0. \tag{2}$$

The scaling coefficients β_{i00}, in contrast, are usually modelled by a vague prior. The selection probabilities π_j and the scaling parameters g_j allow us to place our prior belief by level, producing a simple estimation strategy that is more adaptive than the classical analogues of hard or soft thresholding.

In linear model notation, if $\mathbf{Y}_i=(y_{i,1},\ldots,y_{i,m})$ is the vector of m observations from the ith unit, the regression model is

$$\mathbf{Y}_i=\mathbf{X}\boldsymbol{\beta}_i+\varepsilon_i, \qquad i=1,\ldots,n, \tag{3}$$

where $\boldsymbol{\beta}_i=(\beta_{i00},\beta_{i10},\beta_{i20},\beta_{i21},\ldots)^{\mathrm{T}}$ are the wavelet coefficients for f_i after the discrete wavelet transformation $\mathbf{X}$ and $\varepsilon_i \sim N(0,\sigma_i^2\mathbf{I}_m)$. The selection priors (2) are conveniently incorporated as a scale mixture with latent indicator variables γ_{jk} that equal 1 with probability π_j (Clyde and George, 2000; Clyde *et al.*, 1998; De Canditiis and Vidakovic, 2004). The effective joint prior for the coefficients and the model variance is

$$\boldsymbol{\beta}_i,\sigma_i^2|\mathbf{V}\sim \mathrm{NIG}(\mathbf{0},\mathbf{V};u,v)$$

where NIG denotes the normal–inverse gamma prior—the product of the conditionals $\boldsymbol{\beta}_i|\sigma_i^2,\mathbf{V}\sim N(0,\sigma_i^2\mathbf{V})$ and $\sigma_i^2\sim\mathrm{IG}(u,v)$ with u and v as the usual hyperparameters for the inverse gamma prior.

The diagonal matrix $\mathbf{V}$ can be used to obtain a scale mixture prior; we let

$$\mathbf{V}=\mathrm{diag}(\boldsymbol{\gamma})\,\mathrm{diag}(\mathbf{g})$$

where $\boldsymbol{\gamma}=(\gamma_{00},\gamma_{10},\gamma_{20},\gamma_{21},\ldots)$ is a vector of latent indicator variables for selection of each coefficient and $\mathbf{g}=(g_0,g_1,g_2,g_2,\ldots)$ comprise the corresponding scaling parameters given by

$$\gamma_{j,k}\sim\mathrm{Bernoulli}(\pi_j),$$
$$g_j\sim\mathrm{IG}(r_j,s_j)$$

where (r_j,s_j) are hyperparameters that are specified levelwise. This hierarchical layer is especially useful for modelling sparse wavelet representations, which otherwise requires Laplace-like sharp non-conjugate priors (Vidakovic, 1998). In particular, there is the flexibility of controlling our prior belief about the scaling coefficient β_{i00} by letting $\pi_0=1$ and tuning (r_0,s_0) to vary $\mathrm{var}(g_0)$.

To summarize the hierarchical wavelet model, we have

$$\mathbf{Y}_i|\boldsymbol{\beta}_i,\sigma_i^2 \sim N(\mathbf{X}\boldsymbol{\beta}_i,\sigma_i^2\mathbf{I}_m),$$
$$\boldsymbol{\beta}_i,\sigma_i^2|\gamma,\mathbf{g} \sim \text{NIG}(0,\mathbf{V};u,v),$$
$$g_j \sim \text{IG}(r_j,s_j),$$
$$\gamma_{j,k} \sim \text{Bernoulli}(\pi_j),$$

for $i \in \{1,\ldots,n\}$, $j \in \{0,\ldots,J\}$ and $k \in \{0,\ldots,2^{j-1}\}$.

3. Wavelet clustering model

In the clustering model, the parameters $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i,\sigma_i^2)$ for the underlying functions f_i are elicited by DP priors. DPs are almost surely discrete and comprise a certain partitioning of the parameter space that is needed for clustering. More precisely, the sequence of parameters $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_n$ comes from a random distribution F that is a realization from a DP $D(\alpha,H_\phi)$ depending on a precision parameter α and base prior $H_\phi = E(F)$ with ϕ as the parameters of H. The nonparametric hierarchical model is completed by mixing the base prior for the DP with the hyperpriors of Section 2.1 and is described as

$$\mathbf{Y}_i|\boldsymbol{\beta}_i,\sigma_i^2 \sim N(\mathbf{X}\boldsymbol{\beta}_i,\sigma_i^2\mathbf{I}_m), \tag{4}$$
$$\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_n \sim F,$$
$$F \sim D\{\alpha,\text{NIG}(0,\mathbf{V};u,v)\}, \tag{5}$$
$$g_j \sim \text{IG}(r_j,s_j), \tag{6}$$
$$\gamma_{j,k} \sim \text{Bernoulli}(\pi_j),$$
$$\alpha \sim G(d_0,\eta_0).$$

Here $\phi = \{\mathbf{g},\gamma\}$.

The underlying clustering properties of the DP are easier to appreciate in its Pölya urn representation (Blackwell and MacQueen, 1973). This connection is also used later to perform sequential flexible Gibbs sampling of the clustering parameters $\boldsymbol{\theta}_i$ as in Escobar and West (1995). In a sequence of draws $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots$ from the Pölya urn the nth sample is either distinct with a small probability $\alpha/(\alpha+n-1)$ or is tied to a previous sample with positive probability to form a cluster. Let $\boldsymbol{\theta}_{-n} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n\}\backslash\boldsymbol{\theta}_n$ and d_{n-1} be the number of pre-existing clusters of tied samples in $\boldsymbol{\theta}_{-n}$ at the nth draw; then we have

$$\boldsymbol{\theta}_n|\boldsymbol{\theta}_{-n},\alpha,\phi = \frac{\alpha}{\alpha+n-1}H_\phi + \sum_{i=1}^{d_{n-1}} \frac{n_i}{\alpha+n-1}\delta_{\bar{\boldsymbol{\theta}}_i} \tag{7}$$

where $H_\phi = \text{NIG}(\mathbf{0},\mathbf{V};u,v)$ is the base prior, $\phi = \{\mathbf{g},\gamma\}$ and the ith cluster has n_i tied samples that are commonly expressed by $\bar{\boldsymbol{\theta}}_i = (\bar{\boldsymbol{\beta}}_i,\bar{\sigma}_i^2)$ and

$$\sum_{i=1}^{d_{n-1}} n_i = n-1.$$

In the long term of sequential draws, the number of clusters d_n is much less than n and is determined by the precision α. We also use the set $\mathcal{C}_n$ containing the clustering profile at the nth draw, such that $\mathcal{C}_n(i)$ is the set of indices of all the curves in the ith cluster.

The modelling of the base prior H_ϕ is important and is reflected in all the realizations from the DP which are centred at $H_\phi = \mathrm{NIG}(0, \mathbf{V}; u, v)$. As before, the detail coefficients are conveniently modelled by a selection prior. The scaling coefficients are modelled by using priors for which the hyperparameters are empirically estimated and in Section 6 we show that this in general works better than vague priors.

We consider two models that differ in the way that the error terms are modelled. Model 1 is a heteroscedastic model with $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma_i^2)$ and can be useful to handle potential fluctuations in variability across clusters or to check the normality assumptions in model (3). Model 2 is an oversmoothed homoscedastic model with $\sigma_i^2 = \sigma^2$ for all i, which is computationally more straightforward. The fact that the d_n separate draws of variance in model 1 are replaced by a single draw imparts more stability to the Markov chain. This makes it suitable for cases where the population is not overly heteroscedastic.

The nonparametric model does not allow direct control over the number of clusters, but it offers the parameter α that determines the expected and the asymptotic number of clusters: $d_n/\log(n) \rightarrow \alpha$ almost surely (Korwar and Hollander, 1973). Minor subjective reservations aside, we think that this vagueness does not affect the inference in practice.

3.1. Choice of hyperparameters

The specification of (g_j, π_j) represents our prior belief about the collection of curves at each level. A variety of scale mixture or shrinkage priors (Vidakovic, 1998; Clyde and George, 2000) have been proposed for robust and heavy-tailed modelling of wavelet coefficients. All these specifications comprise different ways of modelling the decay of wavelet coefficients, relating them to functional smoothness. In particular, Abramovich *et al.* (1998) showed that these parameters can be specified such that the functions fall in Besov spaces—a valuable back-bone for modelling a broad range of smoothness and spatial adaptation properties. For the Besov space $\mathcal{B}^l_{p,q}$ $(l > 0, 1 < p, q \leqslant \infty)$, l gives the order of smoothness in $L_p([0,1])$, p controls the spatial inhomogeneity and the parameter q allows fine distinctions in the smoothness of fixed order l.

In general, for a function with an almost surely finite wavelet representation the smoothness is the same as that of the mother wavelet ψ. Suppose that $\psi \in \mathcal{B}^l_{p,q}$ $(1 \leqslant p, q \leqslant \infty)$ has ζ vanishing moments satisfying $\max(0, 1/p - \frac{1}{2}) < l < \zeta$. For $j > 0$, fixing $g_j = 2^{-aj} c_1$ and $\pi_j = \min(1, 2^{-bj} c_2)$ with $c_1, c_2, a > 0$ and $b > 1$ gives an almost surely finite wavelet series and f_i also belongs to $\mathcal{B}^l_{p,q}$. Abramovich *et al.* (1998) extended the equivalence for $b \in [0,1]$, if (a,b) are chosen to satisfy

$$(b-1)/p + (a-1)/2 \geqslant l - 1/p \tag{8}$$

with equality when $q = \infty$ and $1 \leqslant p < \infty$. For convenience, we shall refer to (a,b) as the Besov parameters.

For any (a,b) the realizations lie in a family of Besov spaces that are described by relationship (8). By fixing a, we see a direct relationship between b and the positive range of p. In effect b controls the sparseness and the spatial inhomogeneity of the functional realizations. In a similar manner, a defines the positive range of l and therefore controls the overall decay of the wavelet coefficients and the smoothness. This analogy can be drawn directly by looking at the decay of the wavelet coefficients that are realized by the prior distributions. Loosely, by increasing a or b we realize functions with higher *effective* smoothness, since in inequality (8) $l > 1/p$ implies that functions in B^l_{pq} are continuous. Apart from the interesting connection with smoothness, in practice, greater flexibility is achieved by updating the probabilities π_j. These parameters can also affect the model's ability to identify clusters. For instance, smoother functions on average cluster more tightly and should result in a drop in the number of clusters.

Recall that, in our hierarchical model, the scaling parameters g_j were mixed by a conjugate prior (6) and, in general, this imparts greater flexibility than subjective deterministic specifications. It also allows modelling within the Besov spaces through restrictions that are imposed on the inverse gamma hyperparameters (r_j, s_j). For $j>0$ and a fixed integer $p \geqslant 1$, if the hyperparameters satisfy

$$E(g_j^{p/2})^{1/p} = \frac{r_j^{1/2}}{\{(s_j-2)(s_j-4)\dots(s_j-p)\}^{1/p}} = 2^{-aj}c_1 \quad \text{and} \quad \pi_j = \min(1, 2^{-bj}c_2)$$

with (a,b) satisfying distribution (2), then the Besov correspondence still holds. The proof (see Appendix A) is similar to that of Abramovich *et al.* (1998), except for the use of the marginal t-distribution after averaging out g_j. A simple way to choose the inverse gamma hyperparameters is to fix $s_j = p+1$ and to calculate r_j to satisfy the foregoing relationship for given values of c_1 and a.

