Bayesian Analysis (2016) **11**, Number 1, pp. 71–98

Bayesian Clustering of Functional Data Using Local Features

Adam Justin Suarez[*] and Subhashis Ghosal[†]

Abstract. The use of exploratory methods is an important step in the understanding of data. When clustering functional data, most methods use traditional clustering techniques on a vector of estimated basis coefficients, assuming that the underlying signal functions live in the L_2-space. Bayesian methods use models which imply the belief that some observations are realizations from some signal plus noise models with identical underlying signal functions. The method we propose differs in this respect: we employ a model that does not assume that any of the signal functions are truly identical, but possibly share many of their local features, represented by coefficients in a multiresolution wavelet basis expansion. We cluster each wavelet coefficient of the signal functions using conditionally independent Dirichlet process priors, thus focusing on exact matching of local features. We then demonstrate the method using two datasets from different fields to show broad application potential.

Keywords: Dirichlet process prior, wavelets, exploratory analysis.

1 Introduction

Exploratory analysis of new data is an important first step to understanding many scientific questions. Cluster analysis is a popular tool in exploratory analysis to try to discover underlying group structure present in the data. The idea behind cluster analysis is to define sets of data points which are similar within groups and dissimilar between groups. The main question that is raised is what concept to use to define "similar". Many clustering techniques (especially hierarchical clustering) can be implemented based solely on a matrix of pairwise similarities (equivalently, dissimilarities). One obvious choice for a notion of similarity is that of distance, which is nearly always available since data are most commonly assumed to be elements of some metric space. However, distances are not the only possible choices, and may miss out on some important qualitative features. A similarity index can be chosen to be any function of two arguments, as long as it represents our qualitative view of what makes two observations similar.

The interest of this paper lies in functional data, where each subject under study gives rise to a noisy function observation. The clustering of functional data has applications in many scientific fields, such as clustering gene expression time series. Recently, functional data have received a lot of attention. In terms of clustering functional data, most of the work has been done from a frequentist perspective. One approach is to adapt clustering methods from multivariate analysis to functional data. Tarpey and

[*] North Carolina State University, ajsuarez@ncsu.edu
[†] North Carolina State University, ghoshal@stat.ncsu.edu

DOI: 10.1214/14-BA925

Kinateder (2003) generalized the k-means clustering to the functional data setting, and proved results analogous to those of multivariate k-means. James and Sugar (2003) used a random effects model to cluster sparsely sampled functional data, and dealt with the situation where the data are not observed on the same fixed time grid. One extension we do not pursue is heteroscedasticity, which was studied in Serban (2008). Functional data can often lead to viewing multiple time series in a new light. For a review of clustering of time series see Liao (2005).

We shall not choose to define similarity in terms of distance; instead, we define a similarity function, for functions with a finite wavelet basis expansion, that strongly encourages exact matches of coefficients, meaning that the corresponding observations come from the same component of the population. Since the observed functional data include error, we use a model-based approach to estimate the true functions and use the posterior distribution of the basis coefficients to compute an estimated true similarity index. In a Bayesian setting, Ray and Mallick (2006) used a truncated wavelet basis expansion and a Dirichlet process prior on their unknown joint distribution. Crandell and Dunson (2011) extended this model to species sampling model priors, and also allowed the basis to be unknown. Both of these approaches cluster curves based on all of their basis coefficients jointly. By a well-known clustering property of the Dirichlet process, this implies the prior belief that some of the underlying functions have all wavelet coefficients identical, and hence, the functions themselves are identical. Often, this would not be an acceptable assumption, and this is one aspect in which the current paper differs from most previous work. Petrone, Guindani, and Gelfand (2009) also dealt with this problem using "canonical curves," from which pieces of the observed functions are drawn. For example, Figure 1 shows, for the EEG data discussed later, how very few pairs of data points are within a distance that would be a reasonable estimate of the error standard deviation. If the procedure of Ray and Mallick (2006) were used on this data, very few observations would have positive posterior probability of sharing underlying functions (see Section 8).

We assign priors on the wavelet coefficients independently, where each individual coefficient gets a Dirichlet process prior distribution. The Dirichlet process allows for exact coefficient matches between functions while allowing for new values to arise also. The strength of the model is in the Bayesian approach, where the underlying coefficients across subjects are seen as exchangeable but correlated, and hence, allow for shared learning among them.

In this paper, we present a method for quantifying similarity of functional data that can be used in a hierarchical clustering scheme. Wavelet coefficient parameters are clustered separately, and we define a function that quantifies our preference for exact matching of coefficients. We present theoretical results relating to the interpretation of the center measure of our prior, and, additionally, provide asymptotic justification of the clustering performance of our method by analyzing the small variance performance. We then demonstrate the method's use on real datasets, and show competitive performance on a dataset with a known true clustering.

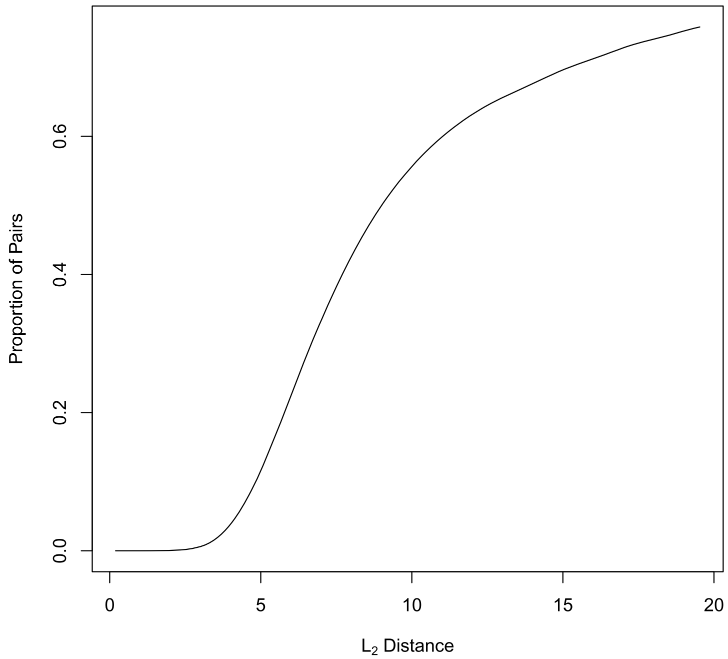

Figure 1: (EEG Data) Plot of the proportion of pairs of data points with distance less than a given value.

2 The Model

There are two common models, closely related in the asymptotic sense, that can be used to describe functional data. First let $\{\phi_k : k \in \mathbb{Z}\} \cup \{\psi_{jk} : j \in \mathbb{N}, k \in \mathbb{Z}\}$ be a given wavelet basis in the multiresolution framework. In particular, we consider the space $L_2[0,1]$ and the family called *wavelets on the interval* (Cohen, Daubechies, and Vial, 1993). The first model can be viewed as a problem of measurement error, where there is a true function, $f_i \in L_2([0,1])$, but when we measure it at a point $t_j \in [0,1]$, we only see a noisy version, so that

$$Y_i(t_j) = f_i(t_j) + \epsilon_{ij}, \tag{1}$$

where ϵ_{ij} is normally distributed with mean 0 and variance σ^2, and independent across i and j. We shall assume that all functions are observed on the same fixed time grid, and that the total number of time points is a power of 2, $n = 2^m$; this is done mainly for computational convenience since we will employ the discrete wavelet transform (DWT). Let $\boldsymbol{Y}_i = (Y_i(t_1), \ldots, Y_i(t_n))^t$, $\boldsymbol{f}_i = (f_i(t_1), \ldots, f_i(t_n))^t$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in})^t$. If we let $\boldsymbol{W}$ denote the $n \times n$ orthogonal matrix corresponding to the DWT for a certain wavelet family, then our model can be transformed to

$$\boldsymbol{W}\boldsymbol{Y}_i = \boldsymbol{W}\boldsymbol{f}_i + \boldsymbol{W}\boldsymbol{\epsilon}_i. \tag{2}$$

One important property of the multivariate normal distribution is its rotational invariance, which implies that $\boldsymbol{W}\boldsymbol{\epsilon}_i \overset{d}{=} \boldsymbol{\epsilon}_i$, meaning equal in distribution. Throughout, we shall use ϕ_σ to represent the Lebesgue density of the normal distribution with mean zero and variance σ^2.

The second model is the so-called Gaussian white noise model, given by

$$dY_i(t) = f_i(t)dt + \sigma dB_i(t), \tag{3}$$

where $B_i(\cdot)$ are independent Wiener processes (Brownian motions) on $[0, 1]$. This corresponds to ideal observations of continuously sampled functions. Now let

$$\begin{aligned} a_k^{(i)} &= \int_0^1 \phi_k(t)dY_i(t), & \alpha_k^{(i)} &= \int_0^1 \phi_k(t)f_i(t)dt, \\ b_{jk}^{(i)} &= \int_0^1 \psi_{jk}(t)dY_i(t), & \beta_{jk}^{(i)} &= \int_0^1 \psi_{jk}(t)f_i(t)dt, \\ \bar{e}_k^{(i)} &= \sigma\int_0^1 \phi_k(t)dB_i(t), & e_{jk}^{(i)} &= \sigma\int_0^1 \psi_{jk}(t)dB_i(t). \end{aligned}$$

Due to the properties of stochastic integrals with respect to the Wiener process, all of the $\bar{e}_k^{(i)}$ and $e_{jk}^{(i)}$ are independent and normally distributed with mean 0 and variance σ^2. The model implied on the wavelet coefficients by (3) is then

$$a_k^{(i)} = \alpha_k^{(i)} + \bar{e}_k^{(i)}, \qquad b_{jk}^{(i)} = \beta_{jk}^{(i)} + e_{jk}^{(i)}. \tag{4}$$

For the finite (measurement error) model, our decomposition is for $k = 0, \dots, 2^j - 1$, and $j = 0, \dots, m - 1$. Note that m is not a parameter, but the assumed length (in terms of its base 2 logarithm) of each observation vector. In the infinite (random function) model, j can range over the natural numbers. In both models, we only need $k = 0$ for the scaling coefficient, $\alpha_0^{(i)}$. Our procedure focuses mainly on the detail coefficients, $\{\beta_{jk}^{(i)}\}$, so we shall rarely mention the scaling coefficient. On the detail coefficients, for both models, the observed coefficients can be represented as

$$b_{jk}^{(i)} \overset{\text{ind}}{\sim} N(\beta_{jk}^{(i)}, \sigma^2). \tag{5}$$

In the results section of this paper, we shall also consider the case where the variance decreases to 0, which corresponds to the case of independent and identically distributed (*i.i.d.*) replications of the entire experiment. Because of this unifying framework, we shall not make much distinction between the finite and infinite model. We shall state results mainly for the infinite model; however, they can be shown to be true for the finite model using minor modifications.

