Bayesian nonparametric modeling for functional analysis of variance

XuanLong Nguyen and Alan E. Gelfand [1]

Abstract

Analysis of variance is a standard statistical modeling approach for comparing populations. The functional analysis setting envisions that mean functions are associated with the populations, customarily modeled using basis representations, and seeks to compare them. Here, we adopt the modeling approach of functions as realizations of stochastic processes. We extend the Gaussian process version to allow nonparametric specifications using Dirichlet process mixing. Several metrics are introduced for comparison of populations. Then we introduce a hierarchical Dirichlet process model which enables comparison of the population *distributions*, either directly or through functionals of interest using the foregoing metrics. The modeling is extended to allow us to switch the sampling scheme. There are still population level distributions but now we sample at levels of the functions, obtaining observations from potentially different individuals at different levels. We illustrate with both simulated data and a dataset of temperature vs depth measurements at different locations in the Atlantic Ocean.

Key words: Dirichlet processes, Gaussian processes, global and local clustering, hierarchical models, random distributions

1 Introduction

In this paper we consider response models where the responses are functions indexed by groups, with the goal to learn if the functions differ across groups and, if so, how they differ. It is natural to refer to this setting as a functional analysis of variance (ANOVA) problem, recognizing the challenges in comparing surfaces (uncountable dimensional response) across populations rather than scalars (usual ANOVA) or vectors (MANOVA).

[1]XuanLong Nguyen (xuanlong@umich.edu) is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107. Alan E. Gelfand (alan@stat.duke.edu) is Professor, Department of Statistical Science, Duke University, Durham, NC 27708-0251. This work was partially supported by NSF grants No. 0940671 and No. 1047871 (XN).

Applications on R^1 typically have time as the argument, for example progesterone levels for groups of women (MacLehose and Dunson, 2009; Nguyen and Gelfand, 2011), mass spectroscopy data for different groups over time (Morris and Carroll, 2006), dose response (white blood cell counts) indexed by cancer treatments (DeIorio et al., 2004), and temperature profiles indexed by climate model (Kaufman and Sain, 2010). As a different example, Rappold et al. (2007) and Rodriguez et al. (2009) examine temperature vs. depth over different regions in the Atlantic Ocean. On R^2, we find investigation of brain images (Petrone et al., 2009) and, more generally, image analysis (Nguyen and Gelfand, 2011).

The contribution of this paper is to formulate the functional ANOVA problem in a fully Bayesian nonparametric framework using suitable hierarchical modeling. In particular, we begin with the Gaussian process (GP), then extend to the spatial Dirichlet process (SDP) (Gelfand et al., 2005). Then, we introduce a novel hierarchical and nested Dirichlet process (HDP) specification, which models the (random) distributions which generate the functions, by adopting and extending the hierarchical modeling of Teh et al. (2006) and Nguyen (2010). The novelty in our modeling framework is in the ability to switch the sampling scheme. This is particularly applicable to the setting for functional ANOVA which may require sampling at both functional level and the levels of functions. In our modeling for functional ANOVA, we still have random population-level distributions but now we sample at levels of the function, obtaining observations from potentially different individuals at different levels. We discuss metrics for comparing populations which are applicable under any of these modeling specifications.

Notably, we work in the setting where we do not have a large number of observations of the functions over the domain of the argument. Hence, we do not seek to learn about the functions at fine detail, at high resolution of the argument. For the latter setting, it might be advantageous to use special basis representations such as wavelets (Morris and Carroll, 2006). Rather, we seek to interpolate the functions over their domain, not an activity of the high resolution work. Hence we are drawn to GP's and processes that extend GP's. Though we may not have many observations of the function, we do not find a MANOVA model to be appropriate. For MANOVA, the components of the vector need not be the same measurement variable. So, general covariance matrices are introduced and these matrices are partitioned to obtain *variance* components. For us, the components are measurements all on the same variable and we introduce "structured" dependence

between them.

The ANOVA setting presumes that the individual-level functions are "pre-clustered", i.e., they are already indexed by a population label. So, unlike usual Dirichlet process settings, we are not primarily seeking to determine clusters that create groups of functions. (Such adaptive clustering or mixture modeling is often the reason for adopting DP specifications.) In fact, we are proposing to use the DP structure primarily to compare the groups. Within a GP framework, we can not talk about groups being the *same* since this happens with probability 0; instead, we employ metrics to measure closeness; we develop such ideas below. If we move to SDP's, we can have ties. That is, for a pair of groups, now the curves are either identical everywhere or nowhere. Finally, when we work with the HDP and the nested HDP we move to a comparison of the distributions that generate population-level features. Now, ties are possible (only global with the HDP, local with the nested HDP) for realized curves across populations. In the spirit of customary ANOVA hypothesis testing, priors that allow such *ties* are natural for this setting; they capture the same vein as familiar "spike and slab" priors for variable selection (see, e.g., Ishwaran and Sunil-Rao (2005) and references therein) which allow parameters (or differences in parameters) to have positive prior probability (hence positive posterior probability) of being 0.

As in usual ANOVA settings, replications are required; in order to assess differences between populations, we need to learn about the variability within populations. Similarly, in the functional ANOVA setting, we are not seeking to cluster individuals within populations. Rather, we are seeking to learn about the variability of individual observations, for us, individual curves, within a population, again, to facilitate comparison of curves across populations.

Modern nonparametric ANOVA moves away from Gaussian error assumptions, adopting population models that allow skewness, heavier tails, and multimodalities. It also considers comparing other functionals, such as quantiles, across populations. We are in this contemporary camp but in the setting of curves rather than scalars. In fact, in our HDP and nested HDP versions, we compare distributions across populations where such comparison can be done based upon local functionals (i.e., at the arguments of the curves) yielding global functions.

The field of functional data analysis has benefitted from the seminal books of Ramsay and Silverman (2006) (cf. Chapter 13) and Ferraty and Vieu (2006). This work proceeds through the use of orthonormal basis representations for functions, typically spline bases. As noted above,

usually the functions of interest are over space and/or time and the literature is substantial. Notable alternative applications include (Brumback and Rice, 1998), (Spitzner et al., 2003), Wang et al. (2005). These basis representations provide explicit forms for the functions, i.e., finite dimensional parametric representations of the function.

Our approach is to view the entire function as unknown and to view it as a realization of a stochastic process. In this regard, Gaussian processes are a customary place to begin (Cressie, 1993; Banerjee et al., 2004). Since we work within the Bayesian framework we use such processes and extensions of them as priors for the functions we model and use the available data to update to posterior estimates of the functions. By introducing nonparametric specifications, we move beyond the work of Kaufman and Sain (2010). They confine themselves to the use of GP's in their Bayesian functional ANOVA formulation and imitate classical ANOVA modeling by incorporating constraints on the functions in order to identify them. They introduce pointwise and global credible intervals for comparison of curves, employing deviations relative to an appropriate "average" curve. Our DP-based framework yields a much different model construction, resulting in a different approach for comparison. The recent book chapter of Dunson (2010) (Sec. 7.3) provides a review of various Bayesian nonparametric approaches to the modeling of functional data.

As noted above, we build our modeling in a sequential fashion and, as a by-product, offer comparison between the GP and DP extensions of the GP. We employ simulated data as a proof of concept, to demonstrate the benefits of our more flexible modeling. We also analyze a real dataset which considers the temperature vs. depth relationship for four different regions in the Atlantic Ocean.

The plan for the paper is as follows. In Section 2 we briefly review the Gaussian and Spatial Dirichlet processes we will use to model realizations of functions. In Section 3 we move these models to our functional ANOVA setting, discussing summaries of individual functions and comparison of functions. Section 4 proposes a new functional ANOVA model based upon hierarchical Dirichlet processes. Section 5 takes up the simulated and real examples while Section 6 closes with a summary and future investigations.

2 Stochastic process models for random functions

As noted in the Introduction, we model our unknown functions as realizations of stochastic processes. Gaussian processes are convenient to work with in this regard since consistent specification of finite dimensional distributions for GP's only requires specification of a mean function and a valid covariance function. Formally, we will write that $\theta(x)$ follows a GP over the set $x \in D$ and specify $\mathbb{E}(\theta(x)) = \mu_\theta(x)$ and $cov(\theta(x), \theta(x')) = C(x, x')$ where C is valid over D. Here we confine ourselves to stationary forms and write $C(x, x')$ as $\sigma_C^2 \rho(x - x'; \phi_C)$ where ρ is a valid stationary correlation function. For example, an exponential covariance function takes the form $C(x, x') = \sigma_C^2 \exp\{\|x - x'\|^2/\phi_C\}$.

Next, we turn to the spatial Dirichlet process, introduced by Gelfand et al. (2005). We first recall the Dirichlet process (Ferguson, 1973) which provides a random probability measure on spaces of distribution functions. A constructive definition was introduced by Sethuraman (1994). In the univariate case, let $\{\omega_k, k = 1, 2, ...\}$ and $\{\boldsymbol{\phi}_k, k = 1, 2, ...\}$ be independent sequences of i.i.d. random variables. Let $\omega_k \sim$ Beta$(1, \gamma)$, γ a positive precision parameter and $\boldsymbol{\phi}_k \sim H$, H a parametric *base* distribution. Define $\beta_1 = \omega_1$, $\beta_k = \omega_k \prod_{j=1}^{k-1}(1 - \omega_j)$, $k = 2, 3, ...$ Notationally, we will write that $\boldsymbol{\beta} \sim$ GEM(γ). Then, a realization from DP(γ, H) is almost surely of the form $\sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\phi}_k}$. We note that we may also specify that $k = 1, 2, ..., K$ where $K < \infty$, referred to as the finite DP or DP$_K$, and the weights are drawn from a K-dimensional Dirichlet distribution (Ishwaran and Zarepour, 2002).

We can immediately extend this definition to accommodate a realization of a spatial random field. Replace $\boldsymbol{\phi}_k$ with $\boldsymbol{\phi}_{k,D} = \{\phi_k(x) : x \in D\}$. Here, H can be a stationary Gaussian random field and each $\boldsymbol{\phi}_{k,D}$ is a realization from G_0, i.e., a random *surface* over D. Hence, we create a random process over D of the form $G = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\phi}_{k,D}}$, centered at the process H and write $G \sim$ DP(γ, H). G describes a stochastic process of random distributions and, since they were working in the spatial setting with $D \subseteq R^2$, Gelfand et al (2005) called this class of processes spatial or SDP's. We will use this terminology as well though, for us, x need not index geographic space. We can directly verify that the set, G_D, as a collection of random measures is a dependent Dirichlet process (DDP) (MacEachern, 1999). Furthermore, if H produces a.s. continuous realizations then the a.s. representation of G_D ensures that $G(\theta(x)) - G(\theta(x')) \to 0$ a.s. as $||x - x\prime|| \to 0$. In other words, smoothness of realizations from $DP(\gamma, H)$ is determined by the choice of the covariance function

of H. Conditions for almost sure or mean square continuity are discussed in Kent (1989) and Stein (1999), respectively.

For $\boldsymbol{\theta}_D = \{\theta(x) : x \in D\}$ a realization from G, it is straightforward to verify that $\mathbb{E}(\theta(x) \mid G) = \sum \beta_k \boldsymbol{\phi}_k(x)$ and $cov(\theta(x), \theta(x') \mid G) = \sum \beta_k \boldsymbol{\phi}_k(x)\boldsymbol{\phi}_k(x') - \{\sum \beta_k \boldsymbol{\phi}_k(x)\}\{\sum \beta_k \boldsymbol{\phi}_k(x')\}$. We smooth out the point masses of G by mixing against a white noise process $\mathcal{K}$ (with mean 0 and variance τ^2) resulting in a random process over D with continuous support. Operating formally, if $\boldsymbol{\theta}_D \mid G \sim G$ and $\boldsymbol{Y}_D - \boldsymbol{\theta}_D \mid \tau^2 \sim \kappa$, κ a density then $f(\boldsymbol{Y}_D \mid G, \tau^2) = \int \kappa(\boldsymbol{Y}_D - \boldsymbol{\theta}_D \mid \tau^2)\, G(d\boldsymbol{\theta}_D)$. Hence, ignoring the mean, $Y(x) = \theta(x) + \epsilon(x)$ where $\theta(x)$ is from the SDP and $\epsilon(x)$ is white noise.