Abramovich *et al.* (1998) used fixed values of (a,b) based on the prior knowledge about the function's regularity, followed by method-of-moments estimators to calculate the constants c_1 and c_2. We maximize the marginal likelihood (Clyde and George, 2000) with respect to the base prior H_ϕ to estimate the constants while fixing a and b:

$$L(c_1, c_2) \propto -\log|\mathbf{V}_n^*| - (v+mn)\log\left\{u + \sum_{i=1}^n \mathbf{Y}_i'\mathbf{Y}_i - \boldsymbol{\mu}_n^{*\prime}(\mathbf{V}_n^*)^{-1}\boldsymbol{\mu}_n^*\right\}$$

where

$$\boldsymbol{\mu}_n^* = \mathbf{V}_n^* \sum_{i=1}^n \mathbf{X}'\mathbf{Y}_i$$

and

$$\mathbf{V}_n^* = (\mathbf{V}^{-1} + n\mathbf{I}_m)^{-1}.$$

For fixed values of a and b, a gridded maximization procedure can lead to estimates of c_1 and c_2. Moreover, an objective selection of the hyperparameters (a,b) can be performed by extending the maximization procedure to a grid in (a,b), as illustrated in Section 6.5. Ideally, the marginal likelihood can be calculated for the DP priors by using the collapsed sequential importance sampling methods of Basu and Chib (2003). This procedure can be computationally intensive depending on the size of the data set. In general, for small data sets, it leads to estimates that are comparable with the marginal maximum likelihood estimates using H_ϕ.

4. Posterior inference

Adopting base priors that are conjugate to the likelihood expedites the posterior sampling of the clustering parameters $\boldsymbol{\theta}_i$ from the Pölya urn. We also retain the computational advantage of the scale mixture form of the base prior and the conditional posteriors for all the indicator variables γ_{jk} and the scale parameters g_j are in standard form.

The conditional posterior distributions for the heteroscedastic case are derived here. The corresponding conditionals for the special homoscedastic case easily follow from these derivations.

4.1. Conditional distributions for clustering

To update the clustering parameters $\boldsymbol{\theta}_n = (\boldsymbol{\beta}, \sigma^2)_n$, we combine likelihood (4) with the Pölya urn prior (7):

$$(\boldsymbol{\beta}, \sigma^2)_n | (\boldsymbol{\beta}, \sigma^2)_{-n}, \alpha, \boldsymbol{\phi}, \mathbf{Y}_n \propto q_{n0}\, H^*_{\phi,n}(\beta, \sigma^2) + \sum_{i=1}^{d_{n-1}} q_{ni} \delta_{(\bar{\boldsymbol{\beta}}, \bar{\sigma}^2)_i} \tag{9}$$

where $H^*_{\phi,n} = \mathrm{NIG}(\boldsymbol{\mu}^*, \mathbf{V}^*; u^*_n, v^*_n)$ is H_ϕ *a posteriori* with

$$\begin{aligned} \mathbf{V}^* &= (\mathbf{V}^{-1} + \mathbf{I}_m)^{-1}, \\ \boldsymbol{\mu}^* &= \mathbf{V}^* \mathbf{X}' \mathbf{Y}_n, \\ u^*_n &= u + \mathbf{Y}'_n \mathbf{Y}_n - \boldsymbol{\mu}^{*\prime} (\mathbf{V}^*)^{-1} \boldsymbol{\mu}^*, \\ v^*_n &= v + m. \end{aligned}$$

The weights $q_{n\cdot}$ follow from the likelihood and its marginals in the normal–inverse gamma conjugate family (Escobar and West, 1995) and determine the posterior inclination towards new distinct samples. We have

$$q_{ni} \propto n_i\, \phi(\mathbf{Y}_n | \mathbf{X} \bar{\boldsymbol{\beta}}_i, \bar{\sigma}^2_i \mathbf{I}_m)$$

as the conditional distribution of $\mathbf{Y}_n$ given the ith cluster or distinct sample $(\bar{\boldsymbol{\beta}}, \bar{\sigma}^2)_i$ and

$$q_{n0} \propto \alpha\, t_m \{\mathbf{0}, u(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}')\}$$

is the marginal distribution.

Similarly, setting $H_\phi = N(\mathbf{0}, \sigma^2 \mathbf{V})$ for the homoscedastic model gives

$$H^*_{\phi,n} = N(\boldsymbol{\mu}^*, \sigma^2 \mathbf{V}^*)$$

with

$$\begin{aligned} q_{n0} &\propto \alpha\, \phi\{\mathbf{0}, \sigma^2(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}')\}, \\ q_{ni} &\propto n_i\, \phi(\mathbf{Y}_n | \mathbf{X}\bar{\boldsymbol{\beta}}_i, \sigma^2 \mathbf{I}_m). \end{aligned}$$

The posterior distribution of σ^2 is simply $(\sigma^2 | \mathbf{Y}, \phi) \sim \mathrm{IG}(u^*, v^*)$ where

$$\begin{aligned} u^* &= u + \sum_{i=1}^{d_n} \Big\{ \sum_{j \in \mathcal{C}_n(i)} \mathbf{Y}'_j \mathbf{Y}_j - \boldsymbol{\mu}^{*\prime}_i (\mathbf{V}^*_i)^{-1} \boldsymbol{\mu}^*_i \Big\}, \\ v^* &= v + mn, \\ \boldsymbol{\mu}^*_i &= \mathbf{V}^*_i \sum_{j \in \mathcal{C}_n(i)} \mathbf{X}' \mathbf{Y}_j, \\ \mathbf{V}^*_i &= (\mathbf{V}^{-1} + n_i \mathbf{I}_m)^{-1}. \end{aligned}$$

Sampling requires computation of the mixture probabilities q_{ni} for the distinct pre-existing parameter values, which are small compared with n. The sequential update of model parameters from the Pölya urn model can be randomized to preclude any ordering-related bias. This is followed by a resampling step (Bush and MacEachern, 1996) that expedites model mixing by literally shaking up the converged mixture model.

4.2. Posterior sampling of the mixing parameters

For notational convenience, the parameters are grouped by the dyadic levels j of the wavelet decomposition $-\gamma_j = \{\gamma_{jk} : \forall k\}$ and $\bar{\beta}_{ij} = \{\bar{\beta}_{ijk} : \forall k\}$ for $j \geqslant 0$. To obtain the conditional posteriors for the scaling parameters g_j and the indicator variables γ_{jk}, we exploit the conditional independence of the distinct cluster parameters $\{(\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2)\}_{i=1}^{d_n}$ given the clusters $\mathcal{C}_n$. Korwar and Hollander (1973) showed that conditional on $\mathcal{C}_n$ the distinct parameters are IID H_ϕ.

For the heteroscedastic case, the scaling parameters g_j are updated levelwise by combining the base prior and distribution (4):

$$g_j | \gamma_j, \mathcal{C}_n, \{\bar{\sigma}_i^2, \bar{\boldsymbol{\beta}}_{ij}\}_{i=1}^{d_n} \sim \text{IG}(r_j^*, s_j^*),$$

where

$$s_j^* = d_n \sum_{k=0}^{2^{j-1}} \gamma_{jk} + s_j,$$

$$r_j^* = \sum_{i=1}^{d_n} \bar{\sigma}_i^{-2} \sum_{k=0}^{2^{j-1}} \gamma_{jk}^2 \bar{\beta}_{ijk}^2 + r_j.$$

For model 2, the $\bar{\sigma}_i^2$s are replaced by a single σ^2. Observe that these updates combine the information at the jth resolution across all the distinct functions and the average shrinkage is conservative and depends on the total variation at any resolution.

Similarly, for each level, the indicators $\boldsymbol{\gamma}_j$ are updated conditionally on the indicators at other levels $\boldsymbol{\gamma}_{-j}$:

$$f(\boldsymbol{\gamma}_j | \boldsymbol{\gamma}_{-j}, \mathbf{g}, \mathcal{C}_n, \mathbf{Y}) \propto \prod_{i=1}^{d_n} f(\{\mathbf{Y}_j\}_{j \in \mathcal{C}_n(i)} | \boldsymbol{\gamma}, \mathbf{g}, \mathcal{C}_n) \pi_j$$

where

$$f(\{\mathbf{Y}_j\}_{j \in \mathcal{C}_n(i)} | \boldsymbol{\gamma}, \mathbf{g}, \mathcal{C}_n) \propto C_i \frac{|\mathbf{V}_i^*|^{1/2}}{|\mathbf{V}|^{1/2}} \left\{ u + \sum_{j \in \mathcal{C}_n(i)} \mathbf{Y}_j' \mathbf{Y}_j - \boldsymbol{\mu}_i^{*\prime} (\mathbf{V}_i^*)^{-1} \boldsymbol{\mu}_i^* \right\}^{-(v+mn_i)/2} \tag{10}$$

is the marginal likelihood for the ith cluster with

$$C_i = \frac{\Gamma\{(v+mn_i)/2\}}{\pi^{n_i m/2}}$$

and $\mathbf{Y}$ depicts the collection of all responses $\{\mathbf{Y}_i\}_{i=1}^n$. The update is similar in form for model 2. Again, the selection of γ_{jk} is more conservative and is decided by the proportion of variation explained by the coefficients $\bar{\beta}_{ijk}$ at location (j,k) for all i.

4.3. Posterior sampling of precision α

We update the precision α as in Escobar and West (1995). The posterior distribution is derived by combining a gamma prior for α with the distribution of d_n:

$$f(d_n | \alpha, n) = c_n(d_n) \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \alpha^{d_n} n!. \tag{11}$$

It resembles the Cauchy formula for counting permutations and has been calculated independently by several researchers, including Antoniak (1974). Combining equation (11) with $f(\alpha)$, the prior for α, it can be shown that posterior

$$f(\alpha | d_n, n) \propto f(\alpha) \alpha^{d_n - 1} (\alpha + n) \int_0^1 x^\alpha (1-x)^{n-1} \, \mathrm{d}x.$$

Equivalently, there is a beta random variable η such that

$$f(\alpha|d_n,n)=\int f(\alpha,\eta|d_n,n)\,\mathrm{d}\eta.$$

Thus α can be updated in two steps. First, conditionally on α and d_n, update η. Second, conditionally on the last sampled value of η and d_n, draw α. When $\alpha\sim G(d_0,\eta_0)$, both conditional distributions are in standard form, given by

$$(\alpha|\eta,d_n)\sim\rho_n G\{d_0+d_n,\eta_0-\log(\eta)\}+(1-\rho_n)\,G\{d_0+d_n-1,\eta_0-\log(\eta)\},$$
$$(\eta|\alpha,d_n)\sim\mathrm{beta}(\alpha+1,n)$$

where

$$\frac{\rho_n}{1-\rho_n}=\frac{d_0+d_n-1}{n\{\eta_0-\log(\eta)\}}.$$

5. Properties of the clustering model

When sequentially sampling from the Pölya urn (9) the mixture probabilities $q_{n\cdot}$ are based on the l_2-distance between the nth observed function and the d_{n-1} previously sampled functions. Most classical clustering algorithms such as k-means and neural network clustering, or even statistical models with normal likelihoods (and priors), inherently use the l_2-distance as a measure of distance between two curves. Likewise, in James and Sugar (2003) the decision of choosing a cluster was based on the squared distance between spline coefficients.

In this section, we ascertain whether, in the long term of sequential draws, there is a minimum squared distance that would ensure a distinct sample. This distance may be viewed as the eventual minimum separation between clusters and is referred to as the *sampling resolution*. As n becomes larger the collection of sampled functions becomes populated and we expect the resolution to grow, i.e. for a new sample to be distinct it must distinguish itself more clearly from increasing population as n becomes large. The rate of this growth gives an idea about the adaptation of the clustering model. The following theorem, which is proved in Appendix A, characterizes the resolution of the DP without any mixing, i.e. while fixing the parameters $\mathbf{g}$, γ and α.

Theorem 1. For the homoscedastic model with the base prior H_ϕ such that $\|\boldsymbol{\beta}_i\|_2<\infty$ almost surely $\forall i=1,2,\ldots$, the posterior sampling resolution is $\sigma^2\,O[\log\{\log(n)^{1+\delta}\}]$, for any $\delta>0$.