3 Prior Distributions

One of the attractive aspects of using a wavelet expansion for modeling is that, for many functions, the coefficients are sparse. This knowledge is easily incorporated in the prior distributions placed on the wavelet coefficients. If the error is normally distributed, a conjugate prior on the coefficients is given by independent normal priors. Since we know that some of the coefficients are identically zero, we can incorporate a point mass at 0 into the prior. Specifically, the prior, for $i = 1, \dots, N$,

$$\beta_{jk}^{(i)} \overset{\text{ind}}{\sim} \pi_j N(0, \tau_j^2) + (1 - \pi_j)\delta_0, \tag{6a}$$

$$\sigma^2 \sim \text{IG}(a, b), \tag{6b}$$

where δ_0 is a point mass at 0 and IG stands for inverse gamma. The first coefficient, α_0, is known as the scaling coefficient, and is usually modeled differently, with a vague prior. In our case, we simply assume that the observations have been detrended, so that the value of α_0 is identically 0. Abramovich, Sapatinas, and Silverman (1998) showed that under certain conditions on the mother wavelet, choices of the hyperparameters in this model will guarantee that the corresponding random functions almost surely lie in specific Besov spaces (denoted by $B_{p,q}^s$). In particular,

$$\tau_j^2 = \nu_1 \sigma^2 2^{-\gamma_1 j}, \qquad \pi_j = \min(1, \nu_2 2^{-\gamma_2 j}), \tag{7}$$

for $j = 0, 1, \dots,$ and constants $\nu_1, \nu_2, \gamma_1, \gamma_2$, provide the desired interpretation. To be specific, if $\gamma_2 \geq 1$ and $\gamma_1 \geq 0$, then the random function drawn from this prior, f_i, is almost surely an element of $B_{p,q}^s$ if and only if

$$s + \frac{1}{2} - \frac{\gamma_2}{p} - \frac{\gamma_1}{2} < 0, \text{ if } q < \infty, \tag{8a}$$

$$s + \frac{1}{2} - \frac{\gamma_2}{p} - \frac{\gamma_1}{2} = 0, \text{ if } q = \infty. \tag{8b}$$

The main result of Abramovich et al. (1998) is given for any fixed value of the scaling coefficient.

Recently, Ray and Mallick (2006) extended the prior of Abramovich et al. (1998) to the setting of clustering functional data in the finite model case. Their approach was to assign a prior on the sequence of vectors $\boldsymbol{\beta}_i = \{\beta_{jk}^{(i)}\}_{j,k}$ in the following manner:

$$\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N \overset{\text{ind}}{\sim} F, \qquad F \sim \text{DP}(M, G_0), \tag{9}$$

where G_0 was the product of the priors from Abramovich et al. (1998) over j and k, and $\text{DP}(M, G_0)$ stands for the Dirichlet process with center measure G_0 and concentration M. This induces a posterior distribution over partitions of the data, but also implies the prior belief that some true functions are identically equal. When this is not a reasonable assumption, other choices should be made.

Our strategy for placing priors on the true wavelet coefficients is to do so independently for each coefficient using Dirichlet process priors, with center measures corresponding to the usual parametric models often used for modeling by wavelets. Thus, instead of modeling all the coefficients jointly, as in Ray and Mallick (2006), they will be done so independently. We also consider σ^2 to be unknown, and for this reason, we scale the variance in the base measure in the traditional manner. The full model is therefore, $\forall\ j,k$,

$$b_{jk}^{(i)}|\beta_{jk}^{(i)},\sigma \overset{\text{ind}}{\sim} N(\beta_{jk}^{(i)},\sigma^2), \tag{10a}$$

$$\beta_{jk}^{(1)},\ldots,\beta_{jk}^{(N)}|G_{jk} \overset{\text{iid}}{\sim} G_{jk}, \tag{10b}$$

$$G_{jk} \sim \text{DP}(M, G_{jk}^0), \tag{10c}$$

$$G_{jk}^0 = \pi_j N(0,\sigma^2\tau_j^2) + (1-\pi_j)\delta_0, \tag{10d}$$

$$\sigma^2 \sim \text{IG}(a,b). \tag{10e}$$

Across levels of (j,k), the random variables, $\{G_{jk}\}$, are independent. If $X_1, X_2, \ldots |F \overset{\text{iid}}{\sim} F$ where $F \sim \text{DP}(M, G_0)$, then the predictive distribution of the sequence satisfies:

$$P(X_{n+1} \in \cdot|X_1,\ldots,X_n) = \frac{M}{M+n}G_0 + \sum_{j=1}^{k(n)} \frac{m_{j,n}}{M+n}\delta_{X_j^*}, \tag{11}$$

where $X_1^*,\ldots,X_{k(n)}^*$ are the $k(n)$ distinct points in the first n observations, and $m_{j,n} = \#\{i : X_i = X_j^*\}$, for $j = 1,\ldots,k(n)$. This is the so-called Pólya urn representation of the Dirichlet process (Blackwell and MacQueen, 1973). The Dirichlet model is particularly useful when it is known that there is a "typical" cluster with some smaller "abnormal" groups.

This model tries to capture our belief that the functional data share local features that are expressed in their wavelet expansions. We also want to incorporate the knowledge of the possibility of an exactly zero wavelet coefficient, and we do that within the base measure of the Dirichlet process prior.

Remark 1. *In Section 5, we show that all but finitely many coefficients are zero from any realization of the prior. This motivates a different approach to constructing a prior. For all $\{j,k\}$ such that $j > J_*$, let $(\beta_{jk}^{(1)},\ldots,\beta_{jk}^{(N)}) = \mathbf{0}$. We then allow J_* to be random and have a Poisson distribution with parameter λ. The decay of π_j and τ_j are now not essential, since the resulting wavelet series is always convergent. As before, we have that the number of coefficients and levels are almost surely finite. This allows more freedom in the choice of π_j since the quick decay is no longer needed to give us this property. It is still useful, though, to keep the point mass at zero to account for reasonable prior beliefs about the wavelet expansion. This prior seems to be a much more natural choice, and even yields the later results more easily, but comes at the price of increased computational complexity.*

4 The Similarity Matrix and Clustering

With the goal of comparing the similarity between functions, we have many choices. Recalling that there are $n = 2^m$ sampled time points, and excluding one corresponding to the scaling coefficient, we choose to quantify the similarity between two functions using the similarity index

$$S(i,i') = (2^m - 1)^{-1} \sum_{j=0}^{m-1} \sum_{k=0}^{2^j-1} \mathbb{1}(\beta_{jk}^{(i)} = \beta_{jk}^{(i')}), \tag{12}$$

the average number of shared wavelet coefficients. This quantity is meaningful in our model since the Dirichlet process will give positive probability to this value being nonzero. The matrix is easily estimated using posterior samples from Markov Chain Monte Carlo (MCMC) output.

Once an estimate is constructed, the obtained matrix can be used with any clustering method taking a (dis)similarity matrix as its input. In fact, given a clustering procedure, our method can be viewed as providing a posterior distribution on dendrograms (for example). Primarily, however, we shall employ the posterior mean matrix to provide a single output from the chosen clustering algorithm.

5 Interpretation of Prior Characteristics

In this section, we explore and review some of the properties of the previous model of Abramovich et al. (1998) for a single function (the nonparametric regression setting). Instead of studying the model under a fixed value for the hyperparameter, π_j, we consider the limiting case where $\gamma_2 \to 1$ from above, where π_j is also scaled by a factor. In the following, we let $\gamma_2 = 1 + \delta$ and consider $\delta \to 0$, so we have

$$\pi_j = \nu_2 \delta 2^{-(1+\delta)j} \quad \text{where } \delta > 0,\ \delta \to 0. \tag{13}$$

The reason behind this choice is to approximate the situation where $\gamma_2 = 1$ in the original hyperparameter choice, while keeping almost surely finiteness of the number of terms in the wavelet expansion. It is needed that ν_2 be scaled by δ so that, in the limit, the quantities of interest remain finite, else they would diverge without it to balance the growth.

The following proposition would be useful for prior elicitation in the case where the approach mentioned in Remark 1 was taken. It motivates and justifies the use of a Poisson prior on the number of nonzero coefficients and resolution levels, and provides an interpretation of their hyperparameters in this setting. For a proof, see Appendix A.

Proposition 1. *For the infinite product of the priors specified as above, the following hold:*

1. *The number of nonzero wavelet coefficients is almost surely (a.s.) finite, and this number converges in distribution to a Poisson random variable with mean $\nu_2/\log(2)$ as $\delta \to 0$.*

2. *The number of resolution levels with at least one nonzero coefficient is a.s. finite, and this number converges in distribution as $\delta \to 0$ to a Poisson random variable with mean*

$$\lim_{\delta \to 0} \sum_{j=0}^{\infty} \left\{ 1 - (1 - \nu_2 \delta 2^{-(1+\delta)j})^{2^j} \right\} < \infty.$$

6 Convergence Results

In the present situation, we first want to study what happens to our similarity matrix as the noise variance $\sigma^2 \to 0$. This would be the situation where the noisy functional observations are approaching the true underlying functions, respectively. For the purposes of this section, we assume the continuous model of (3), with the full specification in terms of the coefficients being given in (10a).

The asymptotic regime $\sigma^2 \to 0$ can be understood as equivalent with averaging over r *i.i.d.* replications of the observed scheme (10a)–(10d) with $r \to \infty$, thus replacing σ^2 by $\sigma_r^2 = \sigma^2/r$, with σ^2 known. Since σ^2 itself controls the asymptotics in the following, it is essential to treat σ^2 as given, or equivalently, σ^2 as known and $r \to \infty$. Although this setting contrasts with the methodology described, this has little effect when only learning about $\boldsymbol{f}$ is the goal. More generally, it is easy to see that the arguments given below go through if σ^2 is unknown, but has a fixed upper bound. An upper-truncated inverse-gamma prior can still retain the computational conjugacy. If it is desirable to work in full generality without an upper bound for σ^2, we must fully observe all replications since the sample means are sufficient only when σ^2 is known. Below we forgo the full setting and treat σ^2 as known so that it is sufficient to observe the sample mean of $b_{jk}^{(i)}$ over r replications and let $r \to \infty$.

We assume that $\alpha_0^{(i)} = 0$ for all $i = 1, \ldots, N$, and let

$$\|\boldsymbol{f}\|^2 = \sum_{i=1}^{N} \|f_i\|_2^2 = \sum_{i=1}^{N} \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} |\beta_{jk}^{(i)}|^2,$$

where $\boldsymbol{f} = (f_1, \ldots, f_N)$. We also consider the Sobolev norm on the product space, defined by

$$\|\boldsymbol{f}\|_{\mathcal{H}_N^s}^2 = \sum_{i=1}^{N} \sum_{j=0}^{\infty} 2^{2js} \|\beta_{j\cdot}^{(i)}\|_2^2.$$

We shall refer to this space as the N-Sobolev space, $\mathcal{H}_N^s$. Note that since N is fixed, we could have chosen to combine the N Sobolev norms using any norm for $\mathbb{R}^N$. The parameter, s, relates to the number of weak derivatives possessed by the functions which themselves live in $L_2([0,1])$. We use D_r to be the set of all observations.