For the finite set of levels $x_1, ..., x_n$, the induced mixture model becomes

$$f\left(\boldsymbol{Y} \mid G^{(n)}, \tau^2\right) = \int f_{N_n}\left(\boldsymbol{Y} \mid \boldsymbol{\theta}, \tau^2 I_n\right) G^{(n)}(d\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{Y} = (Y(x_1), ..., Y(x_n))'$ and $\boldsymbol{\theta} = \theta^{(n)} = (\theta(x_1), ..., \theta(x_n))'$ yields $f\left(\boldsymbol{Y} | G^{(n)}, \tau^2\right)$ a.s. of the form $\sum_{k=1}^{\infty} \beta_k f_{N_n}(\boldsymbol{y} | \phi_k, \tau^2 I_n)$, a countable location mixture of normals. Given $G^{(n)}$ and τ^2, the resulting covariance matrix becomes $C_{\boldsymbol{Y}} = \tau^2 I_n + C_{\boldsymbol{\theta}}$ with $(C_{\boldsymbol{\theta}})_{i,j} = Cov(\theta(x_i), \theta(x_j) \mid G^{(n)})$.

As evident from the representation of G, the SDP provides a nonstationary, nonGaussian process. From above, given G, two random curves $\theta_1(x)$ and $\theta_2(x)$ agree a.e. with probability $\sum \beta_k^2$ or else they disagree a.e. In the context of functional ANOVA, this allows ties between the population functions.

3 Functional ANOVA using Gaussian processes and spatial Dirichlet processes

We now return to the functional ANOVA problem. We focus on the one-way layout setting, initially specified as

$$Y_{ui}(x) = \theta_u(x) + \epsilon_{ui}(x), \tag{2}$$

for $i = 1, \ldots, n_u$. Here, $u = 1, 2, ..., U$ indexes the *populations/treatments* and i the individuals within the populations. $\boldsymbol{\theta}_u$ denotes the function/surface for population u. Curves for individuals from population u are assumed to be conditionally independent given $\boldsymbol{\theta}_u$, i.e., the $\boldsymbol{\epsilon}_{ui}$ are independent. In fact, for convenience, in the sequel we assume that $\boldsymbol{\epsilon}$ is a white noise process. This implies

that all individual curves are almost surely discontinuous even if the $\boldsymbol{\theta}_u$ are continuous. Our choice here is for simplicity of exposition; in some situations $\boldsymbol{\epsilon}_{ui}$ may be more suitably modeled as GP realizations (e.g., Kaufman and Sain (2010)). Here, individual error ϵ_{ui} is assumed to be a white noise process, i.e., $\epsilon_{ui}(x) \sim N(0, \tau_u^2)$ i.i.d. for $i = 1, \ldots, n_u$.

As in usual ANOVA modeling, we assume that the data from all groups have been *re-centered* around a mean curve $\boldsymbol{\mu}$. So, the $\boldsymbol{\theta}_u$ are deviation curves and, in comparing them, it is the functional variation around $\boldsymbol{\mu}$ that we are interested in. Accordingly, we endow the $\boldsymbol{\theta}_u$ with a prior distribution with a mean curve $\boldsymbol{\mu}$, which may be again endowed with a prior distribution; interest is in the differences between the $\boldsymbol{\theta}_u$'s.

For each population, we are interested in the variation, $var\theta_u(x)$, and correlation, i.e., for levels x_1 and x_2, $corr(\theta_u(x_1), \theta_u(x_2))$, respectively. Additionally, we are interested in summaries of the curves obtained by integration over a given sub-region $B \subseteq D$ of interest:

$$m_1(\boldsymbol{\theta}_u, B) = \int_B \theta_u(x)^2 dx,$$
$$m_2(\boldsymbol{\theta}_u, B) = \int_B \mathbb{I}(\theta_u(x) \geq 0) dx,$$
$$m_3(\boldsymbol{\theta}_u, B) = \int_B (\theta_u(x))_+ dx,$$

In the sequel we suppress B which will often be D. Based on these summaries, to compare $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$, the following measures are considered. Let

$$d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) = m_1(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v);\ d_2(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) = m_2(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v);\ \text{and } d_3(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) = m_3(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v).$$

These "metrics" form the basis for our ANOVA comparisons of functional data. Using the non-parametric curve specifications from the previous section as priors, allows us to specify the prior probability that populations u and v are "the same', as well as to elaborate the nature of their differences using the above metrics. We then use the data to make these comparisons a posteriori.

3.1 Functional ANOVA based on Gaussian processes

Suppose that a priori $\boldsymbol{\theta}_u \sim \text{GP}(\boldsymbol{\mu}, C)$, i.i.d. where C is the covariance function. With observations at levels $x_1, \ldots, x_m$, $\boldsymbol{\theta}_u$ is now distributed as an m-variate normal with mean $(\mu(x_1), \ldots, \mu(x_m))$ and covariance matrix $\boldsymbol{C}$. The common mean curve $\boldsymbol{\mu}$ can be taken to be random, and is endowed

with a suitable prior distribution, e.g., a constant mean Gaussian process: $(\mu(x_1), \ldots, \mu(x_m)) \sim$ GP$(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_m)$. (Here we take the constant mean to be 0 for simplicity). The overall model specification is summarized as follows:

$$
\begin{aligned}
H \equiv \text{GP}(\boldsymbol{\mu}, C), \quad \boldsymbol{\theta}_u | H &\overset{iid}{\sim} H \; u = 1, 2, ..., U \\
\boldsymbol{Y}_{ui} | \boldsymbol{\theta}_u &\overset{iid}{\sim} N(\boldsymbol{\theta}_u, \tau_u^2 \mathbf{I}_m), \quad \text{for all } i = 1, \ldots, n_u
\end{aligned} \tag{3}
$$

where, again, $C(x, x') = \sigma_C^2 \rho(x - x'; \phi_C)$.

Due to conjugacy, conditionally on the data and parameters $\boldsymbol{M} := (\boldsymbol{\mu}, C, \boldsymbol{\tau}, \boldsymbol{\sigma})$, the $\boldsymbol{\theta}_u$'s are independently distributed Gaussian processes with mean $\tilde{\boldsymbol{\mu}}_u$ and covariance $\tilde{C}_u$, respectively. Appendix 1 provides details. Also, conditionally on the data and parameters $\boldsymbol{M}$, $(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v)$ is distributed according to a Gaussian process GP$(\tilde{\boldsymbol{\mu}}_u - \tilde{\boldsymbol{\mu}}_v, \tilde{C}_{u,v})$, where covariance function $\tilde{C}_{u,v} = \tilde{C}_u + \tilde{C}_v$. (To make prior/posterior comparison, the 'no data' versions of the expressions below employ $\mu_u(x)$ and C.)

From Appendix 2, $(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v)^2$ can be expressed as a sum of normal and chi-square variables, and so the expected value:

$$
\mathbb{E}[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}, \boldsymbol{M}] = \int_B (\tilde{\mu}_u(x) - \tilde{\mu}_v(x))^2 dx + \int_B \tilde{C}_{u,v}(x) dx, \tag{4}
$$

$$
\text{Var}[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}, \boldsymbol{M}] = |B| \left(\sum_{k=1}^{\infty} 2\lambda_k^2 + 4\lambda_k \int_B (\tilde{\mu}_u(x) - \tilde{\mu}_v(x)) \psi_k(x) dx \right), \tag{5}
$$

where $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenvalues of the integral operator induced by covariance kernel $\tilde{\boldsymbol{C}}_{u,v}$, while $\boldsymbol{\psi}_k$ are the corresponding eigenfunctions. (We use $\tilde{C}_{u,v}(x)$ to denote $\tilde{C}_{u,v}(x, x)$). The decomposition of the expectation into two terms is worth noting. The first term contributes an integrated squared difference while the second contributes cumulative spatial variation. To obtain $\mathbb{E}[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}]$, one has to integrate out $\boldsymbol{M}$ yielding:

$$
\begin{aligned}
\mathbb{E}[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}] &= \mathbb{E}\left[\int_B (\tilde{\mu}_u(x) - \tilde{\mu}_v(x))^2 dx + \int_B \tilde{C}_{u,v}(x) dx | \text{Data} \right], \\
var[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}] &= var\mathbb{E}[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}, \boldsymbol{M}] + \mathbb{E}[var[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) | \text{Data}, \boldsymbol{M}]]
\end{aligned}
$$

Explicit expressions are no longer available, but the computation can be achieved by sampling over $\boldsymbol{M}$ conditionally on the data.

For d_2, note that for each $x \in B$, conditionally on the data and $\boldsymbol{M}$, we have $Pr(\theta_u(x) - \theta_v(x) > 0) = (1 - \Phi(-(\tilde{\mu}_u(x) - \tilde{\mu}_v(x))/\tilde{C}_{u,v}(x)))$. So,

$$
\begin{aligned}
\mathbb{E}[d_2(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)|\text{Data}, \boldsymbol{M}] &= \int_B 1 - \Phi\bigg(\frac{-(\tilde{\mu}_u(x) - \tilde{\mu}_v(x))}{\tilde{C}_{u,v}(x)}\bigg)dx \\
var[d_2(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)|\text{Data}, \boldsymbol{M}] &= \int_B \bigg[1 - \Phi_2\bigg(\tilde{\mu}_u(x_1) - \tilde{\mu}_v(x_1), \tilde{\mu}_u(x_2) - \tilde{\mu}_v(x_2), \tilde{C}_{u,v}(x_1, x_2)\bigg)\bigg]dx_1 dx_2 \\
&\quad - \bigg[\int_B 1 - \Phi\bigg(\frac{-(\tilde{\mu}_u(x) - \tilde{\mu}_v(x))}{\tilde{C}_{u,v}(x)}\bigg)dx\bigg]^2.
\end{aligned}
$$

Here $\Phi_2(m(x_1), m(x_2)), \rho(x_1, x_2)) := P(Z > 0)$, where Z is a bivariate normal variable with mean $(m(x_1), m(x_2))$ and covariance matrix obtained from the covariance function ρ evaluated at x_1 and x_2 for $x_1 \neq x_2$. For $x_1 = x_2$, $\Phi_2(m(x_1), m(x_2)), \rho(x_1, x_2)) := \Phi(-m(x)/\rho(x_1, x_1))$.

Turning to d_3, it is also simple to obtain the mean expression for d_3 as follows:

$$
\begin{aligned}
\mathbb{E}[d_3(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)|\text{Data}, \boldsymbol{M}] &= \int_B \bigg[1 - \Phi\bigg(\frac{-(\tilde{\mu}_u(x) - \tilde{\mu}_v(x))}{\tilde{C}_{u,v}(x)}\bigg)\bigg] \\
&\quad \bigg[\tilde{\mu}_u(x) - \tilde{\mu}_v(x) + \frac{\phi(-(\tilde{\mu}_u(x) - \tilde{\mu}_v(x))/(\tilde{C}_{u,v}(x)))}{1 - \Phi(-(\tilde{\mu}_u(x) - \tilde{\mu}_v(x))/(\tilde{C}_{u,v}(x)))}\bigg]dx
\end{aligned}
$$

where ϕ is the density for a standard normal variable. The variance expression is unwieldy and is omitted.

If the region B has irregular shape, the foregoing integrals may need to be computed using Monte Carlo integration. Suppose we uniformly sample say, p levels $x_{01}, \ldots, x_{0p} \in B$, while eigenvalues of the integral operator of the covariance function given by the posterior distributions are computed from the induced Gram matrix using levels x_{0t}. In essence, these approximations yield:

$$
\begin{aligned}
\hat{d}_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) &:= \frac{1}{p}\sum_{t=1}^{p}(\theta_u(x_{0t}) - \theta_v(x_{0t}))^2. \\
\hat{d}_2(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) &:= \frac{1}{p}\sum_{t=1}^{p}\mathbb{I}(\theta_u(x_{0t}) - \theta_v(x_{0t}) \geq 0). \\
\hat{d}_3(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) &:= \frac{1}{p}\sum_{t=1}^{p}(\theta_u(x_{0t}) - \theta_v(x_{0t}))_+.
\end{aligned}
$$

Under mild conditions, $d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) - \hat{d}_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) \xrightarrow{\mathcal{P}} 0$ as $p \to \infty$. In fact, $\mathbb{E}\hat{d}_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) \to \mathbb{E}d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$, and $var(\hat{d}_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)) \to var(d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v))$. Note that p does not depend on the available amount of data. Thus we can estimate the expectation and variance expression for d_1 as accurately

as we wish (given that we can obtain exact expressions for $\hat{d}_1$). The same holds for d_2 and d_3. Moreover, by sampling over the posterior distribution of the mean curves $\boldsymbol{\theta}_u$ for all $u \in V$, we can obtain summaries other than the means and variances.