Remark 1. The slow rate of increase suggests good adaptation properties of the model that does not change drastically as n becomes large.

Remark 2. Note that the condition of bounded l_2-norm can be achieved by choosing hyperparameters from distribution (2).

Remark 3. The error variance directly affects the separation between clusters in that a higher variability means that the clusters are more spread out.

6. Examples

We analyse one synthetic and two real data sets to illustrate the practical potential of the functional clustering model. The virtues of wavelet modelling are emphasized by using data sets that exhibit different degrees of smoothness and spatial inhomogeneity. The Besov class of priors that was mentioned in Section 3.1 easily accommodates the three extremes that are listed below in that b controls the inhomogeneity and the overall smoothness can be attributed to a.

(a) *Smooth but inhomogeneous*: the synthesized Doppler signals, although intrinsically continuous, the finite sampling along with the changing intensity of oscillations make the function spatially inhomogeneous.
(b) *Not smooth and inhomogeneous*: yeast cell cycle gene expression profiles in the second example may be far from continuous and are characterized by varying degrees of temporal fluctuation.
(c) *Not smooth but homogeneous*: the third example analyses a meteorological precipitation data set that exhibits a certain homogeneity in the prevailing bumpiness.

For evaluating performance the number of clusters and the misclassification rate (in supervised conditions) are reported. In addition, the robustness to missing observations is checked for synthetic data sets with different amounts of missing points and a yeast cell cycle microarray data set. All the results reported are averaged over 100 simulations with 10000 iterations per simulation and a burn-in period of 1000. In general, these specifications must vary depending on n, m and the prior estimate of the model variance. The Markov chain Monte Carlo (MCMC) algorithm mixes fairly well in all these examples and the chains seem to converge much before the allotted burn-in time. Excepting the microarray data set that is analysed below, a nearly symmetric Symmlet wavelet basis with eight vanishing moments was used in all the experiments.

6.1. Mixed effects spline model

In some cases, the results are compared with the mixed effects spline model of James and Sugar (2003) that is given by

$$\mathbf{Y}_i = \mathbf{S}\boldsymbol{\beta}_{1i} + \mathbf{S}\boldsymbol{\beta}_{2i} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2 \mathbf{I}_m), \tag{12}$$

where $\mathbf{S}$ $(m \times p)$ is a natural cubic spline design with suitably chosen knots, $\boldsymbol{\beta}_{1i}$ are cluster-specific coefficients (i.e. all responses in a cluster have the same $\boldsymbol{\beta}_{1i}$) and coefficients $\boldsymbol{\beta}_{2i}$ account for individual variations in the functions within each cluster. We do not consider the additional dimension reducing transformation that was used by James and Sugar (2003) but instead compensate with a smaller number of knots to fit higher order splines. Since the original EM implementation of this model was unavailable, Bayesian modelling was used to expedite the comparison and the following conjugate priors were used:

$$\boldsymbol{\beta}_{1i} \sim P, \qquad P \sim \mathrm{DP}\{\alpha, N(0, \sigma_i^2 \boldsymbol{\Gamma}_1)\}, \quad \boldsymbol{\Gamma}_1 \sim \mathrm{IW}(\mathbf{R}_1, s_1),$$
$$\boldsymbol{\beta}_{2i} \sim N(0, \sigma_i^2 \boldsymbol{\Gamma}_2), \qquad \boldsymbol{\Gamma}_2 \sim \mathrm{IW}(\mathbf{R}_2, s_2).$$

The posterior conditionals for $\boldsymbol{\beta}_{1i}$ and $\boldsymbol{\Gamma}_1$ follow closely from the derivations in Section 4 and inverse Wishart posteriors take the place of the inverse gamma posteriors. Conditional on the clusters—the distinct β_{1i}s and $\boldsymbol{\Gamma}_1$, there are n additional conjugate normal draws of $\boldsymbol{\beta}_{2i}$ followed by another inverse Wishart draw of $\boldsymbol{\Gamma}_2$. In the simulations, $\mathbf{R}_1 = 10.0\mathbf{I}$ and $s_1 = 5$, to give a diffuse prior for $\boldsymbol{\Gamma}_1$. Averaging out $\boldsymbol{\beta}_{1i}$ with respect to the base prior, and $\boldsymbol{\beta}_{2i}$ with respect to its normal prior, the hyperparameters $\mathbf{R}_2$ and s_2 are estimated by empirical Bayes methods.

6.2. Priors for scaling coefficients

Prior modelling of the scaling coefficients determines the sensitivity to the locational differences in the data set. We compare the traditional choice of vague or diffuse priors with a prior for which the hyperparameters have been empirically estimated.

The prior scale parameter g_0 for the scaling coefficients β_{i00} follows an inverse gamma distribution with parameters (r_0, s_0). For diffuse priors these parameters are adjusted for large prior variance. An approximate empirical estimation of these hyperparameters can be carried out by marginalizing the likelihood with respect to the base prior (of the DP) individually for each curve and then applying moment matching on the population of estimated g_0s to estimate (r_0, s_0). The marginal likelihood follows by combining the model likelihood (4) $\tilde{\beta}_{ijk} \sim N(\beta_{ijk}, \sigma^2)$, written in terms of the empirical wavelet coefficients $\tilde{\beta}_{ijk}$, with the prior $\beta_{ijk} \sim N(0, \gamma_{jk} g_j \sigma^2)$. For the scaling coefficients, we have $\tilde{\beta}_{i00} \sim N\{0, (1+g_0)\sigma^2\}$ and g_0 for the ith curve is simply estimated as $\tilde{\beta}^2_{i00}/\sigma^2 - 1$, where σ^2 can be estimated from the finer levels of wavelet decomposition. The median and the one-third range of the population of n such estimates are matched with

$$E(g_0) = \frac{r_0}{s_0 - 2},$$

$$\mathrm{var}(g_0) = \frac{2r_0^2}{(s_0-2)^2(s_0-4)}$$

respectively, to generate rough estimates of (r_0, s_0).

From a comparative study of the two priors applied to various example data sets, the empirically estimated prior seems to be a more reliable choice and to lead to better cluster estimates than the diffuse prior. In all the examples that are furnished below, empirically estimated priors were used for the scaling coefficients. For illustrative purposes, a comparative study is also discussed in one of the examples.

6.3. Model choice

The state space of possible clustering combinations can be very large and some model selection criterion is required to decide between the best models. Recently, Quintana and Iglesias (2003) provided a search algorithm to approach the best model by minimizing a penalized risk; however, for large data sets the computational constraints can be prohibitive. Traditional approaches, such as using model marginal likelihoods for model comparison, seem to be more practicable considering the large data sets that are commonly encountered in clustering problems.

The marginal likelihoods conditional on the specific clustering configurations follow directly from the calculations in Section 4.2. For a fixed cluster configuration $\mathcal{C}$, simple Monte Carlo averaging of marginal distributions (10) gives

$$f(\mathbf{Y}|\mathcal{C}) \approx \frac{1}{N} \sum_{k=1}^{N} \prod_i f(\{\mathbf{Y}_j\}_{j\in\mathcal{C}(i)} | \gamma^{(k)}, \mathbf{g}^{(k)}, \mathcal{C}), \tag{13}$$

where N is the total number of MCMC samples. Here a large state space of the indicators $\gamma \in \{0,1\}^m$ would ideally require a very large number of MCMC samples. In general, it is observed that the Markov chain of the indicator variables mixes well like most hierarchical wavelet models, with little change in the convergent states across simulations. This is essentially due to the sharp localization of features on the wavelet scale. In such cases, a reasonable estimate of the marginal likelihood (13) is achieved by averaging over the Markov chain for a previously estimated γ.

6.4. Missing data interpolation

It is common to encounter missing values in clustering and to supplement the model with a Bayesian imputation step is useful. There has been some work for wavelet methods on unequispaced grids (Kovac and Silverman, 2000; Pensky and Vidakovic, 2001); however, we limit ourselves to the case where points are missing from a fixed equispaced grid.

Missing data imputation is expedited by Gibbs sampling under the *a priori* independence of the d_n distinct parameters. Gibbs sampling starts with a random-clustering configuration and randomly imputed missing points. Conditionally on the imputed data set, the Pölya urn samples the cluster parameters $\boldsymbol{\theta}_i$. After n such steps of sequential sampling the initial clustering configuration $\mathcal{C}_n$ is completely updated. Next, conditionally on $\mathcal{C}_n$ the posterior predictive distribution is used to perform the imputation. For instance, for an incomplete response (say the jth response) that was assigned to the ith cluster in the preceding step of the Gibbs sampling, we use

$$\mathbf{Y}_j|\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2 \sim N(\mathbf{X}\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2\mathbf{I}_m)$$

and

$$\bar{\boldsymbol{\beta}}_i \overset{\text{IID}}{\sim} N(0, \bar{\sigma}_i^2\mathbf{V}),$$

to write the marginal

$$\mathbf{Y}_j|\bar{\sigma}_i^2, \mathbf{g}, \gamma \sim N\{\mathbf{0}, \bar{\sigma}_i^2(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}')\}.$$

Let $\mathbf{Y}_{j1}$ and $\mathbf{Y}_{j2}$ be the known and missing parts of $\mathbf{Y}_j$ respectively; then from standard normal distribution theory the posterior predictive distribution is given by

$$(\mathbf{Y}_{j2}|\mathbf{Y}_{j1}, \bar{\sigma}_i^2, \mathbf{g}, \gamma, \mathcal{C}_n) \sim N\{\Lambda_{21}\Lambda_{11}^{-1}\mathbf{Y}_{j1}, \bar{\sigma}_i^2(\Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12})\},$$

where Λ_{ij} are obtained by partitioning the marginal covariance as

$$\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}' = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

We can use this technique to predict unobserved portions of any curve with uncertainty intervals accurately. The effectiveness of our method is shown in one of the examples.

6.5. *Shifted Doppler signals*

The first data set is motivated by similar examples in Donoho and Johnstone (1994, 1995) and consists of 200 shifted Doppler signals with the common form

$$f_{t_0}(t) = -0.025 + 0.6\surd\{t(1-t)\}\ \sin\{2.10\pi/(t-t_0)\}$$

and the phase t_0 is continuously varied in eight disjoint intervals equally interspersed in [0, 1] to generate the 200 signals. In one of these intervals, three functions were perturbed to assess the model sensitivity to small local fluctuations. This created a total of nine distinct classes of functions that have been equisampled on a common grid of 128 points. For simulation, noisy data were generated by adding independent normal noise $\varepsilon_{i,k} \sim N(0, \sigma^2)$ to each of the 200 Doppler signals at 128 points. Later, with a fixed probability p_m, points were randomly selected and dropped from each function to evaluate the robustness to missing observations.

Table 1 shows the estimated number of clusters $\hat{d}_n$ and the percentage of misclassifications for various amounts of randomly missing data across $\sigma = 0.1, 0.06, 0.02$ when the data are fitted with model 2. These figures were averaged over the best models from 100 simulations for each combination of σ and missing data probability p_m.

Fig. 1 shows the model log-marginal-likelihoods for various values of p_m when $\sigma = 0.06$. The most favoured models on the basis of log-marginal-likelihoods had nine clusters followed by models with 7–11 clusters. First, the case $p_m = 0.0$ with no missing observations is discussed. Fig. 2 shows the nine clusters estimated in one of the simulations at $\sigma = 0.1$. In most of the

Table 1. Performance of the wavelet model at different signal-to-noise ratios and percentage of missing observations†

σ	*$\hat{d}_n$ for the following values of p_m:*				*Misclassifications (%) for the following values of p_m:*			
	0.0	*0.1*	*0.2*	*0.3*	*0.0*	*0.1*	*0.2*	*0.3*
0.1	8.4	7.9	6.5	5.5	10.5	14.5	20.0	26.1
0.06	9.1	8.9	8.4	7.5	8.3	9.8	12.2	17.4
0.02	9.1	9.1	8.7	8.1	3.1	5.2	7.5	10.5

†p_m is the proportion of missing points in each curve. σ is the noise standard deviation. $\hat{d}_n$ is the estimated number of clusters. The actual number of clusters is 9.