Before proceeding, we need some additional notation. The measure of similarity which holds our interest is dependent on how we believe the data to be partitioned. We will thus be interested in knowing how our beliefs about the partition structure of the data change as $r \to \infty$. Let $\mathfrak{P}$ be the set of all partitions of $\{1, \ldots, N\}$, and let a typical element be denoted by $\mathfrak{p} = \{A_0, \ldots, A_M\}$. For a given j, k, let $\mathcal{P}_{jk} = \{A_0^{jk}, A_1^{jk}, \ldots, A_{M_{jk}}^{jk}\}$ be a random partition of $\{1, \ldots, N\}$, which is a function of $\boldsymbol{\beta}_{jk}$ defined in the following way:

$$\beta_{jk}^{(a)} = 0 \iff a \in A_0^{jk}, \text{ and} \tag{14}$$

$$\beta_{jk}^{(a)} = \beta_{jk}^{(b)} \neq 0 \iff a, b \in A_i^{jk} \text{ for some } i \in \{1, \dots, M_{jk}\}. \tag{15}$$

By our prior specification, it is clear that any partition structure has positive probability *a priori*. Let $\mathfrak{p}_0$ represent the "true" partition generated by the true values of the parameters. By a *compatible model*, we mean a collection of all parameter values corresponding to a single partition which is finer than $\mathfrak{p}_0$. By an *incompatible model*, we mean any collection that is not a *compatible model*.

The following result on consistency of the posterior will be useful for studying the asymptotic properties of clustering. The techniques used in the proofs are both similar to, and certainly inspired by Lian (2011). For proofs of the following, see Appendix A.

Theorem 1. *Let $\gamma_1 > 2s + 1$, and assume that the true underlying functions satisfy $\boldsymbol{f}_0 \in \mathcal{H}_N^s$. Then the posterior is norm-consistent, i.e., for any $\epsilon > 0$, $\Pi(\|\boldsymbol{f} - \boldsymbol{f}_0\| < \epsilon | D_r) \xrightarrow{p} 1$ as $r \to \infty$.*

Lemma 1. *Assume that the true vector of functions lies in $\mathcal{H}_N^s$, $\gamma_1 > 2s + 1$, and let $\mathfrak{p}_{jk,0}$ be the true partition of the data for a given coefficient indexed by j, k. Then*

$$\Pi(\mathcal{P}_{jk} = \mathfrak{p}_{jk,0} | D_r) \xrightarrow{p} 1 \text{ as } r \to \infty. \tag{16}$$

Finally, we consider neighborhoods of the true full model, that is, $\mathfrak{p}_0 = \{\mathfrak{p}_{jk,0}\}_{jk}$, in the product topology. Each $\mathfrak{p}_{jk,0}$ lives in the space, $\mathfrak{P}$, of all possible partitions of $\{1, \dots, N\}$, which is finite and endowed with the discrete topology. Note that the entire model space is uncountable. When considering the product space, a basic neighborhood in the product topology consists of the product of finitely many singleton sets in $\mathfrak{P}$ with infinitely many copies of $\mathfrak{P}$. Because of this, we easily obtain the following theorem.

Theorem 2. *Let $\mathfrak{p}_0$ be the true model. Then, for any neighborhood in the product topology, $N(\mathfrak{p}_0)$, we have that $\Pi(N(\mathfrak{p}_0)|D_r) \to 1$ in probability as $r \to \infty$.*

Proof. First notice that $N(\mathfrak{p}_0)$ consists of the product of finitely many single point sets with an infinite number of copies of the whole space. Thus, the probability of this neighborhood is the finite product of the probabilities of each point set, each of which tends to 1 by Lemma 1. Thus, the result is proved. □

7 Computation

Computation is done using the urn representation of the Dirichlet process prior, and we follow the procedure of Navarrete, Quintana, and Müller (2008). The main problem in posterior computation will be the fact that we employ an atomic base measure, and this means that in the urn representation, a value can be 0, either because it is tied to a previous value, or because it was drawn from the base measure. To simplify notation, we focus on updating one particular wavelet coefficient across observations, so we fix j, k and let β_i, for $i = 1, \dots, N$, be the parameter for observation i. Due to exchangeability in the Dirichlet process model, it suffices to describe the conditional posterior draws from $\beta_N | \left(\{\beta_i\}_{i=1}^{N-1}, \{b_i\}_{i=1}^{N}, \sigma^2\right)$ and $\sigma^2 | \left(\{\beta_i\}_{i=1}^{N}, \{b_i\}_{i=1}^{N}\right)$. Similar to

Section 2, let $\beta_1^*, \ldots, \beta_{k(N-1)}^*$ be the $k(N-1)$ unique values among the first $N-1$ parameters. Let $\boldsymbol{m}_{N-1} = \{m_{1,N-1}, \ldots, m_{k(N-1),N-1}\}$, where $m_{j,N-1} = \#\{1 \leq i \leq N : \beta_i = \beta_j^*\}$. When needed, we shall additionally subscript quantities to designate the (j,k) level. The Gibbs sampling algorithm executes the following steps:

- Set β_N equal to β_l^* with probability proportional to

$$\frac{m_{l,N-1}}{M+N-1}\phi_\sigma(b_i - \beta_l^*) = \frac{m_{l,N-1}}{M+N-1}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(b_N - \beta_l^*)^2\right\}.$$

- With probability proportional to

$$\begin{aligned}\frac{M}{M+N-1}\int \phi_\sigma(b_N - \beta)dG_0(\beta) = \\ \frac{M}{M+N-1}(2\pi\sigma^2)^{-1/2}e^{-1/(2\sigma^2)} \\ \times\left[1 - \pi_j + \pi_j\left(\frac{\tau^2}{1+\tau_j^2}\right)^{1/2}\exp\left\{\frac{b_N^2\tau_j^2}{2\sigma^2(1+\tau_j^2)}\right\}\right],\end{aligned}$$

  sample β_N from the following distribution

$$\pi^*\delta_0 + (1-\pi^*)N\left(\frac{b_N\tau_j^2}{1+\tau_j^2}, \frac{\sigma^2\tau_j^2}{1+\tau_j^2}\right),$$

  where

$$\pi^* = \left(1 + \frac{\pi_j}{1-\pi_j}\left(\frac{1}{1+\tau_j^2}\right)^{1/2}\exp\left\{\frac{\tau_j^2 b_N^2}{\sigma^2(1+\tau_j^2)}\right\}\right)^{-1}.$$

- Sample σ^2 from the following distribution:

$$IG\left(a + Nn/2 + \sum_{j,k}k_{ij}(N)/2, b + \frac{1}{2}\sum_{i,j,k}(b_{jk}^{(i)} - \beta_{jk}^{(i)})^2 + \sum_j \frac{1}{2\tau_j^2}\sum_k\sum_{i=1}^{k_{jk}(N)}{\beta_{jk,i}^*}^2\right).$$

- Finally, update any hyperparameters that have been added to the prior specification.

The main issue to notice is that both possibilities for β_N can lead to a value of 0 (either being tied to an existing point which happens to be 0, or drawing a 0 from the base measure), and this needs to be taken into account when fitting the model, which simply requires careful bookkeeping. After each draw from these full conditionals, we need to update the unique points, along with $k(N-1)$, and $\boldsymbol{m}_{N-1}$.

8 Applications to Data

In this section, we present the usefulness of the above method by analyzing two different data sets. When presenting results, often it is convenient to display the similarity matrix after it has been used in a deterministic hierarchical clustering scheme. In particular, we apply Ward's method of clustering (Ward Jr., 1963) to the dissimilarity matrix defined by $\{1/S^*(i,i')\}_{i,i'=1}^{N}$. This method is an agglomerative method, in which a single element is joined with an existing group, so that the sum of the variances of all groups is minimized. Other hierarchical clustering methods are also possible.

During the analysis, it was noted that the choice of ν_1 in the prior was both difficult to make *a priori* and also strongly influenced the results. This situation was addressed by use of a hyperprior for ν_1 of a conditionally conjugate inverse gamma distribution. This did not cost much in terms of computation, and also provided more robust results.

In both examples, there are rational preconceived notions of how reasonable results should appear. This type of example was chosen to establish confidence that, when used for purely exploratory analysis, we have the potential to find meaningful relationships between observations. All three sets of data also fit well into the model.

Although we explored the use of the method of Ray and Mallick (2006) on these data, we do not present the results of that analysis. To get a meaningful number of nonzero entries in the corresponding similarity matrix, or matrix of pairwise probabilities of shared group membership, either the Dirichlet process concentration parameter is required to be nearly zero, or the *a priori* probability of a zero wavelet coefficient is required to be very large. Since the method of Ray and Mallick (2006) is not intended for data for which the belief of identical true functions does not hold, we do not show the comparison in this section.

The method was coded in `C` and made use of the GNU Scientific Library (Galassi et al., 2009). It is available at the author's website (`http://www4.ncsu.edu/~ajsuarez`), in addition to the supplemental materials.

8.1 EEG Data

The first dataset is from a study by Andrzejak et al. (2001), which is freely available online (`http://epileptologie-bonn.de/cms/front_content.php?idcat=193`). The data consist of 500 electroencephalography (EEG) time series, each of length 4096, corresponding to a sampling rate of 173.61 Hz. Because of the periodic nature of the data and computational considerations, only the first 128 time points were used. For the MCMC algorithm, $10,000$ steps were used, including $1,000$ steps for burn-in. On a 3.6 GHz AMD Bulldozer-powered desktop computer running single-threaded, this chain took approximately 6.5 hours to run.

The data are combined from 5 separate groups of data, which Andrzejak et al. (2001) label as A–E. Sets A and B came from measurements on healthy individuals, while the rest are from patients who suffered from seizures, and who had later been treated with corrective surgery. The important fact about the data is that the observations from set E are all from known seizure activity.