3.2 An example

We illustrate the functional ANOVA from the previous subsection. We consider a two-population problem. Using Gaussian process modeling, we generated two groups of curves with sample size $n_1 = n_2 = 20$. The mean curves $\boldsymbol{\theta}_u$ are random draws with mean $\boldsymbol{\mu} = \mathbf{0}$ and the covariance function takes an exponential form with $\sigma_\theta = .5$ and $\omega_\theta = .02$. The white noise variance $\tau_1 = \tau_2 = .2$ for both groups. The samples are two groups of curves $Y_{ui}(x)$ where $x = 1, 2, \ldots, 50$, $u = 1, 2$ and $i = 1, \ldots, n_u$. Fig. 1 shows the two sets of *observed* curves.

The posterior inference procedure is described in detail in Appendix 1. For prior specification, we set $a_{\tau_u} = 2$ and $b_{\tau_u} = 3$ for $u = 1, 2$; $a_{\sigma_u} = 2$ and $b_{\sigma_u} = 2$ for $u = 1, 2$, and $a_\mu = 10$ and $b_\mu = .01$. We utilized distance measures d_1, d_2 and d_3 described earlier, using illustrative domains B of the form $[x, x + 10]$, for $x = [2, 4, 6, \ldots, 40]$. The posterior distributions of relevant parameters were obtained by MCMC sampling, which were run for 10000 iterations, the last 5000 iterations of which were used for the computation of the posterior distributions. See Fig. 1 (right panel) for an estimate and credible intervals for the mean curves $\boldsymbol{\theta}_u$. The posterior distributions for distance measures can be obtained in two ways, either through MCMC samples for mean curves evaluated at 50 new levels uniformly generated from B, or through analytic expression of conditional expectations given parameters, where the parameters were obtained through MCMC samples. We employ the latter, "Rao-Blackwellized" computation.

To illustrate the spatially varying posterior behavior of the distance measures proposed in the previous subsection see Fig. 2, where a point x on the X axis is associated with the interval $[x, x + 10]$. For small values of x, d_1 has small but strictly positive posterior mean. As x slides to the middle region in the domain (e.g., $x = 18$), the posterior mean for d_1 increases to around 1 with probability close to 1, and as x approaches 30, d_1 decreases to the range of $(.15, .2)$ with high probability. The posterior distribution for d_2 captures the probability that the mean curve of the first population dominates that of the second population. For small values of x, this probability is close to .5, suggesting that the two populations share similar mean curves, and as x approaches

the middle of the interval, the probability decreases to 0, indicating where the first population is dominated by the second population. d_3 also captures where a population is dominated by the other, and by how much. The behavior of these metrics is in accord with the rightmost panel of Fig. 1.

3.3 The SDP case

Here, we replace the GP specification with the SDP, described in Section 2. Applied to the model in (3), the overall hierarchical specification is summarized as:

$$\begin{aligned} H \equiv \text{GP}(\boldsymbol{\mu}, C, \quad G_0|H \sim \text{DP}(\gamma, H), \\ \boldsymbol{\theta}_u|G_0 \overset{iid}{\sim} G_0 \text{ for all } u \in V \\ \boldsymbol{Y}_{ui}|\boldsymbol{\theta}_u \overset{iid}{\sim} N(\boldsymbol{\theta}_u, \tau_u^2 \mathbf{I}_m), \quad \text{for all } i = 1, \ldots, n_u; u \in V. \end{aligned} \tag{6}$$

Priors will be supplied for $\boldsymbol{\mu}$ and C, as well as γ. Under these specifications, the $\boldsymbol{\theta}_u$'s are iid draws from G_0. The distribution G_0 varies around prior H, with the amount of variability governed by γ. It is worth noting that this model specification is richer than and subsumes the one given by (3). In fact, letting $\gamma \to \infty$, the induced prior given by (6) converges in distribution to the one given by (3). Integrating over the random measure G_0, $\boldsymbol{\theta}_u$ is distributed according to a GP distribution H, so that the variance and correlation measures within each group are the same as what we obtained using a GP prior in the previous section.

Next, consider the relationship between two groups u and v. Under the properties of Dirichlet processes, and the fact that the Gaussian process distributions are non-atomic, we obtain, a priori, that $P(\boldsymbol{\theta}_u = \boldsymbol{\theta}_v|\gamma) = \frac{1}{1+\gamma}$. Furthermore,

$$corr(\theta_u(x_1), \theta_v(x_2)|\gamma, C) = \frac{\rho(x_1, x_2)}{(1+\gamma)\rho(x_1)\rho(x_2)}. \tag{7}$$

Turning to our metrics, for say d_1, we have:

$$\begin{aligned} P(d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) = 0|\gamma) = \frac{1}{1+\gamma}, \\ \mathbb{E}[d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)|\gamma, C] = \frac{\gamma}{1+\gamma} \int_B C(x, x) dx. \end{aligned}$$

We can obtain similar expressions for d_2 and d_3. With regard to population comparison, note the difference between the SDP and the GP modeling. With the GP prior, under either d_1, d_2 or d_3, with probability 1 there are no ties between $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$.

4 A new functional ANOVA model

We now specify a Bayesian nonparametric ANOVA model which differs from that of the previous section and allows more detailed population comparison. The novelty comes from now seeking comparison of the G_u's, the the random distributions that generate the curves for individuals in population u. Now, we can compare the G_u's directly or compare features of these distributions, for instance, the functional that is the "mean-at-a-point" functional. Furthermore, as we show below, the comparison can be carried out globally, i.e., an overall comparison of the G_u's (Section 4.2) or locally, i.e., relative to the random distributions at a given x, $G_u(x)$ (Section 4.3).

4.1 The global case

Our development proceeds from the hierarchical Dirichlet process modeling approach of Teh et al. (2006). The idea of this approach is that the random G_u's are i.i.d. draws from a Dirichlet process, $G_u \sim \text{DP}(\alpha, G_0)$, for some base measure G_0, which is also random and is distributed according to another Dirichlet process, i.e., $G_0 \sim \text{DP}(\gamma, H)$. G_0 is a.s. a discrete probability measure, say $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\phi}_k}$. Hence, the specification for G_u implies that the G_u's share the same set of atoms that define G_0. This allows explicit comparison of the populations.

In particular, comparison can proceed through functionals of interest of G_u. Denoting G_u by $\{\pi_{uk}, \boldsymbol{\phi}_{k,D}, k = 1, 2, ...\}$, we have the mean functional, $\mu(G_u) = \sum_k \pi_{uk}\boldsymbol{\phi}_k$ which plays the role of $\boldsymbol{\theta}_u$ of the previous section. The mean functional enables us to make connection with comparisons from the previous section, i.e., we immediately have $m_1(\boldsymbol{\theta}_u, B)$, $m_2(\boldsymbol{\theta}_u, B)$, and $m_3(\boldsymbol{\theta}_u, B)$ and, for populations u and v, we have $d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$, $d_2(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$, and $d_3(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$. Equivalently, we also use notation $d_1(G_u, G_v)$ for $d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$ and so on. It is clear that there can be no ties between the mean functionals; $Pr(\boldsymbol{\theta}_u = \boldsymbol{\theta}_v) = 0$. In computing the expressions associated with these quantities, we only have to plug in the form of the mean functional. For instance, after some minor calculation,

we obtain

$$d_1(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v) = \sum_j \sum_k (\pi_{uj} - \pi_{vj})(\pi_{uk} - \pi_{vk}) \int_B \phi_j(x)\phi_k(x)dx.$$

This is a special case of a L^r norm between G_u and G_v given as follows:

$$d_{L^r}(G_u, G_v) = \left(\int_B |\sum_k (\pi_{uk} - \pi_{vk})\phi_k(x)|^r dx \right)^{1/r}.$$

Other functions, say based upon quantiles can be studied (that is, the function arises as say the qth quantile of the marginal distribution of G_u at x) for individual populations and compared across populations. Also, we can directly compare the G_u's. There is an extensive literature on comparing distributions, e.g. see Dudley (1976). For probability distributions on function spaces, comparisons using divergence measures may be generally difficult due disjoint supports (see, however, Nguyen (2013b)). This issue is circumvented by our hierarchical construction. Indeed, because G_u and G_u share the same support with probability one, the variational distance and the Kullback-Leibler distance between G_u and G_v can be defined by taking the following forms, respectively:

$$\begin{aligned} d_V(G_u, G_v) &= \frac{1}{2} \sum_k |\pi_{uk} - \pi_{vk}|, \\ d_{KL}(G_u, G_v) &= \sum_k \pi_{uk} \log(\pi_{uk}/\pi_{vk}). \end{aligned}$$

It is important to recognize a key difference between this ANOVA specification and that of the previous section. Now, we have $Y_{ui}(x) = \mu + \theta_{ui}(x) + \epsilon_{ui}(x)$. Draws, $\boldsymbol{\theta}_{ui}$ from G_u are realized for each individual, $i = 1, 2, ..., n_u$, within population u. Here, $\boldsymbol{\theta}_u$ is never realized for any population; it is the population mean of these curves. We have a model with random effects and a pure error term but with marginal dependence across the i's and also across the u's. That is, though the G_u's are conditionally independent given G_0, we have $\theta_{ui}(x) = \theta_{vi'}(x)$ if say both draw $\phi_k(x)$ and this happens with probability $\pi_{uk}\pi_{vk}$. We can have ties for the individual-level curves.

The full hierarchical specification is formally as follows:

$$\begin{aligned} H \equiv \text{GP}(\boldsymbol{\mu}, C), \quad G_0|H \sim \text{DP}(\gamma, H), \\ G_u|G_0 \sim \text{DP}(\alpha, G_0), \text{ for all } u \in V \\ \boldsymbol{\theta}_{ui}|G_u \sim G_u \text{ for all } i = 1, \ldots, n_u; u \in V \\ \boldsymbol{Y}_{ui}|\boldsymbol{\theta}_{ui} \sim N(\boldsymbol{\theta}_{ui}, \tau_u^2 \mathbf{I}_m), \text{ for all } i = 1, \ldots, n_u; u \in V \end{aligned} \tag{8}$$

Under this prior specification, the components $\boldsymbol{\theta}_{ui}$ are iid draws from distribution G_u. The distribution G_u varies around G, with the amount of variability governed by α. The distribution G in turn varies around H, with the amount of variability governed by γ. We note here that that the induced prior given by (8) is richer than the one given by (6). Letting $\alpha \to \infty$, the model (8) tends to (6). The distribution H (a Gaussian process) provides the support for a global pool of mean curves, which in turn provide the support for the mean curves for each population. Model fitting is a simple adaptation of Teh et al (2006) to functional data.

Suppose we are interested in a two-way ANOVA, i.e., now we have populations indexed by say factor u with levels $u = 1, 2, ..., U$ and factor w with levels $w = 1, 2, ..., W$. The preceding development is unchanged; we merely replace G_u with $G_{uw} \equiv \{\pi_{uwk}, \boldsymbol{\phi}_k\}$ and $i = 1, 2, ..., n_{uw}$. We draw $\theta_{uw,i}$ from G_{uw} for each individual i at levels u and w.