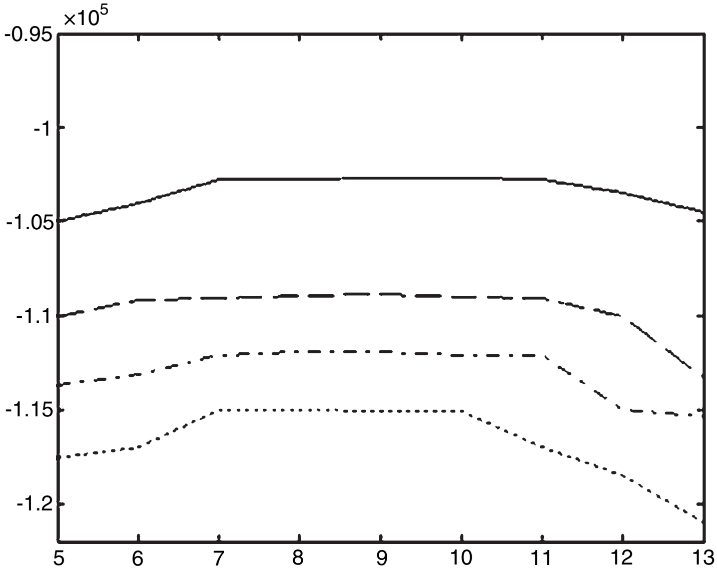

Fig. 1. Model log-marginal-likelihood *versus* number of clusters for the shifted Doppler signals data (σ = 0.06): ———, p_m = 0.0; — —, p_m = 0.1; · – · – ·, p_m = 0.2; -------, p_m = 0.3

situations the wavelet model performs better than the spline model (Table 2) and the estimated number of clusters $\hat{d}_n$ is consistently close to 9. The wavelet model also has lower misclassification rates than the spline model, indicating its robustness in high noise situations.

The Besov parameters (a, b) are obtained by running the maximization procedure that was outlined in Section 3.1 on a 2×2 grid of (a, b) pairs and we obtain the estimates as $a = 1.90$ and $b = 1.20$. We present the log(Bayes factors) to compare the model at $(a, b) = (1.90, 1.20)$ with some other neighbouring values of (a, b) in Table 3.

6.5.1. *Results for the missing data case*

To evaluate the effects of missing data, points from each function were randomly selected and dropped with probabilities p_m. Tables 1 and 2 summarize the results from three separate simulations performed with $p_m = 0.1$, 0.2, 0.3. At $p_m = 0.1$, our method performs similarly to the

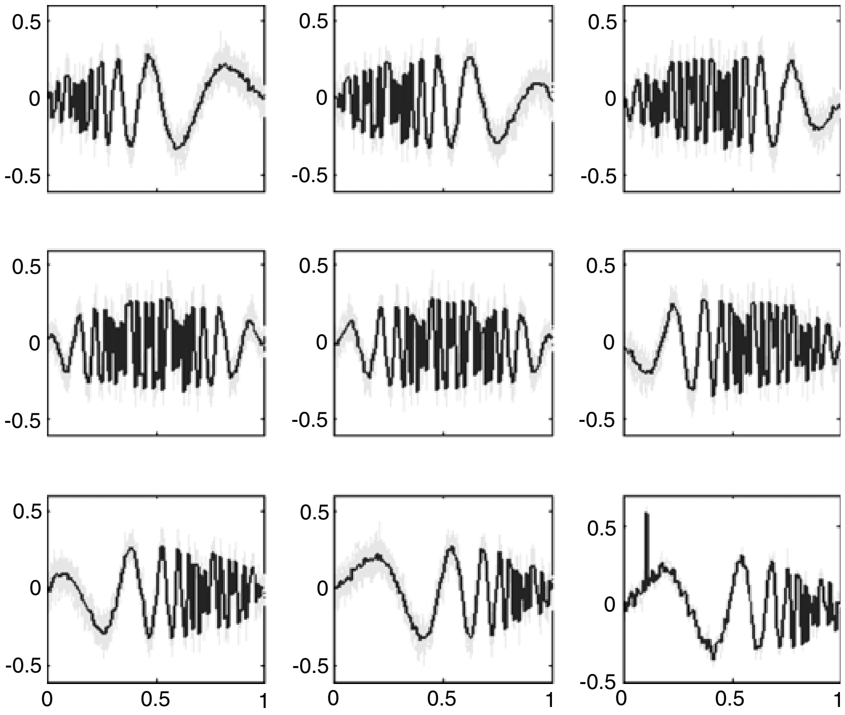

Fig. 2. Nine estimated clusters for the shifted Doppler signals at $\sigma = 0.1$

Table 2. Performance of the mixed effects spline model at different signal-to-noise ratios and percentage of missing observations†

σ	*$\hat{d}_n$ for the following values of p_m:*				*Misclassifications (%) for the following values of p_m:*			
	0.0	*0.1*	*0.2*	*0.3*	*0.0*	*0.1*	*0.2*	*0.3*
0.1	7.3	6.8	5.2	4.8	18.4	22.5	24.9	36.3
0.06	9.5	8.9	7.5	6.3	11.5	16.1	20.4	27.4
0.02	9.1	8.6	7.6	6.9	4.0	7.4	11.0	17.0

†p_m is the probability that a point is missing in each curve; 20 quantile knots. σ is the noise standard deviation. $\hat{d}_n$ is the estimated number of clusters. The actual number of clusters is 9.

complete-data case, with a small deviation in the estimated size and a marginal increase in the misclassifications even at $\sigma = 0.1$. At $p_m = 0.2$ and $p_m = 0.3$, the number of misclassifications increases and there is a drop in $\hat{d}_n$. Notably, the deterioration in the wavelet model (Table 1) with higher amounts of missing data is less drastic than in the spline model (Table 2).

6.5.2. *Predictive inference with missing data*

A major advantage of the functional clustering procedure is that it can accurately predict unobserved portions of a curve. We demonstrate another contrived example where a portion of

Table 3. Estimated 2 log($M_{1.90,1.20}/M_{a,b}$) for the shifted Doppler signal data†

a	*Results for the following values of b:*				
	1.00	*1.10*	***1.20***	*1.30*	*1.40*
1.75	16.78	13.24	9.53	12.63	15.95
1.85	14.07	12.64	3.79	5.98	14.47
1.90	14.52	10.80	—	6.44	13.74
1.95	19.45	17.98	15.84	12.17	15.61

†$M_{a,b}$ is the marginal likelihood of the model with hyperparameters (a, b).

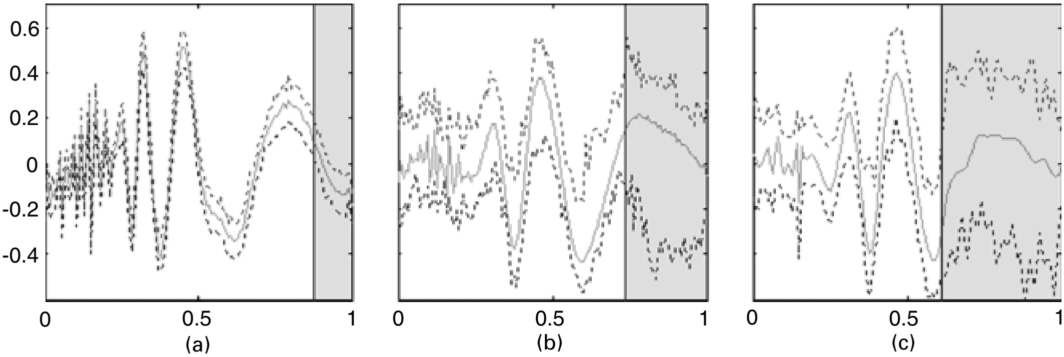

Fig. 3. MCMC prediction bands for a partially observed curve in the shifted Doppler signals example (noise standard deviation $\sigma = 0.1$): (a) 12.5% missing points; (b) 25% missing points; (c) 37.5% missing points

a curve is missing; we examine missing data prediction and its effects on clustering. A fixed portion of the tail from a curve in the shifted Doppler example is dropped and prediction bands are generated from the MCMC samples. We expect the clustering algorithm to be reasonably stable when a few curves are partially observed owing to the shrinkage within clusters and this is confirmed in the prediction bands that are plotted in Fig. 3. With 12.5% missing points the effect of missing data prediction is hardly visible and, although the bands widen with an increasing number of missing points, it is only when this number reaches 37.5% that the curve is occasionally thrown into a distinct cluster.

6.5.3. *Effect of scaling coefficients*

Three types of shifted Doppler data sets are considered (Fig. 4) for the analysis with the first two types (Figs 4(a) and 4(b)) having curves that differ either in their scaling coefficients or detail coefficients. The data set for Fig. 4(c) has differences in both the scaling and the detail coefficients. Each curve is replicated five times in normal noise ($\sigma = 1$) so that the sample size n is 15 for each data set.

For diffuse specifications in all the three cases, we set $r_0 = 2.2$ and $s_0 = 4.2$ with a prior mean $E(g_0) = 1$ and variance $\text{var}(g_0) = 10$. The empirical estimates of (r_0, s_0) for the three data sets calculated in the aforementioned manner are (7.991, 6.695), (5.7961, 11.315) and (5.110, 6.5823).

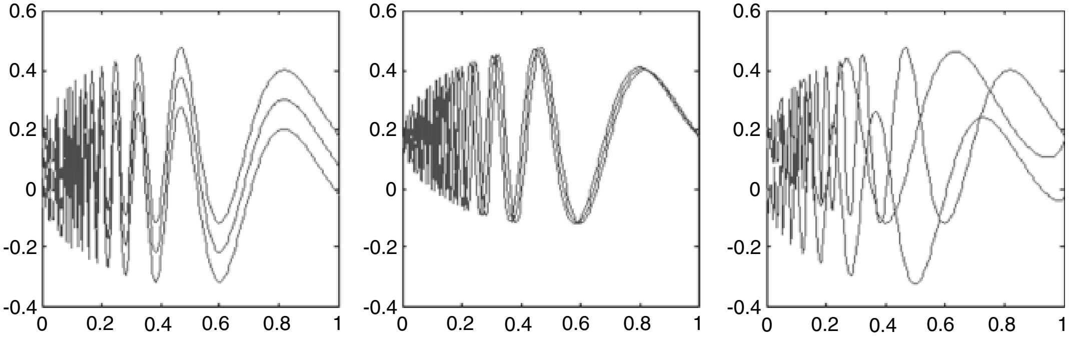

Fig. 4. Three different Doppler signal data sets with three classes used to assess the model sensitivity to scaling coefficients

Table 4. Comparison of prior choices for the scaling coefficients for various Doppler signals†

Set	*Misclassifications (%) for the following priors:*	
	Estimated	*Diffuse*
(a)	7.0	12.2
(b)	13.5	14.1
(c)	4.6	10.5

†The actual number of clusters is 3.

The corresponding priors means of g_0 are 1.7019, 0.6222 and 1.1153 and the prior variances are 2.1491, 0.10584 and 0.96344.

Table 4 summarizes the performance of these priors applied to the three data sets. In data set (a) only the scaling coefficients play a role in the clustering. The tighter empirically estimated prior produces better estimates of d_n with lower misclassification rates than the diffuse prior. The differences between the curves in data set (b) are almost entirely encoded in the detail coefficients and the prior modelling of scaling coefficients does not play a role in the clustering, as shown in Table 4. In data set (c), both the scaling and the detail coefficients differ with clear evidence of three clusters in the latter. Although the role of scaling coefficients is diminished, the empirically estimated prior still manages to outperform the diffuse prior.