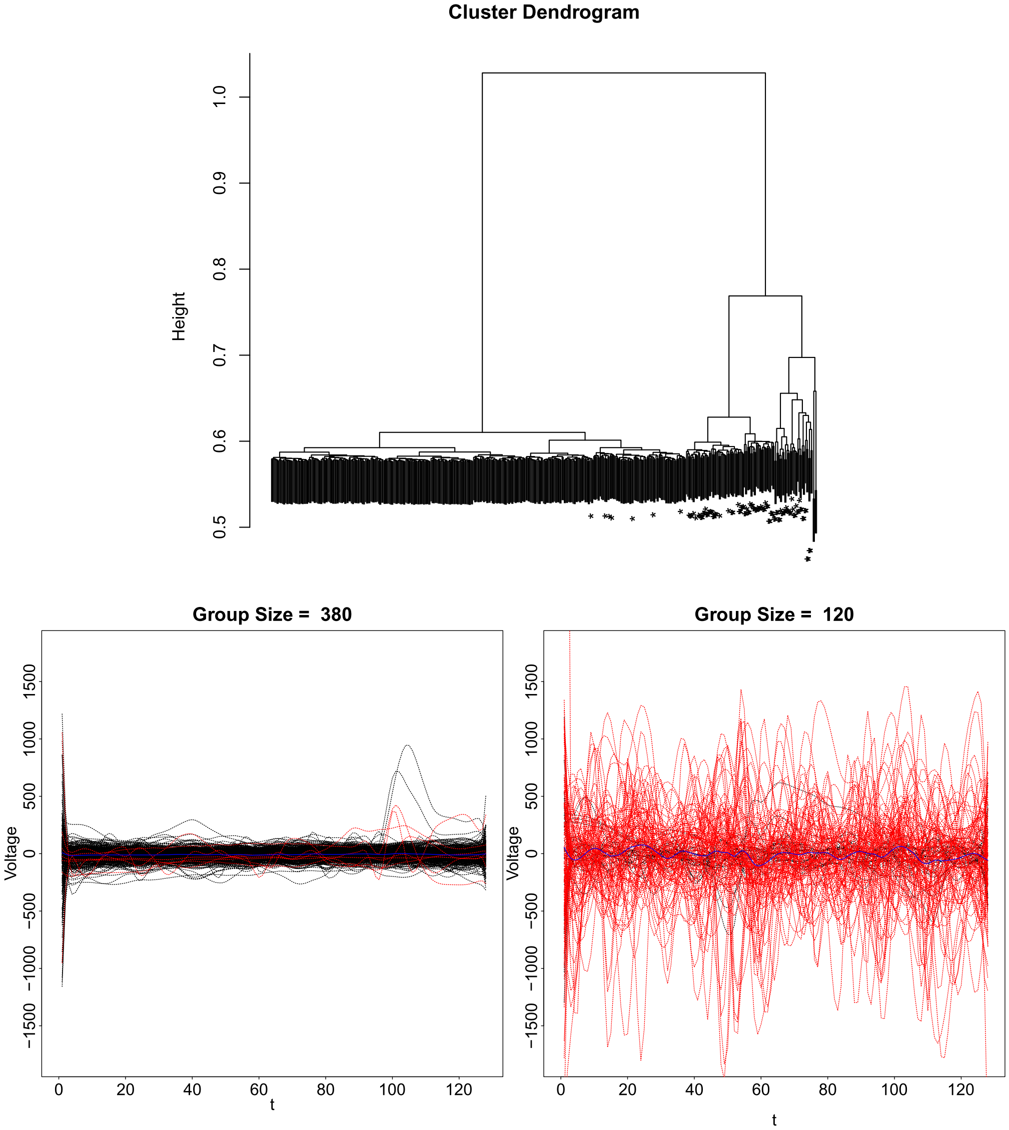

Figure 2: (EEG Data) (Top) Dendrogram generated using the dissimilarity matrix by Ward's method. The stars on the margins represent observations from the fifth group (suspected seizure activity). (Bottom) Groups formed when the dendrogram is cut to yield 2 groups. Dashed lines are pointwise posterior means for the observations, and solid blues lines are the pointwise group average. These figures correspond to $M = 20$.

Figure 2 shows some graphical representations of the results. The leaves of the dendrogram are marked with stars to denote observations coming from set E. The

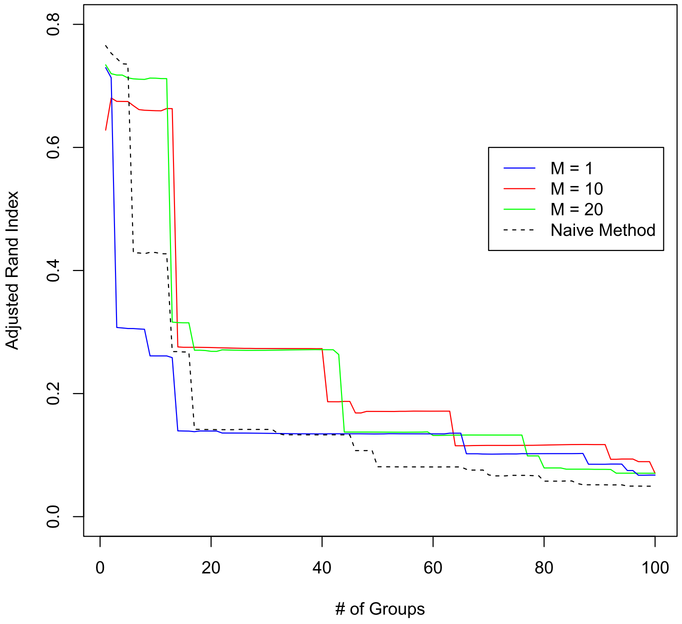

Figure 3: (EEG Data) Adjusted Rand index comparison between 3 hyperparameter choices and a default method. The clustering was evaluated compared to the "true" clustering defined by whether the observation was from known seizure activity.

dendrogram clearly shows a group that seems quite different from the vast majority of the others. In general, this group corresponds to the known seizure activity. We also display the results of obtaining a non-hierarchical clustering by cutting the dendrogram at a given height level. We chose to form 2 groups, and display the results also in Figure 2. Both groups are plotted on the same voltage range. The first group clearly has much lower voltage swings. Large voltage swings are characteristic of seizures (Andrzejak et al., 2001). Thus, the method has yielded a very interpretable result consistent with that known from neurobiology.

Although we apply a clustering algorithm to this data set, there is classification information available since we know which subset came from seizure activity. To evaluate the performance of our method, we use the following procedure: first, obtain a hierarchical clustering using the method described above. Subsequently, starting with 2 groups, cut the tree at various levels, and evaluate the strict clustering by computing the adjusted Rand index compared with the "true" clustering. For this comparison, we used 3 different values of M, and also compared this method to a default method. The default method was to use the same deterministic portion of our method, i.e., Ward's method, but with a dissimilarity matrix defined by the Euclidean distance between observation vectors (an estimated L_2-distance). The results of this comparison are shown in Figure 3. As can be seen from the results, although our method depends on the choice of mass parameter, M, for a moderate number of groups, the results are very similar

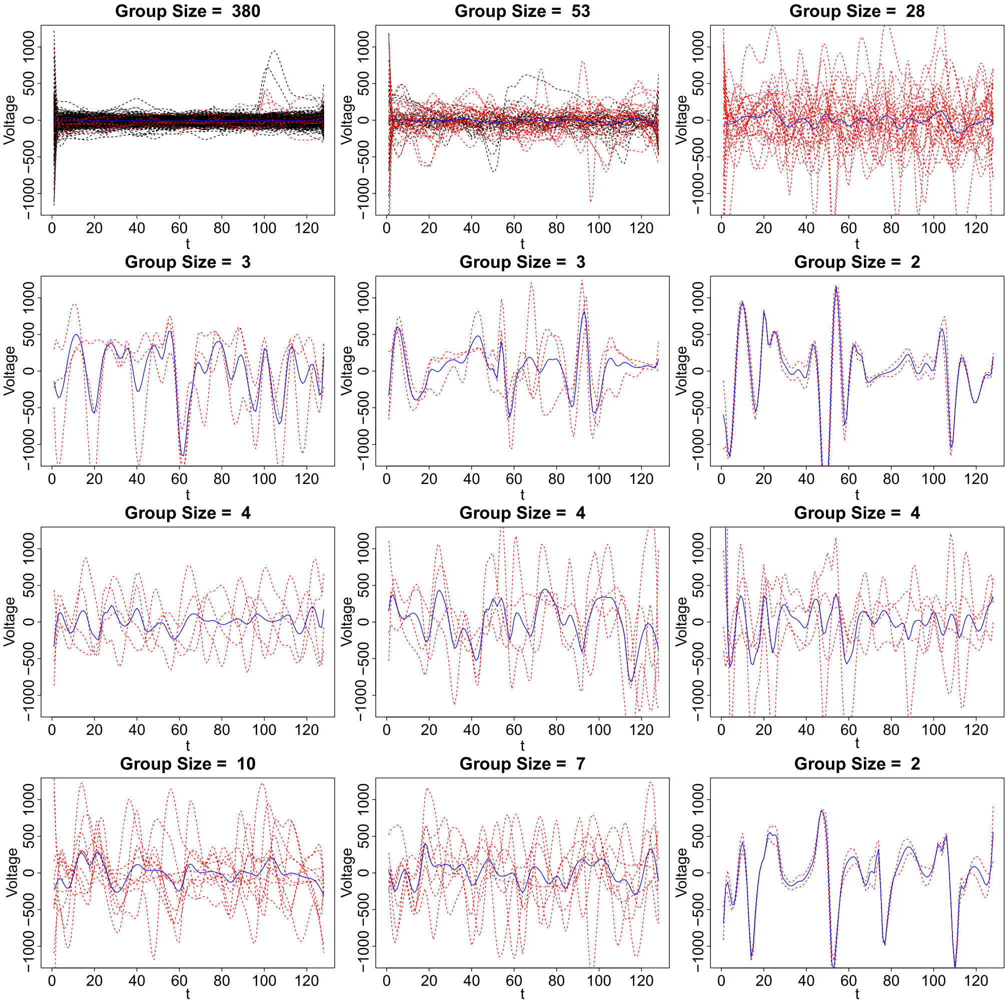

Figure 4: (EEG Data) Clustering the EEG data by cutting the tree at 12 groups. This corresponds to the last cut point that still maintains a relatively high adjusted Rand index.

between choices. For two of the choices, $M = 10, 20$, our method outperforms the default method throughout most choices of cut point, except for the smallest number of groups. Since, in practice, many choices of groups are likely to be explored, this gives reason to believe that our method can certainly aid in this exploration.

Using the adjusted Rand index criteria just described, we present clustering corresponding to cutting the tree at 12 groups for the model $M = 20$. This is the point just before the drop-off in the adjusted Rand index seen in Figure 3. This clustering is shown in Figure 4.

8.2 Canadian Weather Data

The next analysis involves the very popular Canadian weather data, which was obtained via the `fda` package within `R.` These data consist of both average daily temperature and precipitation for 35 Canadian cities. Specifically, we analyzed the precipitation data for the first 256 days of the year. This was done to allow use of the DWT for our method. These data were chosen because of the known spatial and climate correlation for precipitation. As can be seen in Figure 5, the method certainly reproduces this correlation. This is clear from the map, and those familiar with Canadian geography can inspect the heatmap more closely. Although not presented in this paper, a naive approach based on L_2-distance between observations does not nearly show as much spatial clustering. This makes this dataset "harder" than the previous EEG data, in that a naive approach to the EEG data can yield a reasonable, but less clear, description of the data. Again, this example clearly demonstrates the ability of this method to find structure between observations.

Since there is no objectively true clustering for these data, we do not compare to any other methods; however, as in the previous example, we can still analyze the effect that the choice of M has on the results. We focused on three different values, $M = 1, 10, 20$. We present two comparisons of the performance of the methods with these values: first, Figure 6 shows the approximate posterior distribution of number of distinct values for a range of coefficients in the model. As would certainly be expected, the number of groups increases as the mass parameter, M, increases because this controls the prior probability of an exact coefficient match.