Interest would often be in "main" effects which are usually interpreted as *marginal* effects for the levels u and w. In this setting, we can define $G_{u\cdot} = \frac{1}{W}\sum_w G_{uw}$ and $G_{\cdot w} = \frac{1}{U}\sum_u G_{uw}$. That is, $G_{u\cdot} = \sum_k \pi_{u\cdot,k}\delta_{\boldsymbol{\phi}_k}$, similarly for $G_{\cdot w}$. Comparison between $G_{u\cdot}$ and $G_{u'\cdot}$ would be carried out as above. For the mean functional, we immediately have $\boldsymbol{\theta}_{u\cdot} = \frac{1}{W}\sum_w \boldsymbol{\theta}_{uw}$, similarly for $\boldsymbol{\theta}_{\cdot w}$. Lastly, the function $\boldsymbol{\theta}_{u\cdot i} = \frac{1}{W}\sum_w \theta_{uwi}$ is not meaningful. We are interested in marginal features of G_{uw} but not in marginal curves at the individual level.

4.2 Local comparison using a nested hierarchy of Dirichlet processes

From Section 4.1, we have seen that, using the hierarchical DP, we can view the functional ANOVA problem through comparison of G_u's. Here, we maintain the objective of comparison of G_u's but switch the sampling scheme. Now, we sample the functions at levels, i.e., at choices of x, obtaining observations from potentially different individuals at different levels. That is, in some settings, the data is such that, within each population, we choose levels of x and at these levels, we sample individuals; we don't sample curves for individuals. In particular, at level $x \in D$, within population u we have observations $Y_{ui}(x)$ for a set of individuals indexed by i. Associated with $Y_{ui}(x)$ is a $\theta_{ui}(x)$ as in the previous section. But, in the absence of curve level data for individual i, we do not envision drawing an entire $\boldsymbol{\theta}_{ui}$ (though it exists conceptually). Rather, we envision $\theta_{ui}(x)$ drawn from a random *local* distribution Q_{ux} which is centered around G_{ux}, the distribution at x under G_u. In particular, we assume $Q_{ux} \sim DP(\alpha_u, G_{ux})$, nesting the Q's within the G_u's.

Extending the stickbreaking notation of the previous section, we now add $Q_{ux} = \sum_{k=1}^{\infty} \omega_{uxk}\delta_{\phi_k(x)}$.

The implication is local selection of the $\theta_{ui}(x)$. That is, $\theta_{ui}(x_1) = \phi_k(x_1)$ with probability ω_{ux_1k} while $\theta_{ui}(x_2) = \phi_k(x_2)$ with probability ω_{ux_2k}. In the global model described in the previous section, $P(\theta_{ui}(x_1) = \phi_k(x_1)) = P(\theta_{ui}(x_2) = \phi_k(x_2)) = \pi_{uk}$. In different words, were we to realize a set $\{\theta_{ui}(x), x \in D\}$, it would not be one of the $\boldsymbol{\phi}_k$'s but rather, just a locally selected collection of θ's resulting in an everywhere discontinuous surface. However, again, we do not think in terms of modeling a curve for individual i, rather, just a $\theta_{ui}(x)$ at a given $x \in D$. Again, we can have ties across populations but now they are *local*; $\theta_{ui}(x) = \theta_{vi'}(x) = \phi_k(x)$ with probability $\omega_{uxk}\omega_{vxk}$.

We still view this construction as a functional ANOVA problem. Individuals are still pre-clustered to populations. Still there is a population-level distribution G_u. Still we can compare G_u's across u. Still we can employ the same metrics as above to compare the populations. All we have done is introduce another level to the DP specification, as noted above, a level *nested* within the specification for G_u. Such additional flexibility is arguably more appropriate with a sampling scheme that samples at different levels of x.

The overall hierarchical specification is summarized as follows:

$$\begin{aligned} H \equiv \text{GP}(\boldsymbol{\mu}, C), \quad G_0|H \sim \text{DP}(\gamma, H), \\ G_u|G_0 \sim \text{DP}(\alpha_0, G_0), \text{ for all } u \in V \\ Q_{u;x}|G_u \sim \text{DP}(\alpha_u, G_{u;x}), \quad \theta_{ui}(x)|Q_{u;x} \sim Q_{u;x} \text{ for all } i = 1, \ldots, n_u; u \in V \\ \boldsymbol{Y}_{ui}|\theta_{ui} \sim N(\theta_{ui}, \tau_u^2 \mathbf{I}_m), \text{ for all } i = 1, \ldots, n_u; u \in V. \end{aligned} \tag{9}$$

To fit this model, we use a demanding MCMC algorithm. Details are presented in Appendix 4.

We can more explicitly describe the model in ((9)) using a stickbreaking parametrization. Due to the discrete nature of Dirichlet process realizations, the random measures G_0, G_u all share the same support. The random measure Q_{ux} also share the same support as the G_u and G_0 when the

latter two are restricted to level x, for any $x \in D$. Indeed, they can be expressed as follows:

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \\ G_u &= \sum_{k=1}^{\infty} \pi_{uk} \delta_{\phi_k}, \\ Q_{ux} &= \sum_{k=1}^{\infty} \omega_{uxk} \delta_{\phi_k(x)}. \end{aligned} \tag{10}$$

As before, the ϕ_k are iid draws from the base measure H. $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$, $\boldsymbol{\pi}_u = (\pi_{uk})_{k=1}^{\infty}$ and $\boldsymbol{\omega}_{ux} = (\omega_{uxk})_{k=1}^{\infty}$ are stick-breaking weight vectors satisfying the following hierarchical specifications:

$$\begin{aligned} \boldsymbol{\beta}|\gamma \sim \text{GEM}(\gamma), \quad \boldsymbol{\pi}_u|\boldsymbol{\beta} \sim \text{DP}(\alpha_0, \boldsymbol{\beta}) \text{ for all } u \in V \\ \boldsymbol{\omega}_{ux}|\boldsymbol{\pi}_u \sim \text{DP}(\alpha_u, \boldsymbol{\pi}_u) \text{ for all } u \in V, x \in D \qquad (11) \\ \phi_k \sim H \text{ for all } k = 1, 2, \ldots \\ z_{ui}(x)|\boldsymbol{\pi}_u \sim \boldsymbol{\omega}_{ux} \text{ for all } i = 1, \ldots, n_u; u \in V \\ Y_{ui}(x)|z_{ui}, (\phi_k)_{k=1}^{\infty} \sim N(\phi_{z_{ui}(x)}(x)|\tau_u^2) \text{ for all } i = 1, \ldots, n_u; u \in V; x \in D. \qquad (12) \end{aligned}$$

As is the case with parametric hierarchical models, the hierarchical framework in the nonparametric context also lends itself naturally to the decomposition of variation measures for data within each group and between groups. Appendix 3 provides details of the decomposition calculation.

5 Hierarchical Dirichlet process examples

Here we present two simulated examples and one real data analysis. The first simulation example presents an unusual functional ANOVA setting. The second serves as a proof of concept for the local nested HDP modeling. The real data analysis fully illustrates all of the foregoing development.

5.1 Multi-modal non-stationary and non-Gaussian and globally sharing groups of functional data

We consider a two-population setting where the first population uses one functional atom, but the second is associated with two functional atoms with (latent) selection probability $1/2$ for each.

One of the two functional atoms in the second group is shared with the first group. Both functional atoms are generated according to a Gaussian process with mean 0 and covariance specified by $\sigma_\theta = 0.5, \omega_\theta = 0.01$. Data associated with a functional atom are obtained by adding independent an white noise process with variance variance $\tau_u = 0.2$ for both $u = 1, 2$. The sample sizes for the two groups are $n_1 = n_2 = 40$. The sample curves $Y_{ui}(x)$ are observed at 50 levels $x = 1, \ldots, 50$ for $i = 1, \ldots, n_u$. For this data set we use the global model described in Section 4.1. For prior specification, $\tau_u^2 \sim$ InGamma(a_{τ_u}, b_{τ_u}) where $a_{\tau_u} = 2, b_{\tau_u} = 1$. The base measure H is also a mean-0 Gaussian process with $\sigma_\theta \sim$ InGamma(a_σ, b_σ) with $a_\sigma = 2, b_\sigma = 1$. In addition $\omega_\theta = 0.01$. (We use a slightly modified parameterization for covariance function $C(x, x') = \sigma_\theta^2 \exp\{-\omega_\theta \|x - x'\|^2\}$). The concentration parameters are specified as $\gamma = 0.005$ and $\alpha = 0.01$. The posterior distributions of parameters and distance measures of interest are obtained via MCMC samples.

Fig. 4 (left) shows that while the number of functional clusters for group 1 is close to 1 with high probability, for group 2 there are two functional clusters with probability close to 1. Moreover, with high probability there are overall two functional clusters for both groups. This implies that the functional cluster that underlies group 1 is in fact also a functional cluster for group 2. The right panel in Fig. 4 illustrate the posterior distributions for population means $\mu(G_1)$ and $\mu(G_2)$. The tight credible interval bands are due to the effect of averaging implicitly over sample curves. Additional comparisons can be performed on the basis of $\mu(G_1)$ and $\mu(G_2)$ using distance measures such as d_1, d_2 and d_3, but these still do not always fully capture the heterogeneity between and within the two groups. Because of the sharing of functional atoms at each MCMC iteration, a visually appealing method for characterizing the variation between and within each group of functional curves, is to perform pairwise comparisons for sample curves on the basis of the functional atoms that the curves are associated with, using the same distance measures mentioned above.

Fig. 5 (right) produces a heatmap in which each entry represents the posterior probability that two given functional curves share the same functional atom. It shows that all sample curves in group 1 share the same functional atom (cluster) with high probability, and that the first 20 sample curves in group 2 also shares the same cluster as that of group 1, while the remaining 20 sample curves in group 2 share another functional cluster. A more detailed analysis is carried out using distance measure d_1 with varying domains in Fig. 6. Each panel provides a different subregion as indicated and the entries in the heatmaps provide the (posterior) mean of the distance between

the global atoms associated with two given sample curves. The heatmaps reveal the need for differential numbers of curves within each population as well as the variability within each group due to the variation between the functional atoms. The variation is most pronounced for, e.g., the interval $[1, 10]$), and is negligible for, e.g., $[21, 30]$).

5.2 Functional ANOVA with sampling at levels of the functions

This simulation example is motivated by the ocean temperature data set. We employ the nested HDP model developed in Section 4.2. Here, we create data from three populations. The populations are regulated by three functional atoms, say, $\boldsymbol{\phi}_1, \boldsymbol{\phi_2}$ and $\boldsymbol{\phi}_3$. These functional atoms were generated according to a mean-0 Gaussian process with and a covariance function given by parameters $\sigma_\phi = 1, \omega_\phi = .01$. Population 1 uses only functional atom $\boldsymbol{\phi_2}$, population 2 uses $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ with equal probabilities, while population 3 uses $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ with equal probabilities. For each population u and level x, observations $Y_{ui}(x)$ are i.i.d. draws from a mixture of Gaussians with the means given by the associating functional atom ($\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ or $\boldsymbol{\phi}_3$) evaluated at x, and the variance given by τ_u^2. We let $\tau_u = 0.1$ for all u. The number of samples at level x is $n_{ux} = 20$ for all u's and i's. The set of x's is $[1, \ldots, 10]$. Fig. 7 shows the data set.

For prior specification, let $\tau_u^2 \sim$ InvGamma(a_{τ_u}, b_{τ_u}) where $a_{\tau_u} = 5, b_{\tau_u} = 1$. The concentration parameter γ is given a vague prior $\gamma \sim$ Gamma(a_γ, b_γ) where $a_\gamma = 1, b_\gamma = .1$, while other concentration parameters are set to $\alpha_0 = 1$ and $\alpha_u = 1$ for all u. The base measure H is a mean-0 Gaussian process with $\sigma_\phi = 1, \omega_\phi = .01$.