6.6. *Yeast cell cycle data*

Recently there has been huge interest in the analysis of gene expression data from deoxyribonucleic acid (DNA) microarray experiments. When microarray experiments are performed consecutively in time, we call this experimental setting a time course of gene expression profiles. Clustering of the time course data gives insight about genes that behave similarly over the course of the experiment. By comparing genes of unknown function with profiles that are similar to genes of known function, clues to function may be obtained. Hence, the coexpression of genes is of interest.

We analysed a similar data set to that of Spellman *et al.* (1998) which measures the relative levels of messenger ribonucleic acid over time from 6178 genes in α-pheromone synchronized yeast cell cultures. Of interest are the connected genetic regulatory loops controlling the *Saccharomyces cerevisiae* pheromone response pathway (PRP) and whether the genes involved can be identified by their characteristic expression profiles in one or more clusters. In the sexual reproduction of yeast there is an essential role of pheromone response and mating pathways that ultimately target the protein STE12 and bind DNA as a transcriptional activator for a number of other genes. This is a natural choice for our methodology because the intergenic regions in the yeast genome that are bound to STE12 are known from genomic location analysis in the presence of pheromone (Ren *et al.*, 2000). With the induction of the α-pheromone, the expression levels for the PRP genes show a steep rise, until an internal stabilizing mechanism is triggered and the fervour dies down. The result is a spiky event (or, more contextually, a temporal singularity) in an otherwise smooth expression profile which, for the most part, is comparable with the response of genes that do not participate in the pheromone response signalling and yeast mating.

The wavelet transforms give a time–frequency breakdown of information and the coefficients may reveal new patterns in frequency as well as time. For example, Klevecz (2000) performed a detailed wavelet analysis of the yeast cell cycle data and found significant high frequency artefacts that were isolated in time, contrary to the earlier notion that yeast cell cycle profiles are representative of slowly varying biological events. The hierarchical clustering model can easily delineate profiles with increased levels of gene expression occurring uniformly throughout the cycle from profiles that are characterized by sporadic bursts of spiky events (when relatively more messages are synthesized). The latter, being a characteristic of the PRP genes, is more important here. Note that this extremal behaviour is easily accommodated within the Besov spaces.

In the experiments, 16 (of the 18) equisampled measurements over two cell cycles (lasting roughly 140 min) from 600 significantly expressed genes were considered. Some expression profiles were incomplete with a maximum of eight missing expressions per gene. To allow for possible deflections in the error variance in the population of gene expressions, we resorted to a heteroscedastic model (model 1) and found that the estimated variance varies between clusters. This may be attributed to the significant deviations in the sizes of cluster and the relatively short-sized expression profiles.

In general, the normality assumptions do not hold for microarray experiments and some preprocessing steps are necessary. For example, for the yeast data a log-transformation seems to suffice. This is confirmed by a Bayesian analysis (Chaloner and Brant, 1988) of the residuals. The residuals in the normal likelihood (3) are sampled from their posterior distribution conditionally on the clustering configuration. From standard distribution theory, this posterior distribution is normal with mean $\mathbf{Y}_i - \mathbf{X}\boldsymbol{\mu}_i^*$ and covariance $\sigma^2\mathbf{V}^*$ ($\boldsymbol{\mu}_i^*$ and $\mathbf{V}^*$ are defined in expression (9)). A multivariate χ^2-test is then performed to check the normality of the sampled residuals for each curve. The p-values (at a level $\alpha = 0.05$) from each curve are provided in Fig. 5 and show that most of the vector responses satisfy the normality assumptions.

The clustering algorithm showed maximum preference for models with 6–9 clusters (Fig. 6), of which two models with eight clusters dominated the others in terms of the model log-marginal-likelihoods (listed in Table 5). Using a grid maximization procedure, the Besov parameters (a, b) were set to (1.45, 0.5) and this explains the spatial inhomogeneity in the expression profiles that is noticeable in the eight clusters (for one of the best models) that are plotted in Fig. 7. The plots can be divided between periodic (Figs 7(a)–7(d)) and non-periodic (Figs 7(e)–7(h)) patterns. The two clusters in the second category can be identified with the early on–off switch patterns and pertain to almost all the PRP genes that were mentioned above.

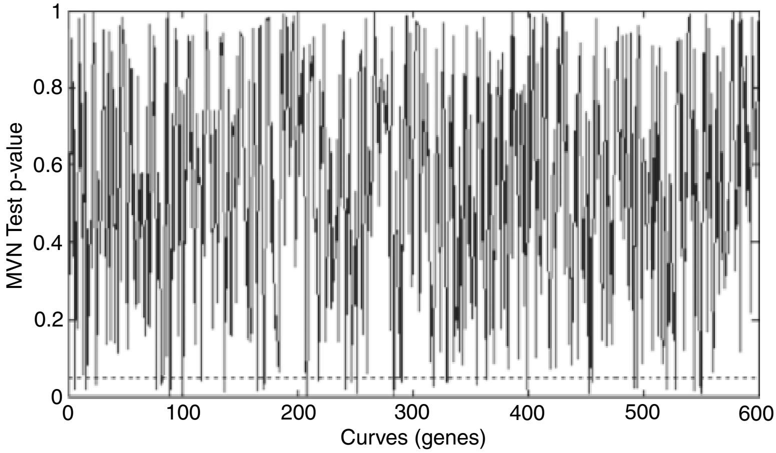

Fig. 5. *p*-values *versus* curves from the multivariate test of normality for the yeast data

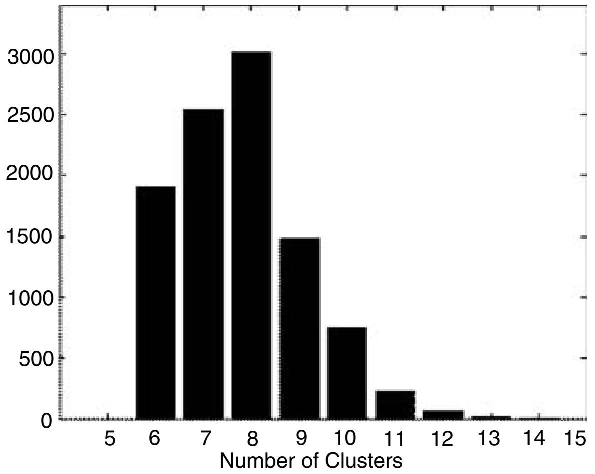

Fig. 6. Histogram showing the preferred number of clusters for the yeast cell cycle data over 10 000 MCMC iterations

6.6.1. Comparison with the spline model

The James and Sugar (2003) model is fitted with four quantile knots. The most preferred models vary in size from 5 to 7 and suggest oversmoothing and the models' incapacity to adjust to sharp fluctuations. The log(Bayes factor) of the best wavelet models compared with the best spline model was much larger than 10. In the results, which are not detailed here, there is a tendency for clusters (b)–(e) and (f)–(h) in Fig. 7 to merge into one cluster. To emphasize this point, we fit three models differing in spatial adaptation to three (of the eight) clusters that were obtained from the wavelet model. The first cluster has periodic and smooth profiles, the second cluster is smooth but not periodic and, finally, the third cluster is totally irregular and comes with a sharp on–off pattern of the PRP genes. Figs 8(a)–8(c), 8(d)–8(f) and 8(g)–8(i) plot the fits by a periodic (Fourier cosine series) basis, a spline basis and a wavelet basis divided in three columns corresponding to the three clusters. In the first column, we see that all three fit equally well, whereas, in the second column, the periodic basis fit (Fig. 8(b)) shows considerable bias at the

Table 5. Log-marginal-likelihoods of the best models for the yeast cell cycle data

d_n	*Best models*	*Log-marginal-likelihood*
8	1,2	-8.697×10^4
	3,4,5	-8.699×10^4
7	1	-8.700×10^4
	2,3	-8.701×10^4
6	1,2	-8.701×10^4
9	1	-8.703×10^4
	2	-8.704×10^4

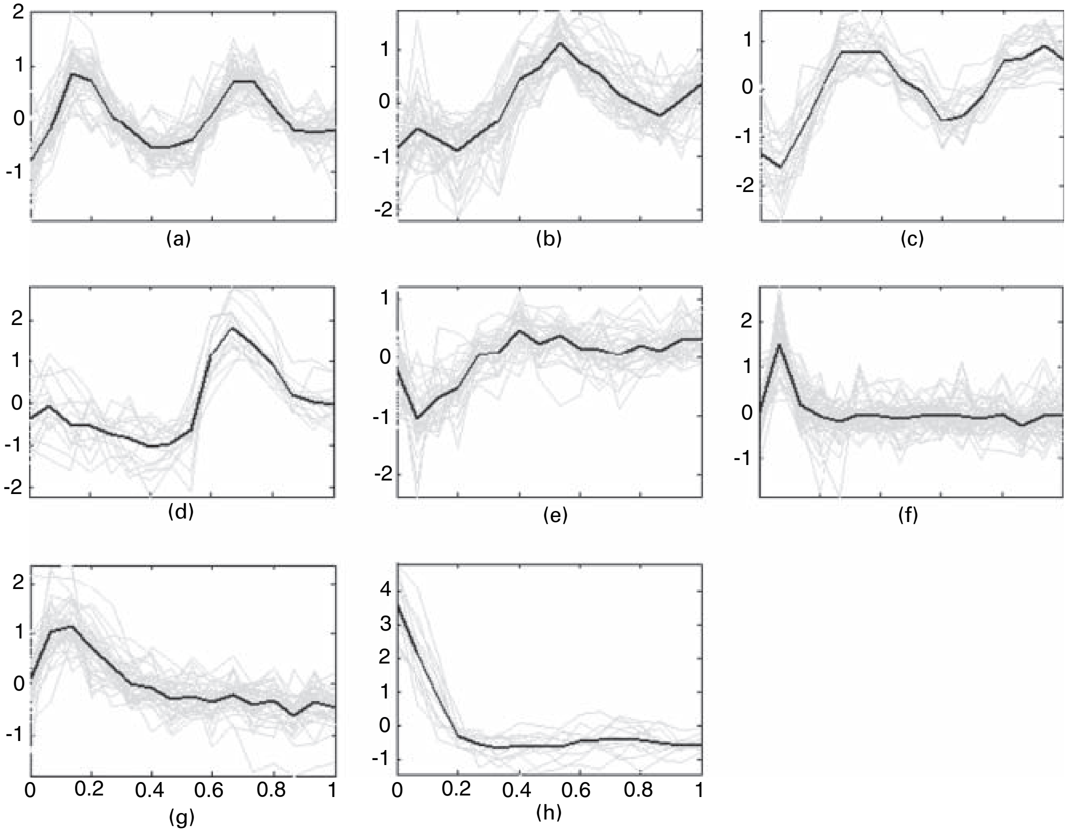

Fig. 7. Clustering of the α-synchronized yeast cell data (eight clusters) for the 600 expression profiles with 18 time points and a maximum of eight missing points: clusters (a)–(d) hold the periodic expression profiles and (e)–(h) hold the non-periodic on–off switching patterns

boundaries. The situation deteriorates further as we move to the third column with a sharp fluctuation at $t=0.1$ which is completely missed by both the periodic and the spline models.