To see how this affects the end result of forming a strict clustering, we show, in Figures 7–9, the end result of cutting the dendrogram obtained by Ward's method at 6 groups. It can be seen from these plots that there is a subtle, but noticeable, effect on the results from different choices of M. Figure 5 shows that they generally correspond to physical proximity between the cities. This could be due to the fact that, although the cities' climate differs in overall trends, local variations are shared, which is something our method was hoping to emphasize.

9 Discussion

There is an important point to note with respect to the ability of this procedure to generalize to priors other than the Dirichlet process used herewithin. In the description of the predictive distribution corresponding to the Dirichlet process, (11), the form suggests the possibility of a generalization to other so-called species sampling models (SSMs). SSMs are random measures for which, under certain conditions, conditionally *i.i.d.* sequences have predictive distributions of the same general form as (11) (Pitman, 1996). However, as pointed out by the Associate Editor, the Dirichlet process is the only member of the class of SSMs for which (11) will be valid when using a base measure with atoms.

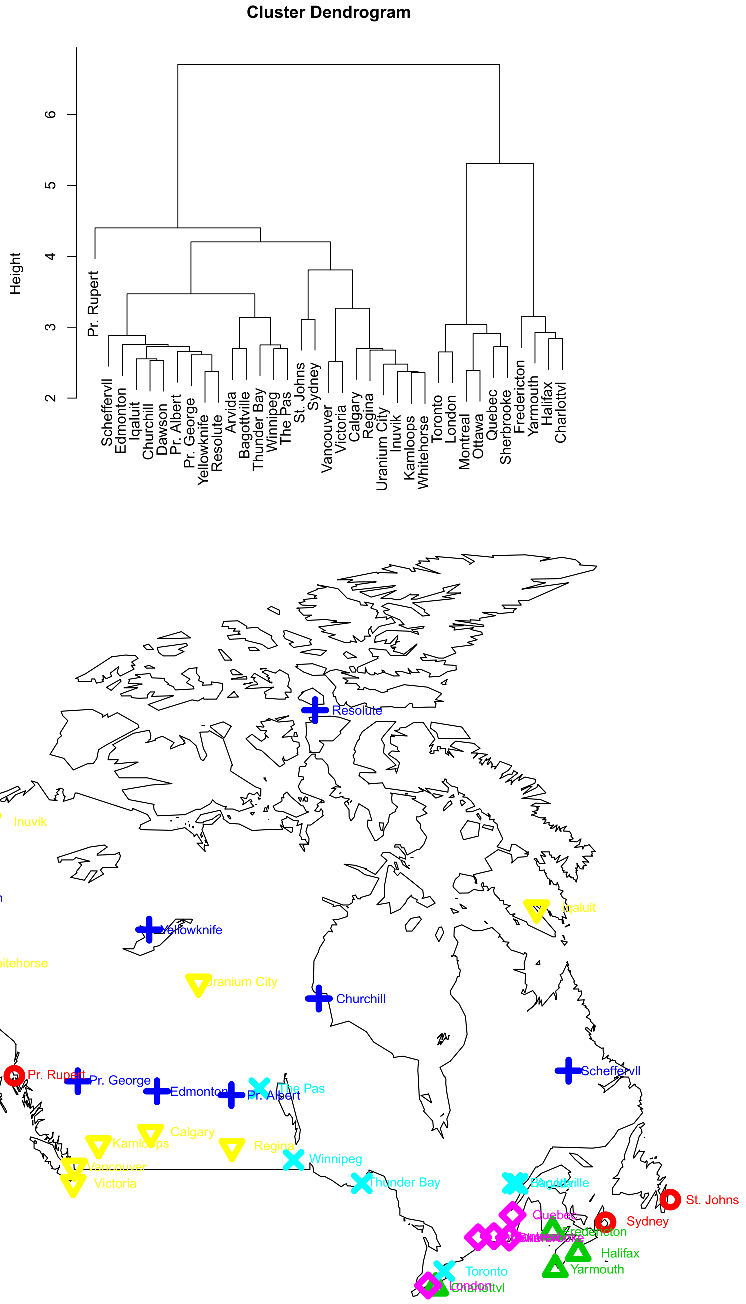

Figure 5: (Weather Data) (Top) Dendrogram created from the dissimilarity matrix by Ward's method. (Bottom) Map of the cities coded by symbol in color to represent the groups formed when the dendrogram is cut to yield 6 groups. These plots correspond to $M = 10$.

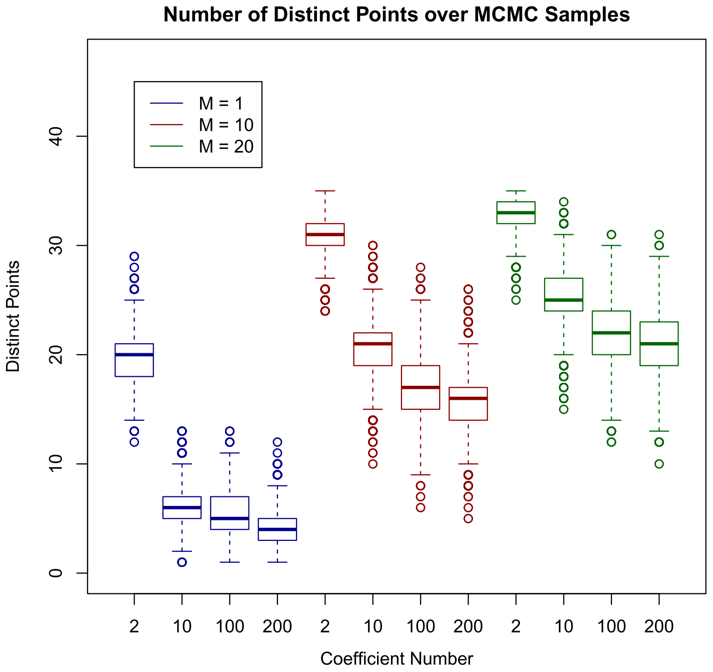

Figure 6: (Weather Data) Boxplots of the number of distinct points in a given MCMC step for a range of coefficients and for different values of the mass parameter, M. The coefficient numbers were chosen arbitrarily to show the clustering at different resolution levels.

Generalization would therefore require some effort in one of two possible directions. First, predictive distributions could be derived for the case of a single atom in the base measure. The other option would be to remove the atom from the base measure and incorporate the probability of a zero value elsewhere in the hierarchy. This second option would provide an advantage that the distribution of zeroes could be chosen arbitrarily instead of being implied by the choice of SSM (which is beta for the Dirichlet process).

We would like to mention a connection between the structure of the prior and the Indian Buffet Process in the special case of the Dirichlet process prior. Typically the most important aspect of a wavelet coefficient is whether it is zero or nonzero since this respectively indicates the absence or presence of the corresponding term, and thus it is an important indicator of the sparsity of the wavelet expansion. When the quantity of interest is only the indicator that a given coefficient is zero or not, for a given observation, this can be viewed as a binary string. Section 5 shows that the distribution of the total number of ones in a given string is distributed as a Poisson random variable with some rate. Because of the use of the Dirichlet process prior, the successive observations are correlated, and, in particular, exchangeable. For our prior, the probability that the

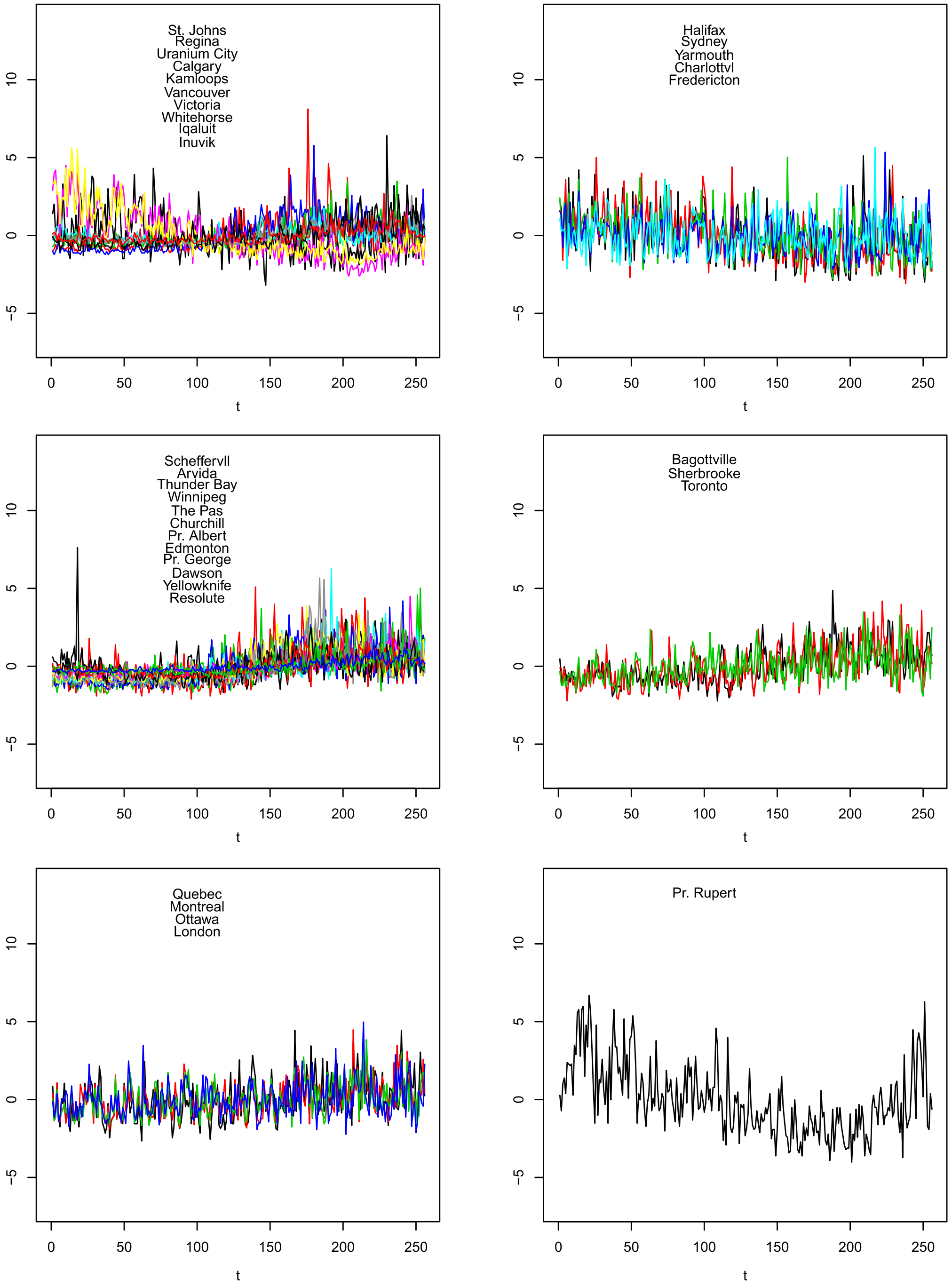

Figure 7: (Weather Data) Clustering formed by cutting at 6 groups for $M = 1$. For each group, the observed functional data are plotted on the same axes.