Again, the data here are a collection of observations $Y_{ui}(x)$; we are not sampling individual curves. However, the underlying assumption of our model is that there exist functional atoms which provide the basis for underlying functional clusters that regulate these groups of data. We are able to estimate not only these functional clusters, but also infer about whether or not they are shared among the populations. Fig. 8 (left panel) depicts the posterior distributions of the number of functional clusters for each of the three populations. Population 3 has 2 functional clusters with high probability, Population 2 is likely to have two functional clusters (as opposed to 1), and Population 1 has either one or two clusters with approximately equal probabilities. For population 1, there seems to be a disagreement with how the data was generated, but a closer look reveals that that the two functional atoms employed by population 2 and the one which is shared with

group (u)	$\boldsymbol{\pi}_{u1}$	$\boldsymbol{\pi}_{u2}$	$\boldsymbol{\pi}_{u3}$
1	**0.9669** (0.0808)	0.0231 (0.0631)	0.0045 (0.0112)
2	**0.4769** (0.1490)	**0.5048** (0.1337)	0.0072 (0.0174)
3	0.0151 (0.0229)	**0.5293** (0.1107)	**0.4520** (0.1108)

Table 1: Posterior mean (and standard deviation) of the mixing proportions for the (dominant) three functional atoms for each group of data.

population 1 are virtually indistinguishable for a significant proportion of levels. Thus in the a posteriori analysis it makes sense to have either of the two functional atoms provide the support for clusters in the data in population 1. Fig. 8 (right panel) depicts the mean estimate and credible intervals for three functional atoms the provide overwhelming support for the data in all three populations. The estimation of the functional atoms is very accurate.

Following Sections 4.1 and 4.2, Table 1 provides the (posterior) mean of the mixing proportions for the three functional atoms with respect to each of the three populations. Accordingly we obtain variational distances between groups: $d_V(G_1, G_2) = 0.49(\pm.11); d_V(G_2, G_3) = 0.52(\pm 0.09); d_V(G_1, G_3) = 0.96(\pm.04)$. The KL distances tend to amplify the difference: $d_{KL}(G_1, G_2) \approx 0.68; d_{KL}(G_1, G_3) \approx 6.08; d_{KL}(G_2, G_1) \approx 3.21; d_{KL}(G_2, G_3) \approx 2.59; d_{KL}(G_3, G_1) \approx 16.79; d_{KL}(G_3, G_2) \approx 8.02$. But it is clear that populations G_1 and G_3 are the most different pair. Turning to local comparisons, Fig. 9 depicts the median and credible intervals for the number of local clusters at each x and population u. There is significantly more variability in population 2 and population 3 than in population 1. Note that for population 1, the median number of local clusters is one for all x, where at most of the $x \geq 4$ there is a probability of having two local clusters. For population 2, the median number of local clusters is two for all x, but for $x \geq 5$ there is also a significant probability that there are only one local cluster. This agrees with the fact that the two functional atoms can be interchanged for $x \geq 4$. For population 3, the median number of local clusters is two, but there are levels with non-negligible probability of having only one or three.

group (u)	$\boldsymbol{\pi}_{u1}$	$\boldsymbol{\pi}_{u2}$	$\boldsymbol{\pi}_{u3}$	$\boldsymbol{\pi}_{u4}$	$\boldsymbol{\pi}_{u5}$	$\boldsymbol{\pi}_{u6}$
1	**0.96** (0.01)	0.00 (0)	0.02 (0.01)	0.00 (0)	0.00 (0)	0.00 (0)
2	**0.02 (0.09)**	**0.74** (0.09)	**0.06** (0.03)	0.00 (0)	**0.12** (0.02)	0.00 (0.02)
3	**0.04** (0.14)	0.01 (0.02)	0.02 (0.02)	**0.90** (0.14)	0.01 (0.02)	0.00 (0)
4	**0.03** (0.13)	0.02 (0.03)	0.02 (0.03)	**0.89** (0.14)	0.01 (0.02)	0.00 (0)

Table 2: Posterior mean (and standard deviation) of the mixing proportions for the (dominant) three functional atoms for each group of data.

5.3 Analysis of an ocean temperature vs depth dataset

We consider a data set consisting of ocean temperature and depth measurements collected at locations in the Atlantic Ocean. The geographic separation naturally divides the locations into 4 distinct groups – see the right panel of Fig. 10 – which we take as our populations. At each location the ocean temperature is recorded, together with the depth and the time where and when the measurement was obtained. The left panel of Fig. 10 illustrates this data set. Because the temperature are recorded at different times (during the days, and across several days), we treat the data not as a collection of functional curves, rather as a collection of temperatures $Y_{ui}(x)$, where $u \in 1, 2, 3, 4$, x indexes the depth level, and i indexes the measurements obtained at that depth level within group u. (There is not enough temporal structure in the dataset to attempt to model time effects.) Again, we are interested in comparison among the 4 groups based on the functional patterns of ocean temperature in terms of ocean depth. There are a total of 4917 such measurements within the first 500 meters of depth. The data set is generally unbalanced: some locations and/or depth levels have more data than others. Moreover, the depths are not equally spaced.

Although locations of measurements obtained within each group are known, due to their close proximity relative to the distances between the groups, we assume that the measurements obtained within depth level are exchangeable. Furthermore, the 4 groups are also viewed as exchangeable. The modeling, inference and analysis were described in Section 4. We grouped the data into 25 equally spaced depth levels, each of which is 20 meters long. The temperature measurements were re-centered around 10^o Celsius, and then re-scaled so that a majority of the measurements fall within $-[1, 1]$. For prior specifications, for the white noise process

	G_1	G_2	G_3	G_4
G_1	0	9.2 (5.5)	6.1 (3.3)	7.1 (3.9)
G_2	8.5 (6.0)	0	6.7 (4.2)	6.3 (5.0)
G_3	7.1 (4.1)	7.0 (3.8)	0	0.3 (0.3)
G_4	7.2 (4.6)	7.0 (4.4)	0.3 (0.4)	0

Table 3: Estimates of $d_{KL}(G_u, G_v)$.

we let $\tau_u \sim$ InvGamma(a_{τ_u}, b_{τ_u}), where $a_{\tau_u} = 5, b_{\tau_u} = 1$. For the concentration parameters, we let $\gamma \sim$ Gamma(a_γ, b_γ) where $a_\gamma = 5, b_\gamma = 1$. We let $\alpha_0 \sim$ Gamma$(a_{\alpha_0}, b_{\alpha_0})$ where $(a_{\alpha_0}, b_{\alpha_0}) = (1, 1)$, while $\alpha_u = 1$ for all $u = 1, \ldots, 4$. The base measure H is specified as a mean-0 Gaussian Process, whose covariance function has the standard exponential form with $(\sigma_\phi, \omega_\phi) = (0.1, 0.5)$.

We next discuss the sensitivity of the hyperparameters in our hierarchical model. Recall that σ_ϕ specifies the variance at a point and ω_ϕ the smoothness of the Gaussian process (which generates functional atoms). Small values of ω_ϕ result in very smooth functional atoms, while large values result in highly distinct (and less smooth) atoms. Both extremes are avoided; our choice of σ_ϕ, ω_ϕ reflects roughly the range of the variance and smoothness based on an exploratory analysis of the data collected at locations 3 and 4 (where the global clusters are mostly unimodal by visual inspection). The posterior inference is found to be robust in this range. Once the levels of smoothness and variance for the functional atoms are specified, the hyperparameters for the white noise variance τ_u are chosen to be highly non-informative. Turning to the Dirichlet processes' concentration parameters, a robust choice is to restrict α_u's and α_0 to relatively small ranges (say, ≤ 1). Once these are fixed, the hyperparameters for γ (the concentration parameter for the top level Dirichlet process) are chosen to be highly non-informative. The rationale behind this hinges on the interactions of the Dirichlet processes in multiple levels of a Bayesian nonparametric hierarchy. Such a theory has emerged only recently (Nguyen, 2013a). Finally, we note that for the purpose of group comparisons using the proposed summaries, the resultant inference appears quite robust to the choice of these hyperparameters, as we will see below.

The Gibbs sampling algorithm described in Appendix 4 is run for 5000 iterations, which took

several days to complete on a personal desktop computer. While this seems sufficient for our purpose, for larger scale data sets that have more depth levels (say, in the order of thousands) approximate variational methods may be considered to help speed up the Gibbs sampler. An example of such a strategy developed for a somewhat related and more complex model can be found in Nguyen and Gelfand (2011).

The posterior distribution of the number of functional atoms associated with the functional temperature-depth trends in the four groups have the support between 5 and 9, with a strong mode at 6, suggesting there are 6 dominant functional patterns. Fig. 11 shows the posterior mean and credible intervals of these functional patterns. Table 2 offers the contributions of each of these individual functional patterns within each group of data. It provides strong evidence regarding the functional variations between the groups, and in some cases, within a group. In particular, group 1 is overwhelmingly associated with functional curve (atom) $\boldsymbol{\phi}_1$. This also implies *single* functional behavior of depth vs temperature within group 1. Group 3 and group 4 also have largely single functional behavior, with most of the contributions (89%) coming functional curve (atom) $\boldsymbol{\phi}_4$, and small contributions coming from $\boldsymbol{\phi}_1$. In fact, the decompositions represented by $\boldsymbol{\pi}$ is are almost indistinguishable between the two groups. Group 2 exhibits very heterogeneous functional behavior; there are contributions from more than 4 functional curves, $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3, \boldsymbol{\phi}_5$. Using KL distance measures to compare between groups in Table 3, it is evident that group 3 and 4 exhibit very similar functional behaviors, while group 2 is most different from the other groups. It is useful to examine the posterior distributions of the variance τ_u^2 for the noise processes associated with the functional curves. In particular, the posterior mean for τ_1 is very small ($0.42(0.10)$), suggesting the highly predictable behavior of temperature in group 1. τ_3 is largest ($2.07(0.25)$), perhaps due to the relative sparsity of measurements obtained within group 3, in spite of the fact that the overall behavior of group 3 and 4 are very similar. For completeness for τ_2 we have $0.93(0.52)$ and for τ_4 we have $1.84(0.22)$.

Fig. 12 illustrates the posterior mean and credible intervals for $d_2(\mu(G_u), \mu(G_v))$, providing detailed comparison in temperature vs depth behavior in the four groups. For instance, group 1 has consistently higher temperature than group 2, while group 3 and group 4 are very similar. Using d_3, it is observed that the difference between group 3 and 2 increases with lower depth. Despite the sparse and unbalanced data in some of the groups, our functional modeling approach provides

relatively fine-scaled comparisons across depth levels. Fig. 13 reveals in more detail the variations in the number of local clusters in each of the 4 groups. Again, the number of local clusters within group 1 is 1 (supported by ϕ_1 with overwhelming probability. The number of local clusters in group 3 and 4 are 1 with overwhelming probability at shallow depths (less than 300 meters for group 3), but at deeper depths they are also associated with local clusters supported by ϕ_1. Group 2 has up to 5 local clusters at shallow depth levels, but the number of local clusters decreases to 2 at deeper depth levels. In other words, more functional variation in the temperature behavior is observed near the ocean surface for group 1 than at the deeper levels.

6 Summary and future work

We have presented a sequence of models for the functional ANOVA problem which enable comparison between populations in ways not previously considered in the literature. In particular, our hierarchical DP versions permit comparison of the (random) functions that define the populations using various metrics and over chosen subdomains. Also, we can provide comparison of the random distributions that generate the functions for individuals within the populations. Through simulation examples and a set of temperature vs. depth data, the rich inferential possibilities have been revealed.

An opportunity for future work is to look at the comparison of the populations dynamically. With a suitable data (e.g., temperature vs. depth relationships for various geographically defined groups collected across years), we can imagine a functional ANOVA model at each time point. Explicit modeling might be developed utilizing a state space specification. Novel inference would include the assessment of how differences between populations are evolving in time.

7 Appendix

7.1 Appendix 1: Inference of mean curves under GP prior

This section provides standard expressions for conditional expectation and variance of population mean curves given a collection of functional data. Suppose that the data $\boldsymbol{Y} = \{Y_{ui}(x)\}$ are observed at the same set of levels $x_1, \ldots, x_m$. In the following we use $\boldsymbol{M}$ to collect all model

parameters, $\boldsymbol{M} = (\boldsymbol{\mu}, \boldsymbol{C}, \boldsymbol{\sigma}, \boldsymbol{\tau})$. Given $\boldsymbol{Y}$ and $\boldsymbol{M}$, $\boldsymbol{\theta}_u = (\theta_u(x_1), \ldots, \theta_u(x_m))$ are independent for $u \in V$. Let $x_{01}, \ldots, x_{0p}$ be p levels that are either placed regularly in B, or uniformly sample from B. For a given population u, we need to derive the posterior distribution for both $\boldsymbol{\theta}_u$, and $\boldsymbol{\theta}_{0u} := (\theta_u(x_{01}), \ldots, \theta_u(x_{0p}))$.