6.6.2. Further simulation study

The yeast cell cycle data is a typical complementary DNA microarray example where high noise levels make inference difficult. Moreover, there is a marked heteroscedasticity in the data as

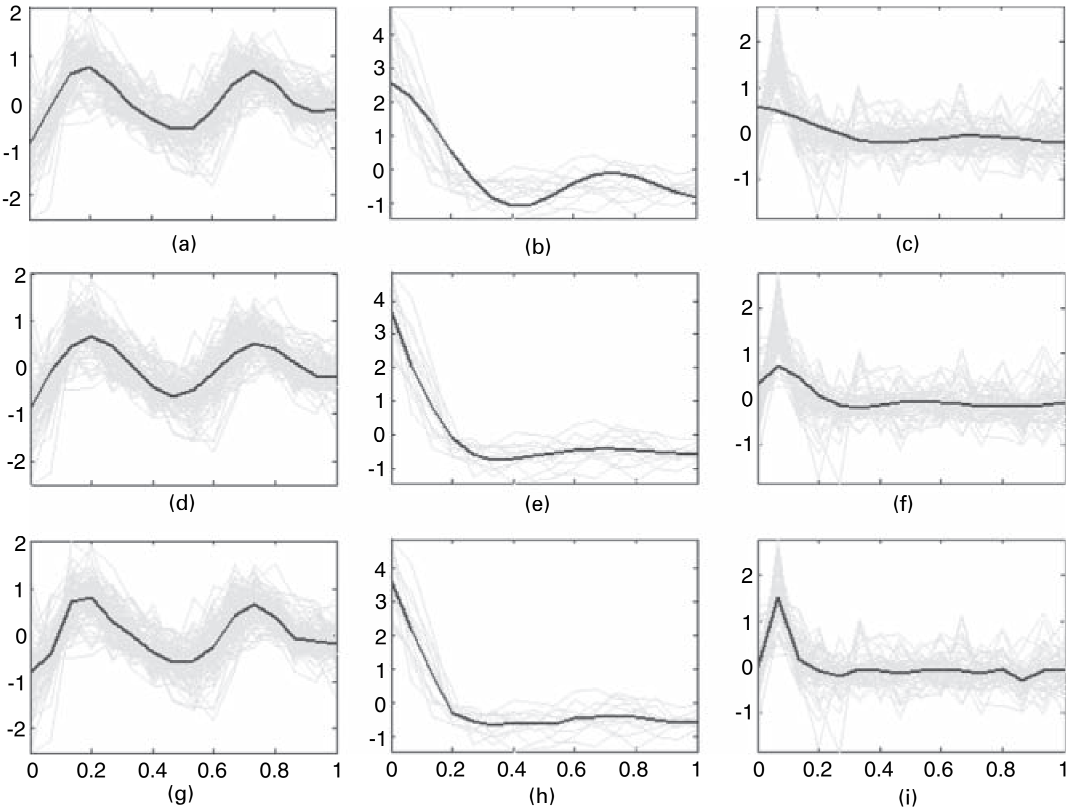

Fig. 8. Comparison of fits for three types of response: (a)–(c) periodic basis fit; (d)–(f) spline fit; (g)–(i) wavelet fit; (a), (d), (g) periodic smooth function (all the bases fit well); (b), (e), (h) non-periodic but smooth function (the periodic basis has problems fitting at the boundary); (c), (f), (i) non-periodic and irregular function (only the wavelet basis captures the sharp fluctuation)

indicated in Table 6, tabulating the estimated variances of the eight clusters of Fig. 7. A follow-up simulation procedure is useful in such situations to validate the results. To simulate a realistic data set for comparing the successful wavelet and spline models, we used the yeast cell cycle data as a prototype. As realistic values of the parameters, we use the representative curves (reproduced from the estimated wavelet coefficients) of the eight clusters that are plotted with thick lines in Fig. 7 and replicate them in IID normal noise with the estimated variances of Table 6 following the structure of our model to generate the responses. In other words, this is an imitation of the original 600 gene expression profiles, to which we want to apply the algorithm and to confirm our findings. This simulation is repeated 100 times to generate 100 different data sets, which are later analysed by using wavelet as well as spline models to obtain the average misclassification rates.

The estimated number of clusters averaged over 100 simulations for the wavelet model is 8.21 with a very low average ‘misclassification rate’ (the deviation from the previously estimated clustering configuration) of 5.38%. In fact, the estimated clusters in these simulations almost always resemble Fig. 7 with differences due to occasional switch-over of curves between clusters (f) and (g), or the formation of new clusters out of (or from combination of) clusters (f), (g) and (h). For the spline model, we see many clusters merging because of oversmoothing and the average number of clusters is 5.96 with a misclassification rate of 36.42%.

Table 6. Size and estimated variance of the eight clusters for the yeast cell cycle data

Cluster	*Number of curves*	*Estimated variance*
(a)	124	0.3519
(b)	163	0.2996
(c)	21	0.8154
(d)	33	0.8862
(e)	59	0.2893
(f)	98	0.3006
(g)	24	0.2694
(h)	78	0.5360

6.7. Precipitation spatial time series data

We consider a National Centers for Environmental Prediction (USA) reanalysis of spatiotemporal data that record the daily precipitation over Oregon and Washington between 1949 and 1994 (Widmann and Bretherton, 2000). The gridded observations represent area-averaged precipitation on a 50 km × 50 km grid. Bi-weekly averages of the daily observations from 179 locations are used, with a total of 512 time points over a time span of roughly 5 years between 1989 and 1994.

The example illustrates the potential application of functional clustering for the topographical categorization of meteorological factors such as precipitation, temperature and snowfall. It is usually difficult to generate topographical contour maps of precipitation *versus* elevation although these are of great interest in climate analysis. A functional clustering model provides a natural way to group similar precipitation patterns viewed as functions and to associate them with elevation. This can also deal with the problem of missing points in precipitation analysis. Missing points are typically interpolated with information from satellite observations and analysis amidst the differences in the measurement errors from two sources can be problematic.

Precipitation maps are formed by using the slope of a simple regression of the average local precipitation and the elevation. The clustered data that are plotted in Fig. 9 show a clear need

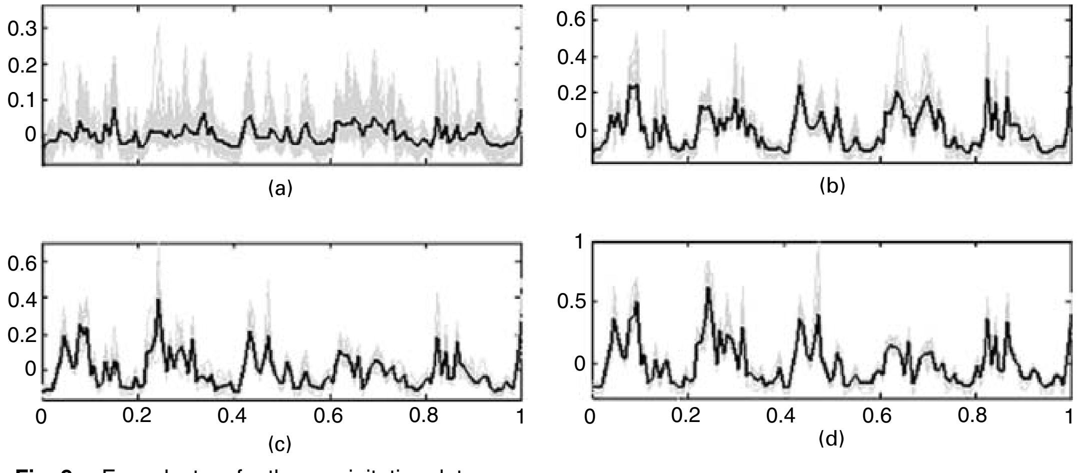

Fig. 9. Four clusters for the precipitation data

for non-linear modelling. This could in fact be used to delineate regions with occasional swings in rainfall patterns—patterns that can be overlooked by other geostatistical methods such as kriging (Rivoirard, 1994).

In many spatial models, a spatial random-effects term viewed as a random-intercept process is introduced to capture the spatial correlation. Previously, the intercept or scaling coefficients β_{i00} in our model were specified by nonparametric priors for clustering. Following Gelfand *et al.* (2003), we introduce a spatial random effect $\alpha(\mathbf{x}_i)$ that can be interpreted as a random spatial adjustment at a location $\mathbf{x}_i$ (in latitude and longitude) to the overall intercept β_{i00}. Thus, for an observed set of locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we write

$$\mathbf{Y}_i = \alpha(\mathbf{x}_i)\mathbf{1} + \mathbf{X}\boldsymbol{\beta}_i + \varepsilon_i,$$

where the overall intercept β_{i00} is an element of $\boldsymbol{\beta}_i$. We assume that the prior distribution for $\boldsymbol{\alpha}$ is a zero-mean Gaussian process with exponential correlation function $\tau^2\boldsymbol{\Phi}_\rho$. Here $\boldsymbol{\Phi}_\rho$ has a special structure in that its (i, j)th entry is $\exp(-\rho\|\mathbf{x}_i - \mathbf{x}_j\|)$ where $\rho > 0$ is a spatial decay parameter. Following Banerjee (2004), we assume a gamma prior for ρ so that the mean of the spatial prior range is half the maximum intersite distance in the data set, and τ^2 is a scaling parameter specified by a vague inverse gamma prior distribution $\mathrm{IG}(0.005, 0.005)$. The dependence between the $\alpha(\mathbf{x}_i)$ makes them identifiable from the other intercept terms without the need for replications.

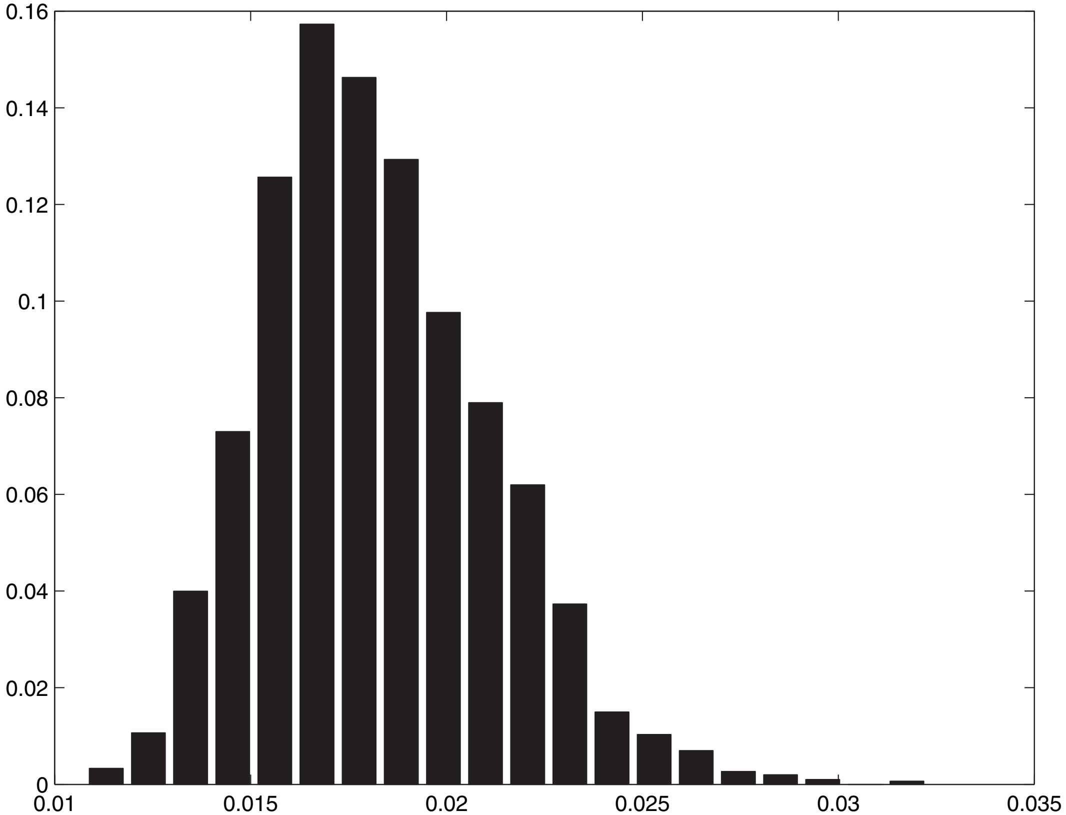

Fig. 10. Posterior distribution of ρ generated from 10000 MCMC iterations

For posterior inference, Gibbs sampling is used to sample $\boldsymbol{\alpha}$ and the clustering parameters alternately. More specifically, for a fixed $\boldsymbol{\alpha}$ let $\mathbf{Y}_i^* = \mathbf{Y}_i - \alpha(\mathbf{x}_i)\mathbf{1}$; then the posterior inference in Section 4 can be performed conditionally on $\mathbf{Y}_i^*$. In model 1, each α_i is updated separately by combining the implied conditional prior $\alpha_i|\alpha_{-i}, \tau^2, \rho$ (from the prior $\boldsymbol{\alpha}|\tau^2, \rho \sim N(0, \tau^2 \boldsymbol{\Phi}_\rho)$) with the likelihood $\mathbf{Y}_i \sim N(\alpha_i \mathbf{1} + \mathbf{X}\boldsymbol{\beta}_i, \sigma_i^2 \mathbf{I})$. For the homoscedastic case, we can directly work with the joint prior and the joint likelihood to draw multivariate samples of $\boldsymbol{\alpha}$. Finally, the parameters ρ and τ^2 are updated by separate Metropolis–Hastings steps by conditioning only on $\boldsymbol{\alpha}$.