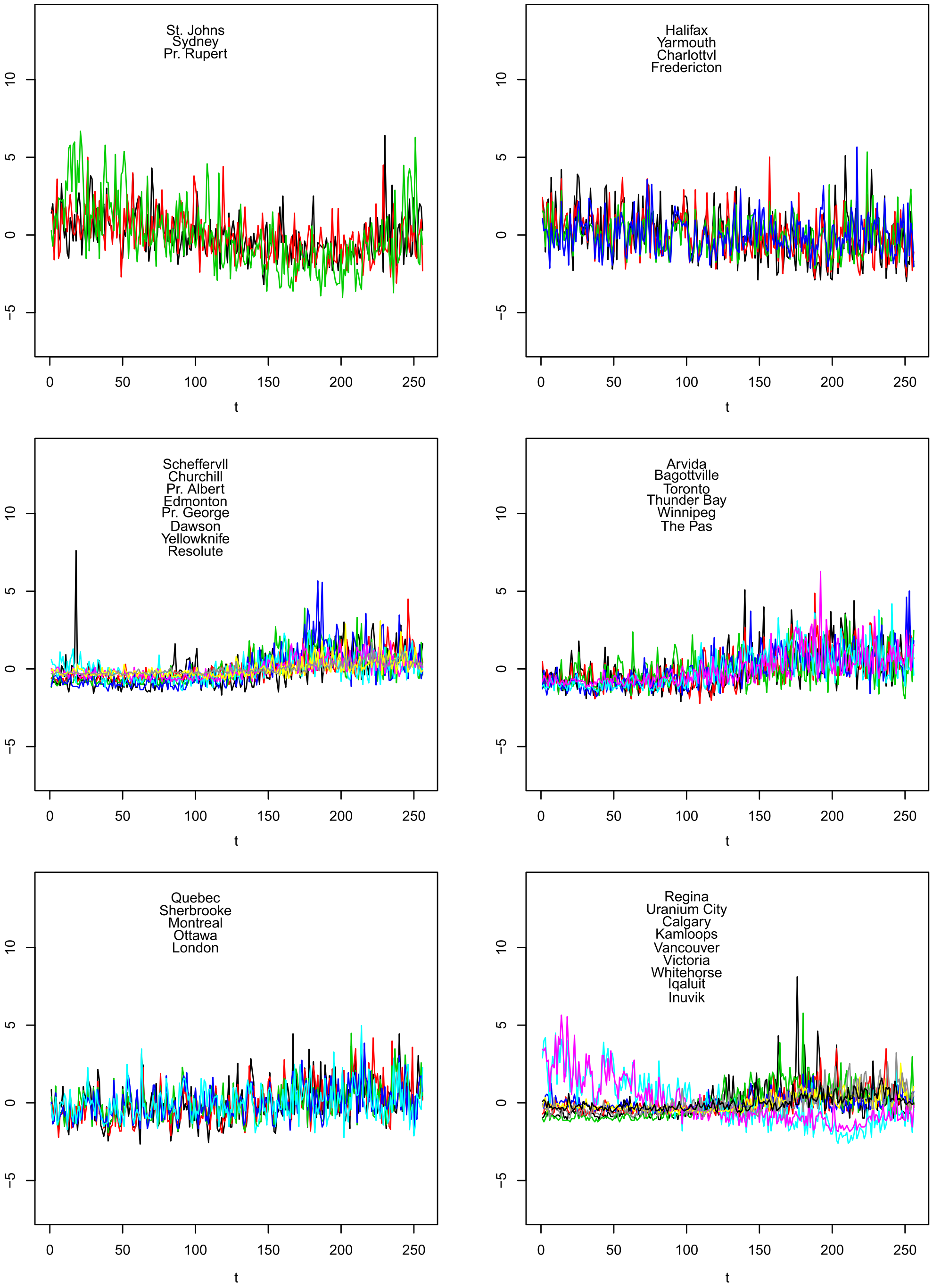

Figure 8: (Weather Data) Clustering formed by cutting at 6 groups for $M = 10$. For each group, the observed functional data are plotted on the same axes.

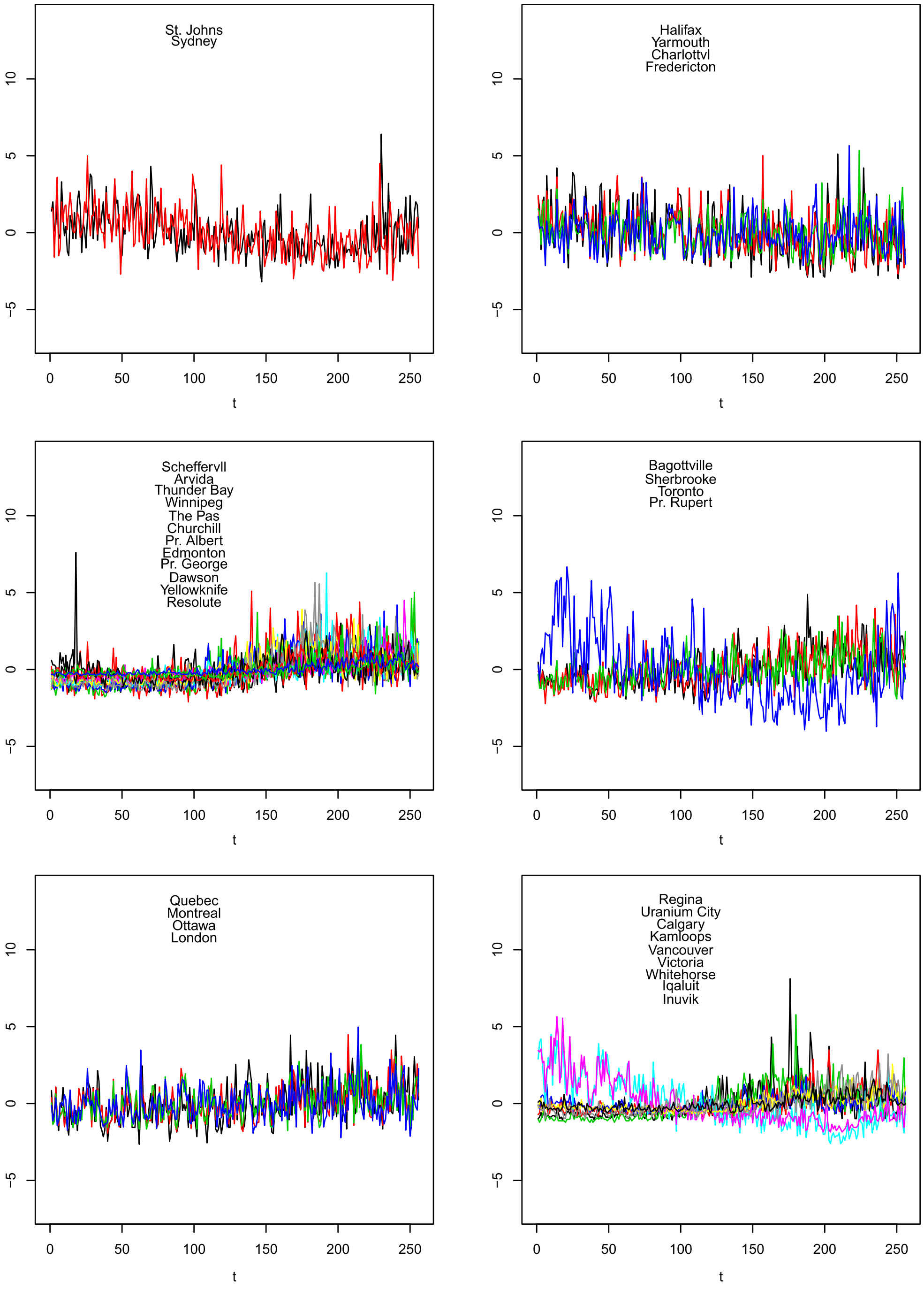

Figure 9: (Weather Data) Clustering formed by cutting at 6 groups for $M = 20$. For each group, the observed functional data are plotted on the same axes.

$(n+1)$th value is 0, given that m_0 out of the first n were zero, is

$$\frac{m_0}{M+n} + (1-\pi_j)\frac{M}{M+n}. \tag{17}$$

With $m_1 = n - m_0$, the probability that the next value is nonzero is given by $(m_1 + \pi_j M)/(M+n)$. If every level is considered separately, this probability coincides with that in a two-parameter Indian Buffet Process defined by Thibaux and Jordan (2007).

However, there is a difference. The Indian Buffet Process is defined for equivalence classes which correspond to rearranging elements of the binary string called the left order. Since it is used mostly in latent feature models, the elements have no inherent meaning, unlike our situation, in which each element corresponds to a particular wavelet coefficient.

The assumption that the functional data are observed on an equally spaced grid of size that is a power of 2 was made to use fast DWT techniques for computation. However, this restriction is not essential for the proposed method. In the case where time points are not equally spaced, or the total of points is not a power of 2, instead of first transforming the data using the DWT, the model can be written as a linear model in which the elements of the design matrix are the point evaluations of the wavelet basis functions at the common grid points. This approach could also be used to handle missing data. When used as a purely exploratory method, since the computation is much quicker in the case of using the DWT, it may be desirable to use a subset of the data rather than increase the computational complexity.

Another extension is to let the hyperparameter M also be given a prior. However, it is known that in Dirichlet mixture models the analysis can be quite sensitive to the choice of prior on M, and for exploratory purposes, running the model separately for various values of M could give a deeper understanding of the data than the results from a marginalized (over M) analysis. With a nonatomic base measure for the Dirichlet process, the standard augmentation procedure of Escobar and West (1995) can be extended to the case of conditionally independent Dirichlet processes with a common concentration parameter M. However, in the present case, because of the point mass at zero in the base measure, the conditional posterior distribution of M depends also on π_j, the size of the mass at 0 in the base measure, so a slight modification of the Escobar and West (1995) procedure will be needed. Alternatively, as the conditional posterior density of M is completely explicit except for its normalizing constant and M is only a one-dimensional parameter, standard sampling procedures can be applied.

To summarize, we have presented a model for functional data and priors which allow for the expression of beliefs related to common shared features (basis coefficients). We have proposed a measure of similarity distinct from the usual metrics, and have shown by theory and application that it yields useful results. In the model of Ray and Mallick (2006), the similarity between functions can be quantified by calculating the probability that two observations are in the same cluster. However, the approach based on separate modeling of the tying pattern of the wavelet coefficients appears to be more appealing.

Appendix A: Proofs

Proof of Proposition 1. First we show that the expected number of nonzero wavelet coefficients is finite a priori. Let $A_{jk} = \{\beta_{jk} \neq 0\}$, $k = 0, 1, \dots, 2^j - 1$ and $j = 0, 1, \dots$. Then, as $\delta \to 0$,

$$\begin{aligned}\sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} P(A_{jk}) &= \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \nu_2 \delta 2^{-(1+\delta)j} = \nu_2 \delta \sum_{j=0}^{\infty} 2^j 2^{-(1+\delta)j} \\ &= \nu_2 \delta \sum_{j=0}^{\infty} 2^{-j\delta} = \frac{\nu_2 \delta}{1 - 2^{-\delta}} < \infty,\end{aligned}$$

Thus, for any $\delta > 0$, by the Borel–Cantelli lemma, the number of nonzero wavelet coefficients is almost surely finite. Note that, as $\delta \to 0$, the expression, $\nu_2 \delta \left(1 - 2^{-\delta}\right)^{-1}$, converges to $\nu_2 (\log 2)^{-1}$ by L'Hôpital's rule.