Let $\boldsymbol{C}_u, \boldsymbol{C}_{0u}$ be the a priori covariance matrices for $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_{0u}$, respectively, while $\boldsymbol{R}_u$ be the covariance matrix of size $m \times p$ for the two as given by the GP with covariance function $\boldsymbol{C}$. We have

$$\begin{aligned}
\boldsymbol{\theta}_u|\text{Data}, \boldsymbol{M} &\sim N_m(\tilde{\boldsymbol{\mu}}_u, \tilde{\boldsymbol{C}}_u), \text{where} \\
\tilde{\boldsymbol{C}}_u^{-1} &= \boldsymbol{C}_u^{-1} + (n_u/\tau_u^2)\mathbf{I}_m \\
\tilde{\boldsymbol{C}}_u^{-1}\tilde{\boldsymbol{\mu}}_u &= \boldsymbol{C}_u^{-1}\boldsymbol{\mu} + (1/\tau_u^2)\sum_{i=1}^{n_u}\boldsymbol{Y}_{ui}.
\end{aligned}$$

We have $\boldsymbol{\theta}_{0u}|\boldsymbol{\theta}_u, \boldsymbol{M} \sim N_p(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{S}})$ where

$$\begin{aligned}
\tilde{\boldsymbol{m}} &= \boldsymbol{m}_{0u} + \boldsymbol{R}_u^T\boldsymbol{C}_u^{-1}(\boldsymbol{\theta}_u - \boldsymbol{\mu}) \\
\tilde{\boldsymbol{S}} &= \boldsymbol{C}_{0u} - \boldsymbol{R}_u^T\boldsymbol{C}_u^{-1}\boldsymbol{R}_u, \text{where} \\
\boldsymbol{m}_{0u} &= (\mu(x_{01}), \ldots, \mu(x_{0p})).
\end{aligned}$$

Due to conditional independence relation, $\boldsymbol{\theta}_{0u} \perp \text{Data}|\boldsymbol{\theta}_u, \boldsymbol{M}$, so we have:

$$[\boldsymbol{\theta}_{0u}|\text{Data}, \boldsymbol{M}] \propto \int [\boldsymbol{\theta}_{0u}|\boldsymbol{\theta}_u, \boldsymbol{M}] \times [\boldsymbol{\theta}_u|\text{Data}, \boldsymbol{M}] d\boldsymbol{\theta}_u$$

Standard calculations yield

$$\begin{aligned}
\boldsymbol{\theta}_{0u}|\text{Data}, \boldsymbol{M} &\sim N_p(\tilde{\boldsymbol{\mu}}_{0u}, \tilde{\boldsymbol{C}}_{0u}), \text{where} \\
\tilde{\boldsymbol{\mu}}_{0u} &= \boldsymbol{m}_{0u} + \boldsymbol{R}_u^T\boldsymbol{C}_u^{-1}(\tilde{\boldsymbol{\mu}}_u - \boldsymbol{\mu}) \\
\tilde{\boldsymbol{C}}_{0u} &= \tilde{\boldsymbol{S}} + \boldsymbol{R}_u^T\boldsymbol{C}_u^{-1}\tilde{\boldsymbol{C}}_u\boldsymbol{C}_u^{-1}\boldsymbol{R}_u.
\end{aligned}$$

Finally, we need to sample $\boldsymbol{M} = (\boldsymbol{\mu}, \boldsymbol{C}, \boldsymbol{\tau}, \boldsymbol{\sigma})$ conditionally on the data. This can be achieved via Gibbs sampling.

1. Conditional for $\boldsymbol{\mu}$: This is normal with covariance matrix and mean specified by:

$$\begin{aligned}
\boldsymbol{C}_\mu^{-1} &= \sum_u (\boldsymbol{C}_u + \tau_u^2\mathbf{I}_m)^{-1} + (1/\sigma_\mu^2)\mathbf{I}_m \\
\boldsymbol{C}_\mu^{-1}\boldsymbol{\mu}_\mu &= \sum_u (\boldsymbol{C}_u + \tau_u^2\mathbf{I}_m)^{-1}\sum_{i=1}^{n_u}\boldsymbol{Y}_{ui}.
\end{aligned}$$

2. Conditional for $\boldsymbol{\tau}_u$, for each u: Endow τ_u with igamma(a_{τ_u}, b_{τ_u}), then the conditional for τ_u^2 is also igamma with updated parameters $b_{\tau_u} := a_{\tau_u} + mn_u/2$ and $b_{\tau_u} := b_{\tau_u} + \sum_{i=1}^{n_u} \|\boldsymbol{Y}_{ui} - \boldsymbol{\theta}_u\|^2/2$.

3. Conditional for σ_μ. Endow σ_μ with igamma(a_μ, b_μ) then the conditional for σ_μ^2 is updated by $a_\mu := a_\mu + m/2$ and $b_\mu = b + 1/2\|\boldsymbol{\mu}\|^2$.

4. Conditional for $\boldsymbol{C}_u$, for each u: $\boldsymbol{C}_u$ is parameterized by exponential form, so that $C_u(x_1, x_2) = \sigma_{C_u}^2 S_u$ where $S_u(x_1, x_2) = \exp -\phi_u(x_1 - x_2)^2$. Endow $\sigma_{C_u}^2$ with igamma(a_{C_u}, b_{C_u}), which is updated via $a_{C_u} = a_{C_u} + m/2$ and $b_{C_u} := b_{C_u} + \frac{1}{2}(\boldsymbol{\theta}_u - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\boldsymbol{\theta}_u - \boldsymbol{\mu})$. ϕ_u is updated via a symmetric Metropolis update, with an acceptance rate equal to $\min(1, \exp -\frac{1}{2\sigma_u^2}(\boldsymbol{\theta}_u - \boldsymbol{\mu})^T(\tilde{\boldsymbol{S}}_u^{-1} - \boldsymbol{S}_u^{-1})(\boldsymbol{\theta}_u - \boldsymbol{\mu}))$.

7.2 Appendix 2: Properties of summary metrics

Suppose that $\boldsymbol{\theta}$ is distributed according to a Gaussian process on a closed domain $B \subset \mathbb{R}$ with mean $\boldsymbol{\mu}$ and covariance function $\boldsymbol{C}$. $\boldsymbol{C}$ can be viewed as a positive semidefinite kernel. Moreover, assume that $\int C(x_1, x_2)dx_1dx_2 < \infty$, and consider the integral operator $L_{\boldsymbol{C}} : L_2(B) \to L_2(B)$ induced by the kernel $\boldsymbol{C}$:

$$L_C f(x) = \int_B C(x, x')f(x')dx'.$$

This is a self-adjoint, positive and compact operator with a countable systems of non-negative eigenvalues $\{\lambda_k\}_{k=1}^\infty$ and associated eigenfunctions $\{\boldsymbol{\psi}_k\}_{k=1}^\infty$ which form an orthonormal basis of $L_2(B)$. By Mercer's theorem, $\boldsymbol{C}$ admits the following decomposition: $C(x, x') = \sum_{k=1}^\infty \lambda_k \boldsymbol{\psi}_k(x)\boldsymbol{\psi}_k(x')$. Here the series converges absolutely for each pair x, x' and uniformly in B. For each $k \in \mathbb{N}_+$, define

$$\eta_k = \int_B (\theta(x) - \mu(x))\boldsymbol{\psi}_k(x)dx.$$

By Karhunen-Loève's theorem applied to Gaussian processes, $\boldsymbol{\theta}$ can be written as: $\boldsymbol{\theta} = \boldsymbol{\mu} + \sum_{k=1}^\infty \eta_k \boldsymbol{\psi}_k$, where the convergence is almost sure and is uniform in x. Moreover, the collection of coefficients $\{\eta_k\}$ are independent mean-0 Gaussian variables with variance $var(\eta_k) = \lambda_k$, for any $k \in \mathbb{N}_+$.

It is simple to obtain that $m_1(\boldsymbol{\theta})$ can be expressed in terms of a sum of chi-square and normal variables:

$$m_1(\boldsymbol{\theta}) = \|\boldsymbol{\mu}\|^2 + \sum_{k=1}^{\infty} \eta_k^2 + 2\sum_{k=1}^{\infty} \eta_k \boldsymbol{\mu}^T \boldsymbol{\psi}_k.$$

Due to the mutual independence of η_k's, we obtain that:

$$\mathbb{E}[m_1(\boldsymbol{\theta})|\boldsymbol{\mu}, \boldsymbol{C}] \;\; = \;\; \|\boldsymbol{\mu}\|^2 + \sum_{k=1}^{\infty} \lambda_k = \|\boldsymbol{\mu}\|^2 + \int_B \boldsymbol{C}(x,x)ds.$$

The variance takes the form:

$$\begin{aligned} var[m_1(\boldsymbol{\theta})|\boldsymbol{\mu}, \boldsymbol{C}] &= \mathbb{E}\left[\left(\sum_{k=1}^{\infty} \eta_k^2 + 2\sum_{k=1}^{\infty} \eta_k \boldsymbol{\mu}^T \boldsymbol{\psi}_k - \sum_{k=1}^{\infty} \lambda_k\right)^2 \middle| \boldsymbol{\mu}, \boldsymbol{C}\right] \\ &= \sum_{k=1}^{\infty} 2\lambda_k^2 + 4\lambda_k (\boldsymbol{\mu}^T \boldsymbol{\psi}_k)^2. \end{aligned}$$

where we have used the fact that $\mathbb{E}\eta_k = \mathbb{E}\eta_k^3 = 0$; $\mathbb{E}\eta_k^2 = \lambda_k$, $\mathbb{E}\eta_k^4 = 3\lambda_k^2$. Although the λ_k and $\boldsymbol{\psi}_k$ are determined directly from $\boldsymbol{C}$, except for some special cases closed forms are not available. In practice one might consider sampling for the variance instead.

7.3 Appendix 3: Decomposition of variance and correlation measures

First, we study the relations among random measures in the model. G_0 is a random measure that varies around $H = \text{GP}(\boldsymbol{\mu}, C)$, where the variation is governed by γ. For each group u, G_u is a random measure that varies around G_0, where the variation is governed by α. For each level x and group u, Q_{ux} varies around G_u, where the variation is governed by α_u. Because $G_0 \sim \text{DP}(\gamma, H)$, due to elementary properties of Dirichlet processes for any measurable set A of functions

$$\begin{aligned} \mathbb{E}[G_0(A)^2|H] &= \frac{1}{\gamma+1} H(A) + \frac{\gamma}{\gamma+1} H(A)^2, \\ var[G_0(A)|H] &= \frac{1}{\gamma+1}(H(A) - H(A)^2). \end{aligned}$$

Turning to the random measures G_u for each $u \in V$,

$$var[G_u(A)|G_0] = \frac{1}{\alpha+1}(G_0(A) - G_0(A)^2).$$

Marginalizing out G_0, we have:

$$\begin{aligned} var[G_u(A)|H] &= \mathbb{E}[var[G_u(A)|G_0]|H] + var[\mathbb{E}[G_u(A)|G_0]|H] \\ &= \frac{1}{\alpha+1}(H(A) - \mathbb{E}[G_0(A)^2|H]) + var[G_0(A)|H] \\ &= \left(\frac{1}{\gamma+1} + \frac{\gamma}{(\gamma+1)(\alpha+1)}\right)(H(A) - H(A)^2). \end{aligned} \tag{13}$$

Next, for the random measures Q_{ux} at each level $x \in D$, for any measurable set A_x, as before:

$$var[Q_{ux}(A_x)|G_u] = \frac{1}{\alpha_u+1}(G_u(A_x) - G_u(A_x)^2),$$

so that

$$\begin{aligned} var[Q_{ux}(A_x)|G_0] &= \mathbb{E}[var[Q_{ux}(A_x)|G_u]|G_0] + var[\mathbb{E}[Q_{ux}(A_x)|G_u]|G_0] \\ &= \frac{1}{\alpha_u+1}\mathbb{E}[(G_u(A_x) - G_u(A_x)^2)|G_0] + var[G_u(A_x)|G_0] \\ &= \left(\frac{1}{\alpha+1} + \frac{\alpha}{(\alpha+1)(\alpha_u+1)}\right)(G_0(A_x) - G_0(A_x)^2). \end{aligned}$$