The NCEP reanalysis data set was fitted with model 1 and as expected there were considerable differences in the estimated variance between clusters. The estimated value of ρ is 0.0182 and its posterior distribution from MCMC sampling is shown in Fig. 10. This suggests a high spatial correlation between different locations. The overall homogeneity that is associated with the annual events and the local bumpiness due to fluctuations in rainfall are described well by the estimated Besov parameters $(a, b) = (0.95, 0.3)$. The histogram of the MCMC samples for the number of clusters from one simulation in Fig. 11 shows clear preference for models with

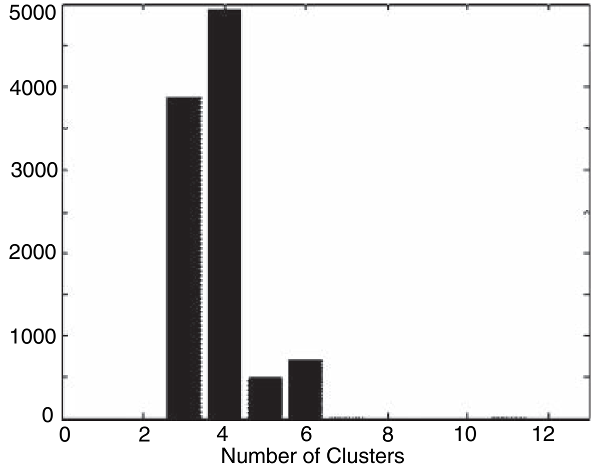

Fig. 11. Histogram showing the preferred number of clusters for the precipitation data over 10 000 MCMC iterations

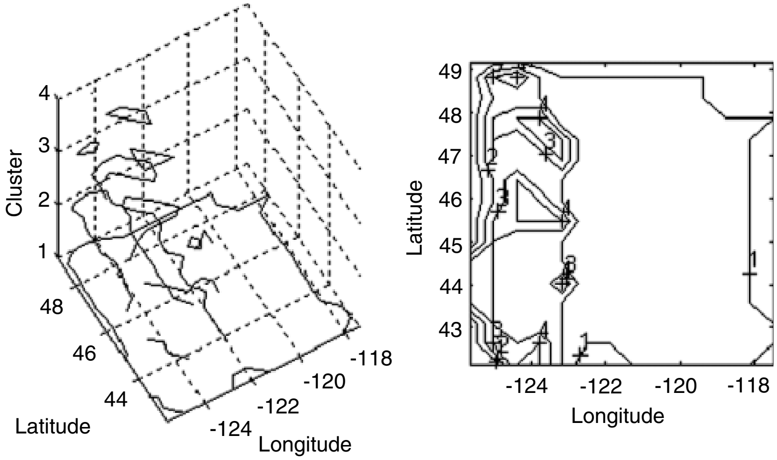

Fig. 12. Topographical distribution of the four clusters shown in two different orientations

four clusters. The estimated clusters from one simulation are shown in Fig. 9 and their distribution on a geographical scale is contour plotted in Fig. 12 in two different orientations. Fig. 9(a) plots the largest cluster and corresponds to a large number of stations outlined by cluster 1 in Figs 9(a) and 9(b). The average annual rainfall in these areas has shown moderate fluctuations over the 5-year period and is notably less than the areas in clusters 2–4. Stations in cluster 3 (Fig. 9(c)) have experienced heavier-than-usual rainfall between 1990 and 1991, but otherwise the average rainfall is comparable with cluster 2 (Fig. 9(b)). Stations in cluster 4 (Fig. 9(d)), although much wetter, share the same pattern as cluster 3, suggesting their geographical proximity, which is confirmed from Fig. 12.

7. Discussion

The nonparametric Bayes model offers a flexible approach to functional clustering and has been shown to perform favourably against other functional clustering methods. Special stress is laid on the overall applicability of the methodology in that we rely on straightforward Gibbs sampling methods that are usable with high dimensional data, employ simple base prior modelling of the wavelet coefficients to encompass a large class of functions and address the missing data problem that is common in real life applications. In addition, the method learns about the number of clusters in an automated manner, unlike other clustering methods where a dimension change comes with a huge computational burden.

In its ability to partition the predictor space into regions of ‘IID’ data, the DP is comparable with product mixture models for clustering. (Indeed, Quintana and Iglesias (2003) showed an equivalence under certain regularity conditions on the DP.) This entails the use of two distinct approaches to Gibbs sampling in this paper. First, the sampling of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ from the Pölya urn allows the update and clustering of these parameters in a unified way and replaces the reversible jump sampler (Green, 1995) that is used in product models for clustering and has a reputation for being complicated. However, conditionally on a sampled configuration of clusters, the remaining parameters are conveniently drawn from the product mixture that is provided by the DP.

The discrete wavelet transform (1) requires that the number of sampled points m be an integer power of 2. The model proposed can be used with more flexible alternatives, such as the lifting scheme (Sweldens, 1996), that do not place restrictions on the discrete support. This could additionally allow the extension of this model to unequispaced data. Another interesting research problem would be to modify Quintana and Iglesias’s (2003) method in this high dimensional functional clustering problem with the presence of missing data.

Acknowledgements

This research was supported by National Cancer Institute grants CA-10462 and CA-57030. The authors gratefully acknowledge the Joint Editor, the Associate Editor and two referees for their constructive suggestions that led to significant improvements in the presentation of the paper.

Appendix A

A.1. Marginal calculations for the heteroscedastic model

Recall that $\mathbf{Y}$ $(n \times m)$ was used as the collection of n functional responses of length m. For convenience, let $\mathbf{Y}_{\mathcal{C}_n(i)}$ $(n_i \times m)$ denote all the responses falling in the ith cluster $\mathcal{C}_n(i)$. We can write the likelihood as

$$f(\mathbf{Y}|\{\bar{\beta},\bar{\sigma}\}_{i=1}^{d_n},\mathcal{C}_n)=\prod_{i=1}^{d_n} f(\mathbf{Y}_{\mathcal{C}_n(i)}|\bar{\boldsymbol{\beta}}_i,\bar{\sigma}_i^2,\mathcal{C}_n)$$

and the marginal as

$$f(\mathbf{Y}|\gamma,g,\mathcal{C}_n)=\prod_{i=1}^{d_n}\int f(\mathbf{Y}_{\mathcal{C}_n(i)}|\bar{\boldsymbol{\beta}}_i,\bar{\sigma}_i^2,\mathcal{C}_n)\,f(\bar{\boldsymbol{\beta}}_i,\bar{\sigma}_i^2)\,\mathrm{d}\bar{\boldsymbol{\beta}}_i\,\mathrm{d}\bar{\sigma}_i^2.$$

Suppose that the ith cluster has n_i responses and $(\bar{\boldsymbol{\beta}}_i,\bar{\sigma}_i^2)\sim\mathrm{NIG}(0,\mathbf{V};u,v)$ *a priori*; the ith (of the d_n) integral inside the product can be written as

$$\begin{aligned}
&\frac{(u/2)^{v/2}}{|\mathbf{V}|^{1/2}(2\pi)^{(n_i m+p)/2}\,\Gamma(v/2)}\int\frac{1}{\bar{\sigma}_i^{(n_i m+v+p+2)/2}}\exp\left\{-\frac{1}{2\bar{\sigma}_i^2}\sum_{i'\in\mathcal{C}(i)}(\mathbf{Y}_{i'}-\mathbf{X}\bar{\boldsymbol{\beta}}_i)^{\mathrm{T}}(\mathbf{Y}_{i'}-\mathbf{X}\bar{\boldsymbol{\beta}}_i)\right\}\\
&\quad\times\exp\left\{-\frac{\mathrm{tr}(\bar{\boldsymbol{\beta}}_i^{\mathrm{T}}\mathbf{V}^{-1}\bar{\boldsymbol{\beta}}_i)+u}{2\bar{\sigma}_i^2}\right\}\mathrm{d}\bar{\boldsymbol{\beta}}_i\,\mathrm{d}\bar{\sigma}_i^2=\frac{|\mathbf{V}_i^*|^{1/2}(u/2)^{v/2}}{|\mathbf{V}|^{1/2}(2\pi)^{n_i m/2}\,\Gamma(v/2)}\int\frac{\exp(-u^*/2\bar{\sigma}_i^2)}{\bar{\sigma}_i^{(n_i m+v+2)/2}}\\
&\qquad\times\int\frac{1}{(2\pi\bar{\sigma}_i)^{p/2}|\mathbf{V}_i^*|^{1/2}}\exp\{(\bar{\boldsymbol{\beta}}_i-\boldsymbol{\mu}_i^*)^{\mathrm{T}}\mathbf{V}_i^{*-1}(\bar{\boldsymbol{\beta}}_i-\boldsymbol{\mu}_i^*)\}\,\mathrm{d}\bar{\boldsymbol{\beta}}_i\,\mathrm{d}\bar{\sigma}_i^2\\
&\qquad=\frac{|\mathbf{V}_i^*|^{1/2}(u/2)^{v/2}}{|\mathbf{V}|^{1/2}(2\pi)^{n_i m/2}\,\Gamma(v/2)}\int\frac{\exp(-u^*/2\bar{\sigma}_i^2)}{\bar{\sigma}_i^{(n_i m+v+2)/2}}\mathrm{d}\sigma_i^2\\
&\qquad\propto\frac{\Gamma\{(v+mn_i)/2\}}{\pi^{n_i m/2}}\frac{|\mathbf{V}_i^*|^{1/2}}{|\mathbf{V}|^{1/2}}(u_i^*)^{-(v+mn_i)/2}
\end{aligned}$$

where

$$\mathbf{V}_i^*=(n_i I+\mathbf{V}^{-1})^{-1},$$

$$\boldsymbol{\mu}_i^*=\mathbf{V}_i^*\sum_{i'\in\mathcal{C}_n(i)}\mathbf{X}^{\mathrm{T}}\mathbf{Y}_{i'},$$

$$u_i^*=u+\sum_{i'\in\mathcal{C}_n(i)}\mathbf{Y}_{i'}^{\mathrm{T}}\mathbf{Y}_{i'}-\boldsymbol{\mu}_i^{*\mathrm{T}}\mathbf{V}_i^{*-1}\boldsymbol{\mu}_i^*.$$

Multiplying over all d_n clusters, we obtain

$$f(\mathbf{Y}|\gamma,\mathbf{g},\mathcal{C}_n)\propto\prod_{i=1}^{d_n}\frac{\Gamma\{(v+mn_i)/2\}}{\pi^{n_i m/2}}\frac{|\mathbf{V}_i^*|^{1/2}}{|\mathbf{V}|^{1/2}}(u_i^*)^{-(v+mn_i)/2}.$$

A.2. Proof for Besov priors

This is an extension of the proof for theorem 2 in Abramovich *et al.* (1998) for the special case of finite Besov scales $p,q<\infty$. This condition ensures that the complete metric parameter space is separable (e.g. Blackwell and MacQueen (1973)). Also, we do not consider a third parameter ρ satisfying $g_j=2^{-aj}c_1 j^{\rho}$.