Similarly, for the events $B_j = \cup_{k=0}^{2^j-1} A_{jk}$, using

$$P(B_j^c) = \prod_{k=0}^{2^j-1} P(A_{jk}^c) = (1 - \nu_2 \delta 2^{1-\delta} j)^{2^j},$$

we get that

$$\begin{aligned}\sum_{j=0}^{\infty} P(B_j) &= \sum_{j=0}^{\infty} \left\{1 - (1 - \nu_2 \delta 2^{-(1+\delta)j})^{2^j}\right\} \\ &\leq \sum_{j=0}^{\infty} 2^j \nu_2 \delta 2^{-j-\delta j} = \nu_2 \delta \sum_{j=0}^{\infty} 2^{-\delta j} < \infty \qquad (18)\end{aligned}$$

so that the number of levels with at least one nonzero coefficient is also almost surely finite.

In order to derive the Poisson limits, we apply Theorem 2 of Le Cam (1960). If $X_1, X_2, \dots, X_n$ are independent Bernoulli random variables with success probabilities $p_1, p_2, \dots$, respectively, then the total variation distance between the distribution of the sum, $Z_n = \sum_{j=1}^n X_j$ and a Poisson random variable is bounded by $\sum_{j=1}^n p_j^2$. Specifically, if Q_n is the measure on $\mathbb{N}$ induced by $\sum X_j$, and Q_n^* is the Poisson measure with rate $\lambda_n = \sum_{j=1}^n p_j$, with $Y_n \sim Q_n^*$. Let $\|Q_n - Q_n^*\| = \sup_{|f| \leq 1} |\mathrm{E}_{Q_n} f(Z_n) - \mathrm{E}_{Q_n^*} f(Y_n)|$ and $|f| = \sup_{x \in \mathbb{N}} |f(x)|$. Then $\|Q_n - Q_n^*\| \leq \sum_{j=1}^n p_j^2$. Therefore, it suffices to bound $\sum_{j,k} \{P(A_{jk})\}^2$ and $\sum_j \{P(B_j)\}^2$.

Now,

$$\sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \{P(A_{jk})\}^2 = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \nu_2^2 \delta^2 2^{-(1+\delta)2j} = \nu_2^2 \delta^2 \sum_{j=0}^{\infty} 2^{-(1+2\delta)j} = \frac{\nu_2^2 \delta^2}{1 - 2^{-(1+2\delta)}} \to 0,$$

as $\delta \to 0$.

Since the priors were specified independently across coefficients, the number of levels for which there is at least one nonzero coefficient (also the number of nonzero coefficients) follows the Poisson-binomial distribution, that is, the distribution of the sum of independent Bernoulli trials, but with varying parameters. Consider the sum of the squared success probabilities

$$\begin{aligned}\sum_{j=0}^{\infty}\{P(B_j)\}^2 &= \sum_{j=0}^{\infty}\left\{1-(1-\nu_2\delta 2^{-(1+\delta)j})^{2^j}\right\}^2 \\ &\leq \sum_{j=0}^{\infty}\left\{2^j\nu_2\delta 2^{-j-\delta j}\right\}^2 = \nu_2^2\sum_{j=0}^{\infty}\delta^2 2^{-2j\delta} \\ &= \nu_2^2\frac{\delta^2}{2^{2\delta}-1}. \qquad (19)\end{aligned}$$

Notice that both the numerator and denominator of (19) converge to 0 as $\delta \to 0$, so by L'Hôpital's rule the limit of the expression in (19) is equal to the limit of $\nu_2^2 2\delta\left(4^\delta \log 4\right)^{-1}$, which is 0. □

Proof of Theorem 1. Since the true vector of functions lies in $\mathcal{H}_N^s$, it is in a ball of radius B for sufficiently large $B > 0$. Let $\epsilon > 0$ and let J be the smallest integer satisfying both

$$\sum_{i=1}^{N}\sum_{j>J}^{\infty}\sum_{k=0}^{2^j-1}|\beta_{jk,0}^{(i)}|^2 < \epsilon^2/8. \qquad (20a)$$

Notice that J exists and is finite since $\|\boldsymbol{f}_0\| < \infty$. With this chosen we have, using $(a+b)^2 \leq 2(a^2+b^2)$, that

$$\begin{aligned}\Pi\left(\sum_{\substack{j>J\\ i,k}}|\beta_{jk}^{(i)}-\beta_{jk,0}^{(i)}|^2<\epsilon^2/2\right) &\geq \Pi\left(\sum_{j>J,i,k}|\beta_{jk}^{(i)}|^2<\epsilon^2/8\right) \\ &\geq 1-\frac{8}{\epsilon^2}\sum_{j>J,i,k}\mathrm{E}_\Pi\left(|\beta_{jk}^{(i)}|^2\right) \\ &= 1-\frac{8}{\epsilon^2}N\nu_1\nu_2\frac{2^{(1-\gamma_2-\gamma_1)J}}{2^{1-\gamma_2-\gamma_1}-1}>0\end{aligned}$$

since $\gamma_1,\gamma_2 > 1$.

Next, notice that

$$\Pi\left(\sum_{j\leq J,i,k}|\beta_{jk}^{(i)}-\beta_{jk,0}^{(i)}|^2<\epsilon^2/2\right) \geq \eta\prod_{i=1}^{N}\Pi\left(\sum_{j\leq J,k}|\beta_{jk}^{(i)}-\beta_{jk,0}^{(i)}|^2<\epsilon^2/2N\right),$$

where $\eta > 0$ is a constant representing the probability that for each $k = 0, \ldots, 2^j - 1$, $j = 0, \ldots, J - 1$, the prior makes all $\beta_{jk}^{(i)}$, $i = 1, \ldots, N$, distinct. Then,

$$\eta \geq \pi^{2^J - 1} \prod_{j \leq J,k} \prod_{i=1}^{N-1} \frac{M}{M+i}.$$

This expression represents the probability that a new unique value must be drawn from the base measure and also must not be assigned to 0, which ensures unique values (it is a lower bound since it does not include the probability of having a single zero value). Now by the positivity of the prior on each $\beta_{jk}^{(i)}$, we have

$$\eta \prod_{i=1}^{N} \Pi \left(\sum_{j \leq J,k} |\beta_{jk}^{(i)} - \beta_{jk,0}^{(i)}|^2 < \epsilon^2/2N \right) > 0.$$

A quantitative estimate of the probability is given in Lian (2011, Theorem 1), but we do not need that here. Combining these results we have that for $\epsilon > 0$,

$$\Pi \left(\sum_{j,i,k} |\beta_{jk}^{(i)} - \beta_{jk,0}^{(i)}|^2 < \epsilon^2 \right) > 0,$$

which shows the positivity of any Kullback–Leibler neighborhood since we have a Gaussian likelihood and the Kullback–Leibler divergence between two normal distributions, $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, is given by $(\mu_1 - \mu_2)^2/2\sigma^2$. Similarly, the Hellinger distance is equivalent to the L_2-distance. Together with the fact that $\mathcal{H}_N^s(B)$ has finite metric entropy by Belitser and Ghosal (2003), we have that, using Theorem 2 from Ghosal et al. (1999), for any given $B > 0$,

$$\Pi \left(\boldsymbol{f} \in \mathcal{H}_N^s(B) : \|\boldsymbol{f} - \boldsymbol{f}_0\| > \epsilon | D_r \right) \to 0 \text{ in probability.}$$

Now, it suffices to show that $\lim_{B \to \infty} \sup_{r > 0} E_0 \Pi(\mathcal{H}_N^s(B)^c | D_r) = 0$. By Markov's inequality and the monotone convergence theorem,

$$\Pi(\mathcal{H}_N^s(B)^c | D_r) \leq B^{-2} \sum_{i=1}^{N} \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2^j - 1} \mathrm{E}\left[|\beta_{jk}^{(i)}|^2 \middle| D_r \right]. \tag{21}$$

We bound the needed expectations as

$$\begin{aligned} \mathrm{E}\left[|\beta_{jk}^{(i)}|^2 \middle| D_r \right] &= \sum_{\mathfrak{p} \in \mathfrak{P}} \mathrm{E}\left[|\beta_{jk}^{(i)}|^2 \middle| \mathcal{P}_{jk} = \mathfrak{p}, D_r \right] \Pi(\mathcal{P}_{jk} = \mathfrak{p} | D_r) \\ &\leq \max_{\mathfrak{p} \in \mathfrak{P}} \mathrm{E}\left[|\beta_{jk}^{(i)}|^2 \middle| \mathcal{P}_{jk} = \mathfrak{p}, D_r \right]. \end{aligned}$$

Now, given $\mathcal{P}_{jk} = \mathfrak{p}$, $i \in A_l^{jk}$ for some $l \in \{0, 1, \dots, M_{jk}\}$ (for definition, see (14)), and the posterior is of conjugate form, so that the expectation is

$$\begin{aligned} \mathrm{E}\left[|\beta_{jk}^{(i)}|^2 \middle| \mathcal{P}_{jk} = \mathfrak{p}, D_r \right] &= \frac{\tau_j^2 \sigma^2}{1 + \tau_j^2 r (\# A_l)} + \left(\frac{\tau_j^2}{1/r + \tau_j^2 (\# A_l)} \right)^2 N \sum_{m \in A_l} \big(b_{jk}^{(m)}\big)^2 . \\ &\leq \frac{\tau_j^2 \sigma^2}{1 + \tau_j^2 r} + \left(\frac{\tau_j^2}{1/r + \tau_j^2} \right)^2 N \sum_{m=1}^{N} \big(b_{jk}^{(m)}\big)^2 , \end{aligned}$$

where $\# A$ is the cardinality of the set A. We now bound the true expectation of this term by

$$\begin{aligned} \tau_j^2 \sigma^2 + \left(\frac{\tau_j^2}{1/r + \tau_j^2} \right)^2 \left[\frac{N \sigma^2}{r} + N \sum_{m=1}^{N} \big(\beta_{jk,0}^{(m)}\big)^2 \right] &\leq \tau_j^2 \sigma^2 + \frac{r N \sigma^2}{\big(\tau_j^{-2} + r\big)^2} + N \sum_{m=1}^{N} |\beta_{jk,0}^{(m)}|^2 \\ &\leq \tau_j^2 \sigma^2 + \frac{N \sigma^2}{\big(\tau_j^{-2} + 1\big)^2} + N \sum_{m=1}^{N} |\beta_{jk,0}^{(m)}|^2 . \end{aligned}$$

Now observe that under the assumption that $\gamma_1 > 2s + 1$, we have

$$\sum_{i=1}^{N} \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2^j - 1} \tau_j^2 = N \nu_1 \sum_{j=0}^{\infty} 2^{j(2s+1-\gamma_1)} < \infty,$$

$$\sum_{i=1}^{N} \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2^j - 1} \frac{N \sigma^2}{\big(\tau_j^{-2} + 1\big)^2} = N \sum_{j=0}^{\infty} \frac{N \sigma^2 2^{j(2s+1)}}{\big(\nu_1^{-1} 2^{\gamma_1 j} + 1\big)^2} \leq N^2 \sigma^2 \nu_1^2 \sum_{j=0}^{\infty} 2^{j(2s+1-2\gamma_1)} < \infty,$$

$$\sum_{i=1}^{N} \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2^j - 1} \sum_{m=1}^{N} |\beta_{jk,0}^{(m)}|^2 = N \| \boldsymbol{f}_0 \|_{\mathcal{H}_N^s}^2 < \infty.$$

Hence, the right hand side of (21) tends to zero as $r \to \infty$, completing the proof. □

Proof of Lemma 1. Theorem 1 implies that for any neighborhood of the true value, $\boldsymbol{\beta}_{jk,0}$, its posterior probability tends to 1. By projection onto the coordinates, the posterior probability of $\beta_{jk}^{(i)}$ lying in any neighborhood of $\beta_{jk,0}^{(i)}$ also tends to 1.