Marginalizing out G_0, we have:

$$\begin{aligned} var[Q_{ux}(A_x)|H] &= \mathbb{E}[var[Q_{ux}(A_x)|G_0]|H] + var[\mathbb{E}[Q_{ux}(A_x)|G_0]|H] \\ &= \left(\frac{1}{\gamma+1} + \frac{\gamma}{(\gamma+1)(\alpha+1)} + \frac{\gamma\alpha}{(\gamma+1)(\alpha+1)(\alpha_u+1)}\right)(H(A_x) - H(A_x)^2) \end{aligned} \tag{14}$$

Next, let A and B are measurable sets with respect to observations at x_1 and x_2, respectively. For $\boldsymbol{\phi} \sim H$, let $H_{x_1}(A) = P(\phi(x_1) \in A|H)$ and $H_{x_1,x_2}(A,B) = P(\phi(x_1) \in A; \phi(x_2) \in B|H)$. Then similar calculation yields, for measure G_0:

$$cov[G_0(A), G_0(B)|H] = \frac{1}{\gamma+1}(H_{x_1,x_2}(A,B) - H_{x_1}(A)H_{x_2}(B))$$

For measure G_u, we have:

$$cov(G_u(A), G_u(B)|H) = \left(\frac{1}{\gamma+1} + \frac{\gamma}{(\gamma+1)(\alpha+1)}\right)(H_{x_1,x_2}(A,B) - H_{x_1}(A)H_{x_2}(B)).$$

Similarly, for Q_{ux}:

$$\begin{aligned} &cov(Q_{ux}(A), Q_{ux}(B)|H) = \\ &\left(\frac{1}{\gamma+1} + \frac{\gamma}{(\gamma+1)(\alpha+1)} + \frac{\gamma\alpha}{(\gamma+1)(\alpha+1)(\alpha_u+1)}\right)(H_{x_1,x_2}(A,B) - H_{x_1}(A)H_{x_2}(B)). \end{aligned}$$

In all expressions above, a priori, the concentration parameters regulate the fraction of variance or correlation that are passed from one level in the Bayesian hierarchy to the next, starting from the base measure H, which regulates the dependence with respect to covariate x. Lastly, we note that similar calculations can be carried out between populations. We omit the details.

7.4 Appendix 4: Posterior computation for the model in (9)

We recall and introduce key notations: $\boldsymbol{\phi}_k$ is a random draw from H, $\boldsymbol{\psi}_t$ a random draw from G_0, $\boldsymbol{\varphi}_{ur}$ a random draw from G_u. Finally, $\theta_{ui}(x)$ is a random draw from Q_{ux}.

Let k_t denote the index of the $\boldsymbol{\phi}_k$ associated with the functional atom $\boldsymbol{\psi}_t$, i.e., $\boldsymbol{\psi}_t = \boldsymbol{\phi}_{k_t}$. Let t_{ur} denote the index of the $\boldsymbol{\psi}_t$ associated with the functional atom $\boldsymbol{\varphi}_{ur}$ in group u, i.e., $\boldsymbol{\varphi}_{ur} = \boldsymbol{\psi}_{t_{ur}}$. Let r_{ui}^x denote the index of the $\boldsymbol{\varphi}_{ur}(x)$ associated with the atom $\theta_{ui}(x)$, i.e., $\theta_{ui}(x) = \boldsymbol{\varphi}_{ur_{ui}^x}(x)$. The local and functional atoms are related by: $\theta_{ui}(x) = \boldsymbol{\varphi}_{ur_{ui}^x}(x) = \boldsymbol{\psi}_{t_{ur_{ui}^x}}(x) = \boldsymbol{\phi}_{k_{t_{ur_{ui}^x}}}(x)$.

Recall that a priori $G_0 \sim \text{DP}(\gamma, H)$. Due to a standard property of a Dirichlet process, conditioning on the global factors $\boldsymbol{\phi}_k$'s and the index vector $\boldsymbol{k}$, the posterior distribution of G_0 is distributed according to a DP: $[G_0|\boldsymbol{k}, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K] \sim \text{DP}(\gamma + q_\cdot, \frac{\gamma H + \sum_{k=1}^K q_k \delta_{\boldsymbol{\phi}_k}}{\gamma + q_\cdot})$, where $q_k = \#\{t : k_t = k\}$ denotes the number of $\boldsymbol{\psi}_t$'s associating with $\boldsymbol{\phi}_k$, and $q_\cdot = \sum_{k=1}^K q_k$. This implies an explicit representation for G_0 as follows:

$$\begin{aligned} G_0 &= \sum_{k=1}^K \beta_k \delta_{\boldsymbol{\phi}_k} + \beta_{\text{new}} G_0^{\text{new}}, \\ \boldsymbol{\beta} &= (\beta_1, \ldots, \beta_K, \beta_{\text{new}}) \sim \text{Dir}(q_1, \ldots, q_K, \gamma) \\ G_0^{\text{new}} &\sim \text{DP}(\gamma, H). \end{aligned} \tag{15}$$

Similarly, conditionally on G_0, the random distributions G_u are independent across the group indices u. In particular, given G_0, $\boldsymbol{k}$, $\boldsymbol{t}_u$ and the $\boldsymbol{\phi}_k$'s, the posterior of G_u is distributed as: $[G_u|G_0, \boldsymbol{k}, \boldsymbol{t}, (\boldsymbol{\phi}_k)_{k=1}^K] \sim \text{DP}(\alpha_0 + m_{u\cdot}, \frac{\alpha_0 G_0 + \sum_{k=1}^K m_{uk}\delta_{\boldsymbol{\phi}_k}}{\alpha_0 + m_{u\cdot}})$, where $m_{uk} = \#\{r : k_{t_{ur}} = k\}$, the number of φ_{ur} associated with $\boldsymbol{\phi}_k$, and $m_{u\cdot} = \sum_{k=1}^K m_{uk}$. This implies the following representation for G_u: $G_u = \sum_{k=1}^K \pi_{uk}\delta_{\boldsymbol{\phi}_k} + \pi_{u\text{new}} G_u^{\text{new}}$, where $G_u^{\text{new}} \sim \text{DP}(\alpha_0\beta_{\text{new}}, G_0^{\text{new}})$ and

$$\boldsymbol{\pi}_u = (\pi_{u1}, \ldots, \pi_{uK}, \pi_{u\text{new}}) \sim \text{Dir}(\alpha_0\beta_1 + m_{u1}, \ldots, \alpha_0\beta_K + m_{uk}, \alpha_0\beta_{\text{new}}). \tag{16}$$

Once more, conditionally on G_u, the random distributions Q_{ux} are independent across levels x. In

particular, given G_u, $\boldsymbol{k}$, $\boldsymbol{t}_u$, $\boldsymbol{r}_u^x$, and the $\boldsymbol{\phi}_k$'s, the posterior of Q_{ux} is distributed as:

$$[Q_{ux}|G_u, \boldsymbol{k}, \boldsymbol{t}_u, \boldsymbol{r}_u^x, (\boldsymbol{\phi}_k)_{k=1}^K] \sim \text{DP}(\alpha_u + n_{ux\cdot}, \frac{\alpha_u G_{ux} + \sum_{k=1}^K n_{uxk}\delta_{\boldsymbol{\phi}_k(x)}}{\alpha_u + n_{ux\cdot}}),$$

where $n_{uxk} = \#\{i : k_{t_{ur_{ui}^x}} = k\}$, the number of $\theta_{ui}(x)$ associated with $\boldsymbol{\phi}_k(x)$, and $n_{ux\cdot} = \sum_{k=1}^K n_{uxk}$. This implies the following representation for Q_{ux}:

$$\begin{aligned} Q_{ux} &= \sum_{k=1}^K \omega_{uxk}\delta_{\boldsymbol{\phi}_k(x)} + \omega_{ux\text{new}}Q_u^{x\text{new}} \\ \boldsymbol{\omega}_{ux} &= (\omega_{ux1}, \ldots, \omega_{uxK}, \omega_{ux\text{new}}) \sim \text{Dir}(\alpha_u\pi_{u1} + n_{ux1}, \ldots, \alpha_u\pi_{uK} + n_{uxK}, \alpha_u\pi_{u\text{new}}) \quad (17) \\ Q_u^{x\text{new}} &\sim \text{DP}(\alpha_u\pi_{u\text{new}}, G_u^{x\text{new}}). \end{aligned}$$

The above characterization suggests a straightforward Gibbs sampling algorithm by constructing a Markov chain for $(\boldsymbol{\phi}_k)_{k=1}^K, \boldsymbol{k}, \boldsymbol{t}, \boldsymbol{r})$. To simplify the implementation by avoiding the bookkeeping steps of the index variables, we will consider a modified block Gibbs sampling algorithm by constructing a Markov chain for the count variables (e.g., $\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{n}$) instead. We will still need the index variable z_{uxi}, which denotes the index of the global atom $\boldsymbol{\phi}_k$ that local atom $\theta_{ui}(x)$ is associated with, i.e., $z_{uxi} = k_{t_{ur_{ui}^x}}$. Note that the likelihood of the data involves only the z_{uxi} variables, and that $\boldsymbol{n}_{ux}$ can be calculated directly in terms of z_{uxi}'s:

$$n_{uxk} = \sum_i \mathbb{I}(z_{uxi} = k).$$

We proceed to describe a block Gibbs sampler by considering a Markov chain for $(\boldsymbol{\phi}, \boldsymbol{q}, \boldsymbol{m}, \boldsymbol{n}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\omega})$.

Sampling $\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\omega}$. Conditional probabilities: $[\boldsymbol{\beta}|\boldsymbol{q}, \gamma] \times \prod_u [\boldsymbol{\pi}_u|\boldsymbol{m}_u, \boldsymbol{\beta}, \alpha_0] \times \prod_u \prod_x [\boldsymbol{\omega}_{ux}|\boldsymbol{n}_{ux}, \boldsymbol{\pi}_u, \alpha_u]$ are given by Eqs. (15)(16)(17).

Sampling of $\boldsymbol{z}$. Note that a priori, $z_{uxi}|\boldsymbol{\omega}_{ux} \sim \boldsymbol{\omega}_{ux}$. Let n_{uxk}^{-uxi} denote the number of data items in the group u and level x, except $y_{ui}(x)$, associating with the mixture component k. Then,

$$p(z_{uxi} = k|\boldsymbol{z}^{-uxi}, \boldsymbol{\omega}, \boldsymbol{\phi}_k, \text{Data}) = \begin{cases} (\alpha_u\pi_{uk} + n_{uxk}^{-uxi})F(y_{ui}(x)|\boldsymbol{\phi}_k(x)) & \text{if } k \text{ is previously used} \\ \alpha_u\pi_{u\text{new}} f_{uxk^{\text{new}}}^{y_{ui}(x)}(y_{ui}(x)) & \text{if } k = k^{\text{new}}, \end{cases}$$

where $f_{uxk^{\text{new}}}^{y_{ui}(x)}(y_{ui}(x)) = \int F(y_{ui}(x)|\boldsymbol{\phi}(x))dH(\boldsymbol{\phi}(x))$ is the prior density of $y_{ui}(x)$.