In univariate notation, the prior on the wavelet coefficients is $f(\beta_{jk}|g_j,\sigma^2)=N(0,\sigma^2 g_j\gamma_{jk}I)$ and $f(g_j)\sim\mathrm{IG}(r_j,s_j)$ implies that $f(\beta_{jk}|\gamma_{jk},\sigma^2)=t_{s_j}(0,r_j\gamma_{jk}\sigma^2)$. We shall need the moments

$$E\|\boldsymbol{\beta}_j\|_p^p=\sum_k E(\beta_{jk}^p),$$

$$E(\|\boldsymbol{\beta}_j\|_p^{2p})=\sum_k E(\beta_{jk}^{2p})+\sum_{k\neq k'}E(\beta_{jk}\beta_{jk'})^p,$$

where $\boldsymbol{\beta}_j$ are the coefficients at the jth resolution. If ν_p is the pth moment of $N(0,1)$, then

$$E(\beta_{jk}^p)=E\{E(\beta_{jk}^p|g_j)\}=2^{-bj}c_2\sigma^p\nu_p\,E(g_j^{p/2}).$$

Thus, we have

$$E(\|\boldsymbol{\beta}_j\|_p^p) = 2^{(1-b)j} c_2 \sigma^p \nu_p \, E(g_j^{p/2}),$$
$$\begin{aligned} E(\|\boldsymbol{\beta}_j\|_p^{2p}) &= 2^{(1-b)j} c_2 \sigma^{2p} \nu_{2p} E(g_j^p) + 2^j (2^j - 1) \times 2^{-2bj} c_2^2 \sigma^{2p} \nu_{2p}^2 \, E(g_j^p) \\ &\leqslant 2^{(1-b)j} c_2 \sigma^{2p} \nu_{2p} (1 + 2^{(1-b)j} c_2 \nu_{2p}) \, E(g_j^p). \end{aligned}$$

Given these moments, by the Chebyshev inequality, we have, for some $\varepsilon > 0$,

$$\Pr\{|2^{-(1-b)j} \|\boldsymbol{\beta}_j\|_p^p - \sigma^p c_2 \nu_p \, E(g_j^{p/2})| > \varepsilon\} \leqslant 2^{-2(1-b)j} \varepsilon^2 \, E(\|\boldsymbol{\beta}_j\|_p^{2p})$$

and applying the Borel–Cantelli lemma

$$\sum_{j=0}^{\infty} \Pr\{|2^{-(1-b)j} \|\boldsymbol{\beta}_j\|_p^p - \sigma^p c_2 \nu_p \, E(g_j^{p/2})| > \varepsilon\} \leqslant \varepsilon^2 \sum_{j=0}^{\infty} 2^{-2(1-b)j} \, E(\|\boldsymbol{\beta}_j\|_p^{2p}) < \infty.$$

Thus $2^{-(1-b)j} \|\boldsymbol{\beta}_j\|_p^p \rightarrow c_2 \sigma^p \nu_p \, E(g_j^{p/2})$ almost surely.

Since this is true for $j = 0, 1, \ldots$, the Besov sequence norm is finite if

$$\sum_{j=0}^{\infty} 2^{j(l+1/2-1/p)q} \times 2^{(1-b)jq/p} \, E(g_j^{p/2})^{q/p} < \infty.$$

The infinite sum on the right-hand side is finite, if $E(g_j^{p/2})^{1/p} \propto 2^{-aj}$ for all $j = 0, 1, \ldots$, where a satisfies $(b-1)/p + (a-1)/2 \geqslant l - 1/p$.

To summarize, if, for all j,

$$E(g_j^{p/2})^{1/p} = \frac{r_j^{1/2}}{\{(s_j - 2)(s_j - 4) \ldots (s_j - p)\}^{1/p}} = 2^{-aj} c_1$$

where a satisfies $(b-1)/p + (a-1)/2 \geqslant l - 1/p$, then the Besov correspondence holds.

A.2.1. Proof of theorem 1

We consider the conditional posterior of $(\boldsymbol{\beta}_{n+1} | \mathbf{Y}, \boldsymbol{\beta}_{-(n+1)}, \sigma^2)$ in the oversmoothed model. The probability that $\boldsymbol{\beta}_{n+1}$ is not tied to any of the previous samples is

$$q_{n+1} = \frac{\sum_{i=1}^{n} \phi(\mathbf{Y}_{n+1} | \mathbf{X}\boldsymbol{\beta}_i, \sigma^2 \mathbf{I}_m)}{\alpha \, \phi\{\mathbf{Y}_{n+1} | 0, \sigma^2 (\mathbf{I}_m + \mathbf{XVX}')\} + \sum_{i=1}^{n} \phi(\mathbf{Y}_{n+1} | \mathbf{X}\boldsymbol{\beta}_i, \sigma^2 \mathbf{I}_m)} \leqslant \frac{1}{1 + \alpha R_{n+1}^* / n}$$

where

$$R_{n+1}^* = \frac{\phi\{\mathbf{Y}_{n+1} | 0, \sigma^2 (\mathbf{I}_m + \mathbf{XVX}')\}}{\phi(\mathbf{Y}_{n+1} | \mathbf{X}\bar{\boldsymbol{\beta}}, \sigma^2 \mathbf{I}_m)}$$

and $\bar{\boldsymbol{\beta}} \in \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n\}$ subject to $\|\mathbf{Y}_{n+1} - \mathbf{X}\bar{\boldsymbol{\beta}}\|_2$ is minimum. Also, we can write

$$\begin{aligned} \log\{E(R_{n+1}^*)\} &= \log\left[E_\beta \left\{ \frac{\phi(\mathbf{Y}_{n+1} | \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_m)}{\phi(\mathbf{Y}_{n+1} | \mathbf{X}\bar{\boldsymbol{\beta}}, \sigma^2 \mathbf{I}_m)} \right\}\right] \geqslant E_\beta \left[\log\left\{ \frac{\phi(\mathbf{Y}_{n+1} | \mathbf{X}\beta, \sigma^2 \mathbf{I}_m)}{\phi(\mathbf{Y}_{n+1} | \mathbf{X}\bar{\beta}, \sigma^2 \mathbf{I}_m)} \right\}\right] \\ &= \frac{1}{2\sigma^2} \{\|\mathbf{Y}_{n+1} - \mathbf{X}\bar{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{Y}_{n+1}\|_2^2 - \sigma^2 \operatorname{tr}(\mathbf{V})\} \equiv \log(R_{n+1}). \end{aligned}$$

Finally, writing $q_{n+1} \leqslant (1 + \alpha R_n / n)^{-1}$, we obtain

$$\begin{aligned} E_{\mathbf{Y}_{n+1}}(q_{n+1}) &\leqslant E\left(\frac{1}{1 + \alpha R_n / n}\right) \\ &\approx \frac{1 + \alpha \, E(R_n / n)}{\exp[2 \, E\{\log(1 + \alpha R_n / n)\}]}. \end{aligned}$$

Since R_n is small for large n,

$$\exp[2 \, E\{\log(1 + \alpha R_n / n)\}] = \exp\{2\alpha \, E(R_n / n)\} - O(n^{-2})$$

and, if $\mathbf{X}\boldsymbol{\beta}_{n+1}$ is the actual function underlying $\mathbf{Y}_{n+1}$, we have

$$\begin{aligned} E_{\mathbf{Y}_{n+1}}(q_{n+1}) &\leqslant \frac{1+\alpha\, E(R_n/n)}{\exp\{2\alpha\, E(R_n/n)\}}, \\ E(R_n) &= \exp(5\|\bar{\boldsymbol{\beta}}\|_2^2/8\sigma^2)\exp\{-\mathrm{tr}(\mathbf{V})/2\}\,\exp\left(-\frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\beta}_{n+1}\right) \end{aligned} \qquad (14)$$

since $E\{\exp(\mathbf{u}^{\mathrm{T}}\mathbf{x})\} = \exp(\mathbf{u}^{\mathrm{T}}\boldsymbol{\mu} + \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}/2)$ for $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We assume that $\|\bar{\boldsymbol{\beta}}\|_2, \|\boldsymbol{\beta}_{n+1}\|_2 < \infty$ almost surely by prior specification and let $\rho_{n+1} = \bar{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\beta}_{n+1}$ be the inner product of the actual functional and the closest available functional. For large n the sample space becomes dense and we expect ρ_{n+1} to increase. Then from expression (14)

$$\sum_{n=1}^{\infty} E_{\mathbf{Y}_{n+1}}(q_{n+1}) \leqslant \sum_{n=1}^{\infty} \frac{1+(C_3/n)\exp(-\rho_{n+1}/2\sigma^2)}{\exp\{(C_4/n)\exp(-\rho_{n+1}/2\sigma^2)\}} < \infty, \qquad C_3, C_4 > 0,$$

if $\rho_{n+1} \approx \sigma^2 \log\{\log(n)^{1+\delta}\}$ for some $\delta > 0$. By the Borell–Cantelli lemma, this means that the new sample is almost surely distinct if the inner product or the l_2-distance is respectively less than or greater than $\sigma^2\, O[\log\{\log(n)^{1+\delta}\}]$.

References

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

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.

Banerjee, S. (2004) On geodetic distance computations in spatial modelling. *Biometrics*, **61**, 617–625.

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Basu, S. and Chib, S. (2003) Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Am. Statist. Ass.*, **98**, 224–235.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Polya urn schemes. *Ann. Statist.*, **1**, 353–355.

Bush, C. S. and MacEachern, S. N. (1996) A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **83**, 221–227.

Chaloner, K. and Brant, R. (1988) A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–660.

Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B, **62**, 681–698.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–402.

Daubechies, I. (1992) *Ten Lectures in Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

De Canditiis, D. and Vidakovic, B. (2004) Wavelet Bayesian block shrinkage via mixtures of normal inverse gamma priors. *J. Comput. Graph. Statist.*, **13**, 383–398.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, **90**, 577–588.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

Gelfand, A. E., Kim, H. K., Sirmans, C. F. and Banerjee, S. (2003) Spatial modelling with spatially varying coefficient processes. *J. Am. Statist. Ass.*, **98**, 387–396.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

James, G. and Sugar, C. (2003) Clustering for sparsely sampled functional data. *J. Am. Statist. Ass.*, **98**, 397–408.

Klevecz, R. R. (2000) Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition of expression array data. *Functnl Integr. Genom.*, **1**, 186–192.

Korwar, R. M. and Hollander, M. (1973) Contributions to the theory of Dirichlet Processes. *Ann. Probab.*, **1**, 705–711.

Kovac, A. and Silverman, B. W. (2000) Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Statist. Ass.*, **95**, 172–183.

Medvedovic, M. and Sivaganesan, S. (2002) Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Pensky, M. and Vidakovic, B. (2001) On non-equally spaced wavelet regression. *Ann. Inst. Statist. Math.*, **53**, 681–690.

Quintana, F. A. and Iglesias, P. L. (2003) Bayesian clustering and product partition models. *J. R. Statist. Soc.* B, **65**, 557–574.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. J., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and Non-linear Geostatistics*. Oxford: Clarendon.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the Yeast Saccharomyces cerevisiae by microarray hybridization. *Molec. Biol. Cell*, **9**, 3273–3297.

Sweldens, W. (1996) The lifting scheme: a custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, **3**, 186–200.

Vidakovic, B. (1998) Nonlinear wavelet shrinkage with Bayes rule and Bayes factors. *J. Am. Statist. Ass.*, **93**, 173–179.

Wakefield, J., Zhou, C. and Self, S. (2003) Modelling gene expression over time: curve clustering with informative prior distributions. 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.

Widmann, M. and Bretherton, C. S. (2000) Validation of mesoscale precipitation in the NCEP reanalysis using a new grid-cell data set for the northwestern United States. *J. Clim.*, **13**, 1936–1950.

Yeung, K., Fraley, C., Raftery, A. and Ruzzo, W. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.