Let $\mathcal{P}_{jk} = \mathfrak{p}$ be an incompatible model, as defined in Section 6. Then there is an incorrect assignment for some i, that is, either $\beta_{jk}^{(i)} = 0$ when $\beta_{jk,0}^{(i)} \neq 0$, or $\beta_{jk}^{(i)} = \beta_{jk}^{(l)}$ for some l when this is not the case. For instance, consider the situation that the partition assigns $\beta_{jk}^{(i)} = 0$ when $\beta_{jk,0}^{(i)} \neq 0$. The inequality implies that there is a positive distance between the two points so that there is a neighborhood of $\beta_{jk,0}^{(i)}$ which does not contain 0. The probability of the complement of this neighborhood must tend to 0 as $r \to \infty$, and since the posterior probability of $\mathcal{P}_{jk} = \mathfrak{p}$ must be less than or equal to this value, its posterior probability must also tend to 0.

Thus, we can focus our attention only on compatible models. Now let $\mathcal{P}_{jk} = \mathfrak{p} = \{A_0, A_1, \ldots, A_m\}$ be a compatible model (we drop the j, k notation for the sets in the partition for this section). Letting H be the distribution function for the inverse-gamma prior, we then have that

$$\begin{aligned} f(\boldsymbol{b}_{jk}|\mathcal{P}_{jk} = \mathfrak{p}) &= \left(\prod_{l\in A_0}\phi_\sigma(b_i)\right)\prod_{i=1}^{m}\left[\int\left(\prod_{l\in A_i}\phi_\sigma(b_i-\beta)\right)\phi_{\tau_j}(\beta)d\beta\right] \\ &= \left(\frac{2\pi}{r}\right)^{-N/2}(\tau_j^2)^{-m/2}\left[\prod_{i=1}^{m}\left(rc_i+\frac{1}{\tau_j^2}\right)^{-1/2}\right](\sigma^2)^{\left(a+\frac{N}{2}-1\right)} \\ &\quad\times\exp\left\{-\frac{1}{\sigma^2}\left[b+\frac{r}{2}\sum_{i=0}^{N}b_i^2-\frac{r^2}{2}\sum_{i=1}^{m}\left(rc_i+\frac{1}{\tau_j^2}\right)^{-1}\left(\sum_{l\in A_i}b_l\right)^2\right]\right\}, \end{aligned}$$

where we have dropped the j, k subscripts, and let $c_l = \#A_l$ and $b_l = b_{jk}^{(l)}$.

Now let $\mathcal{P}_{jk,0} = \mathfrak{p}_{jk,0} = \{A_0^0, A_1^0, \ldots, A_{m_0}^0\}$ be the true model. Since $\mathcal{P}_{jk} = \mathfrak{p}$ is compatible (finer), assume that $m > m_0$, and that $A_i \subset A_i^0$ for $i = 0, \ldots, m_0$. Now,

$$\begin{aligned} &\frac{f(\boldsymbol{b}_{jk}|\mathcal{P}_{jk} = \mathfrak{p})}{f(\boldsymbol{b}_{jk}|\mathcal{P}_{jk} = \mathfrak{p}_0)} \\ &= (\tau_j^2)^{(m_0-m)/2}\sqrt{\frac{\prod_{i=1}^{m_0}\left(rc_i^0+\frac{1}{\tau_j^2}\right)}{\prod_{i=1}^{m}\left(rc_i+\frac{1}{\tau_j^2}\right)}} \\ &\times\exp\left\{\frac{r}{\sigma^2}\left[\sum_{i=1}^{m}\left(\frac{\tau_j^2}{1/r+c_i\tau_j^2}\right)\left(\sum_{l\in A_i}b_l\right)^2-\sum_{i=1}^{m_0}\left(\frac{\tau_j^2}{1/r+c_i^0\tau_j^2}\right)\left(\sum_{l\in A_i^0}b_l\right)^2\right]\right\}. \end{aligned}$$

To examine the behavior of this expression note that $m > m_0$, and $c_i < c_i^0$ for $i = 1, \ldots, m_0$. Now, the expression under the square root converges to 0 as $r \to \infty$, since $m > m_0$ implies that the denominator is of a higher order.

Consider the term inside the square brackets of the exponential:

$$\sum_{i=1}^{m}\left(\frac{\tau_j^2}{1/r+c_i\tau_j^2}\right)\left(\sum_{l\in A_i}b_l\right)^2-\sum_{i=1}^{m_0}\left(\frac{\tau_j^2}{1/r+c_i^0\tau_j^2}\right)\left(\sum_{l\in A_i^0}b_l\right)^2.$$

It is clear that this expression can be written as a quadratic form, $\boldsymbol{b}^t\boldsymbol{A}\boldsymbol{b}$, in $\boldsymbol{b}$, for some matrix, $\boldsymbol{A}$. It is clear that $\boldsymbol{A}$ is $O_r(1)$ as $r \to \infty$. The expectation is then $O_r(1)$, and, since the dispersion matrix of $\boldsymbol{b}$ is $\frac{\sigma^2}{r}\boldsymbol{I}_N$, the variance of $\boldsymbol{b}^t\boldsymbol{A}\boldsymbol{b}$ is $O_r(1/r)$ as $r \to \infty$. Thus, the exponential term is bounded in probability as $r \to \infty$.

Thus, we have shown that the probability of any incompatible model goes to zero. Along with the fact that for any compatible model the marginal likelihood ratio tends to 0 implies that the only model that can possibly retain positive probability is the truth. Since, for any fixed j, k, there are only finitely many models for $\{\beta_{jk}^{(i)}\}_{i=1}^{N}$, the probability of the true model must tend to 1. $\square$

References

Abramovich, F., Sapatinas, T., and Silverman, B. (1998). "Wavelet thresholding via a Bayesian approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4): 725–749. MR1649547. doi: http://dx.doi.org/10.1111/1467-9868.00151. 75, 77

Andrzejak, R., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C., et al. (2001). "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state." *Physical Review Series E*, 64(6; PART 1): 61907–61907. 81, 83

Belitser, E. and Ghosal, S. (2003). "Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution." *The Annals of Statistics*, 31(2): 536–559. MR1983541. doi: http://dx.doi.org/10.1214/aos/1051027880. 94

Blackwell, D. and MacQueen, J. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 353–355. MR0362614. 76

Cohen, A., Daubechies, I., and Vial, P. (1993). "Wavelets on the interval and fast wavelet transforms." *Applied and Computational Harmonic Analysis*, 1(1): 54–81. MR1256527. doi: http://dx.doi.org/10.1006/acha.1993.1005. 73

Crandell, J. L. and Dunson, D. B. (2011). "Posterior simulation across nonparametric models for functional clustering." *Sankhya B*, 73(1): 42–61. MR2826319. doi: http://dx.doi.org/10.1007/s13571-011-0014-z. 72

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. doi: http://dx.doi.org/10.1080/01621459.1995.10476550. 91

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., and Rossi, F. (2009). *GNU Scientific Library Reference Manual (3rd Ed.)*, ISBN 0954612078. URL: http://www.gnu.org/software/gsl/. 81

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27(1): 143–158. MR1701105. doi: http://dx.doi.org/10.1214/aos/1018031105. 94

James, G. and Sugar, C. (2003). "Clustering for sparsely sampled functional data." *Journal of the American Statistical Association*, 98(462): 397–408. MR1995716. doi: http://dx.doi.org/10.1198/016214503000189. 72

Le Cam, L. (1960). "An approximation theorem for the Poisson binomial distribution." *Pacific Journal of Mathematics*, 10(4): 1181–1197. MR0142174. 92

Lian, H. (2011). "On posterior distribution of Bayesian wavelet thresholding." *Journal of Statistical Planning and Inference*, 141(1): 318–324. MR2719497. doi: http://dx.doi.org/10.1016/j.jspi.2010.06.016. 79, 94

Liao, T. (2005). "Clustering of time series data – a survey." *Pattern Recognition*, 38(11): 1857–1874. 72

Navarrete, C., Quintana, F., and Müller, P. (2008). "Some issues in nonparametric Bayesian modeling using species sampling models." *Statistical Modelling*, 8(1): 3–21. MR2750628. doi: http://dx.doi.org/10.1177/1471082X0700800102. 79

Petrone, S., Guindani, M., and Gelfand, A. (2009). "Hybrid Dirichlet mixture models for functional data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4): 755–782. MR2750094. doi: http://dx.doi.org/10.1111/j.1467-9868.2009.00708.x. 72

Pitman, J. (1996). "Some developments of the Blackwell–MacQueen urn scheme." *Lecture Notes-Monograph Series*, 245–267. MR1481784. doi: http://dx.doi.org/10.1214/lnms/1215453576. 85

Ray, S. and Mallick, B. (2006). "Functional clustering by Bayesian wavelet methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2): 305–332. MR2188987. doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00545.x. 72, 75, 81, 91

Serban, N. (2008). "Estimating and clustering curves in the presence of heteroscedastic errors." *Journal of Nonparametric Statistics*, 20(7): 553–571. MR2454615. doi: http://dx.doi.org/10.1080/10485250802348742. 72

Tarpey, T. and Kinateder, K. (2003). "Clustering functional data." *Journal of Classification*, 20(1): 93–114. MR1983123. doi: http://dx.doi.org/10.1007/s00357-003-0007-3. 71

Thibaux, R. and Jordan, M. (2007). "Hierarchical beta processes and the Indian buffet process." In: *International Conference on Artificial Intelligence and Statistics*, volume 11, 564–571. 91

Ward Jr., J. (1963). "Hierarchical grouping to optimize an objective function." *Journal of the American Statistical Association*, 236–244. MR0148188. 81

Acknowledgments

Research is partially supported by NSF grant number DMS-1106570.