Sampling of $\boldsymbol{m}_u$. Recall that m_{uk} is the number of functional atoms $\boldsymbol{\varphi}_{ur}$ associated with $\boldsymbol{\phi}_k$ within each group u. This set of functional atoms φ_{ur}'s can be subdivided into disjoint subsets associated with levels $x \in D$ when the functional atoms φ_{ur} are first generated. Let m_{uxk} be the number of such functional atoms corresponding to the level x. To be precise,

$$\begin{aligned} m_{uxk} &= \#\{r_{ui}^x : z_{uxi} = k_{t_{ur_{ui}^x}} = k \text{ for some } i\} \\ m_{uk} &= \sum_{x \in D} m_{uxk}. \end{aligned}$$

m_{uxk} corresponds to the number of partitions among the n_{uxk} atoms $\theta_{ui}(x)$ such that $z_{uxi} = k$. To obtain the distribution of m_{uxk}, consider the distribution of r_{ui}^x conditionally on G_u (i.e., $\boldsymbol{\pi}_u$, $\boldsymbol{\phi}_k$'s). Note that given G_u, the Q_{ux} are independent across x's. For each atom $\theta_{ui}(x)$, the probability of being assigned to an existing atom $\boldsymbol{\varphi}_{ur}(x)$ such that $k_{t_{ur}} = k$ is

$$p(r_{ui}^x = r | k_{t_{ur}} = k, \boldsymbol{r}^{-uxi}, \boldsymbol{\pi}_u) \propto n_{ux\cdot r}^{-uxi}$$

while the probability of being assigned to a new atom $\boldsymbol{\varphi}_{ur^{\text{new}}}(x)$ is

$$p(r_{ui}^x = r^{\text{new}} | k_{t_{ur^{\text{new}}}} = k, \boldsymbol{r}^{-uxi}, \boldsymbol{\pi}_u) \propto \alpha_u \pi_{uk}$$

where $n_{ux\cdot r}^{-uxi} := \#\{i' : r_{ui'}^x = r; uxi \neq uxi'\}$, the number of data items at group u and level x except $y_{ui}(x)$ that are associated with $\boldsymbol{\varphi}_{ur}$. This implies that m_{uxk} is the number of partitions that arise in a population of n_{uxk} data items, whose distribution is distributed according to a Dirichlet process with concentration parameter $\alpha_u \pi_{uk}$. It was shown by Antoniak (1974) that the distribution of m_{uxk} has the form:

$$p(m_{uxk} = m | \boldsymbol{z}, \boldsymbol{m}^{-uxk}, \boldsymbol{\pi}_u) = \frac{\Gamma(\alpha_u \pi_{uk})}{\Gamma(\alpha_u \pi_{uk} + n_{uxk})} s(n_{uxk}, m) (\alpha_u \pi_{uk})^m,$$

where $s(n, m)$ are unsigned Stirling number of the first kind.

Sampling $\boldsymbol{q}$. The conditional distribution of $\boldsymbol{q}$ can be obtained in a similar manner as $\boldsymbol{m}$. It can be shown that $q_k = \sum_{u \in V} q_{uk}$ where $q_{uk} = \#\{t : k_{t_{ur}} = k \text{ for some } r\}$. Moreover, q_{uk} is the number of partitions that arise in a population of m_{uk} atoms, whose distributed according to a Dirichlet process with concentration parameter $\alpha_0 \beta_k$:

$$p(q_{uk} = q | \boldsymbol{z}, \boldsymbol{q}^{-uk}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha_0 \beta_k)}{\Gamma(\alpha_0 \beta_k + m_{uk})} s(m_{uk}, q) (\alpha_0 \beta_k)^q.$$

Sampling ϕ. The conditional distribution for ϕ can be obtained easily. Suppose that the prior distribution H for ϕ_k is given by a mean function $\boldsymbol{\mu}$ and covariance function $\boldsymbol{C}$, which is reduced to a covariance matrix $\boldsymbol{C}_k$ when restricted to a finite number of covariate values for x. Then the posterior distribution for ϕ_k is also Gaussian with mean and covariance expressions given as follows:

$$\begin{aligned}
\tilde{\boldsymbol{C}}_k^{-1} &= \boldsymbol{C}_k^{-1} + \sum_{u \in V} \text{diag}(\cdots, \sum_x n_{uxk}, \cdots)/\tau_u^2 \\
\tilde{\boldsymbol{C}}_k^{-1}\tilde{\boldsymbol{\mu}}_k &= \boldsymbol{C}_k^{-1}\boldsymbol{\mu}_k + \left(\cdots, \sum_{u \in V} \sum_{i=1}^{n_u} Y_{ui}(\cdot)\mathbb{I}(z_{u \cdot i} = k)/\tau_u^2, \cdots \right)^T.
\end{aligned}$$

References

Antoniak, C. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, 2, 1152–1174.

Banerjee, S., Carlin, B., and Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC Press.

Brumback, B. and Rice, J. (1998), "Smoothing spline models for the analysis of nested and crossed samples of curves," *J. Amer. Statist. Assoc.*, 93, 961–980.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley, NY.

DeIorio, M., Muller, P., Rosner, G., and MacEachern, S. (2004), "An ANOVA model for dependent random measures," *J. Amer. Statist. Assoc.*, 99, 205–215.

Dudley, R. M. (1976), *Probabilities and metrics: Convergence of laws on metric spaces, with a view to statistical testing*, Aarhus Universitet.

Dunson, D. (2010), "Nonparametric Bayes applications to biostatistics," *Bayesian Nonparametrics: Principles and Practice*, In N. Hjort, C. Holmes, P. Mueller, and S. Walker (Eds.).

Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

Gelfand, A., Kottas, A., and MacEachern, S. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *J. Amer. Statist. Assoc.*, 100, 1021–1035.

Ishwaran, H. and Sunil-Rao, J. (2005), "Spike and Slab Variable Selection: Bayesian and Frequentist Strategies," *Annals of Statistics*, 33, 730–773.

Ishwaran, H. and Zarepour, M. (2002), "Dirichlet prior sieves in finite normal mixtures," *Statistica Sinica*, 12, 941–963.

Kaufman, C. and Sain, S. (2010), "Bayesian Functional ANOVA Modeling Using Gaussian Process Prior Distributions," *Bayesian Analysis*, 5, 123–150.

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *Proceedings of the Section on Bayesian Statistical Science, American Statistical Association*.

MacLehose, R. F. and Dunson, D. (2009), "Nonparametric Bayes kernel-based priors for functional data analysis," *Statistica Sinica*, 19, 611–629.

Morris, J. R. and Carroll, R. J. (2006), "Wavelet-based functional mixed models," *J. Royal Stat. Soc. B*, 68, 179–199.

Nguyen, X. (2010), "Inference of global clusters from locally distributed data," *Bayesian Analysis*, 5, 817–846.

— (2013a), "Borrowing strength in hierarchical Bayes: convergence of the Dirichlet base measure," *arxiv.org/abs/1301.0802*.

— (2013b), "Convergence of latent mixing measures in finite and infinite mixture models," *Annals of Statistics*, 41, 370–400.

Nguyen, X. and Gelfand, A. (2011), "The Dirichlet labeling process for clustering functional data," *Statistica Sinica*, 21, 1249–1289.

Petrone, S., Guidani, M., and Gelfand, A. (2009), "Hybrid Dirichlet processes for functional data," *Journal of the Royal Statistical Society B*, 71(4), 755–782.

Ramsay, J. O. and Silverman, B. (2006), *Functional Data Analysis*, Springer, 2nd ed.

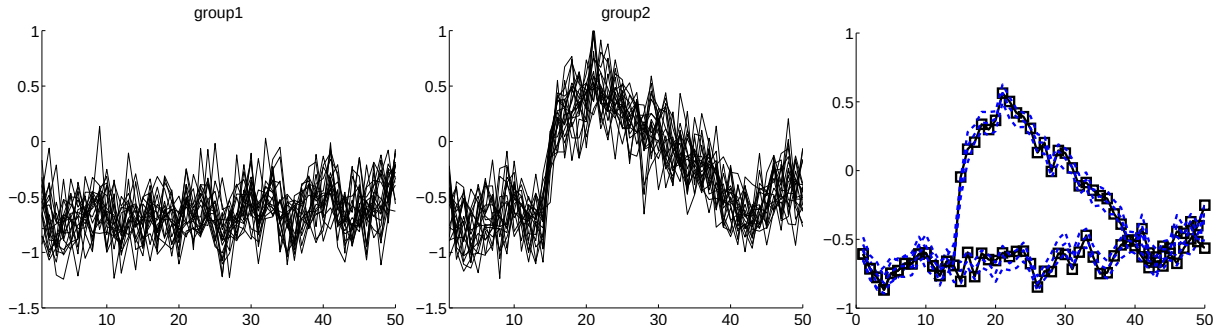


Figure 1: Data set 1. Rightmost panel: Credible intervals and posterior means of mean curves.

Rappold, A., Lavine, M., and Lozier, S. (2007), "Subjective Likelihood for the assessment of trends in the ocean's mixed layer depth," *J. of Amer. Stat. Assoc.*, 102, 771–787.

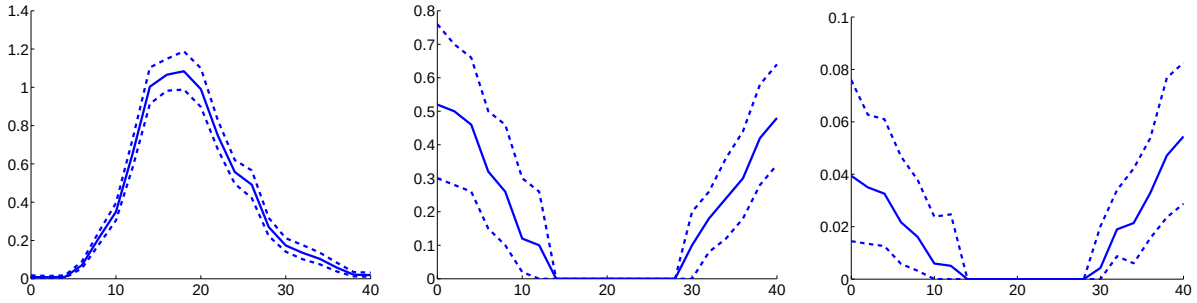
Rodriguez, A., Dunson, D., and Gelfand, A. (2009), "Bayesian nonparametric functional data analysis through density estimation," *Biometrika*, 96(1), 149–162.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Spitzner, D., Marron, J., and Essick, G. (2003), "Mixed-model functional ANOVA for studying human tactile perception," *Journal of the American Statistical Association*, 98, 263–272.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, 101, 1566–1581.

Wang, N., Carroll, R., and Lin, X. (2005), "Efficient semiparametric marginal estimation for longitudinal/clustered data," *J. of Amer. Stat. Assoc.*, 100, 147–157.



(a) d_1 (b) d_2 (c) d_3

Figure 2: Distance measures for group 1 and group 2, using d_1, d_2, d_3.

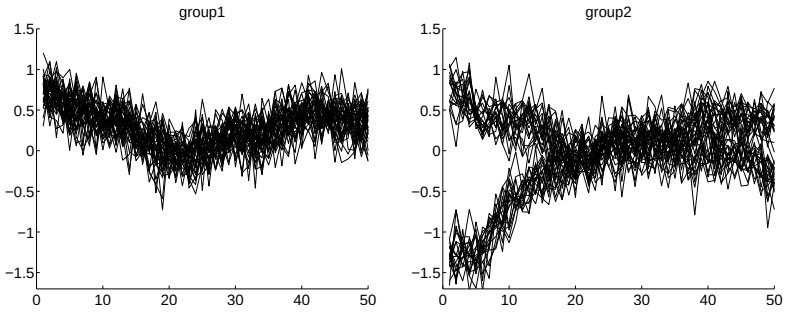


Figure 3: Data set 2.

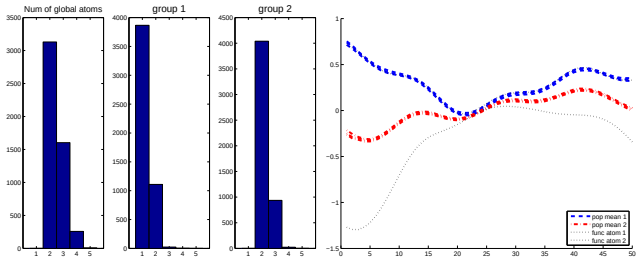


Figure 4: Data set 2. Posterior distribution of the number of functional atoms (left). Right panel: Estimates of population means $\mu(G_u)$ in two dash lines. Estimates of two functional atoms in dotted lines. (Note two of the four lines at the top are almost indistingushable in the plot).

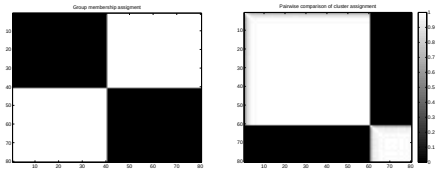


Figure 5: Data set 2. Left panel: Group assignment – the first 40 curves belong to group 1, the second 40 curves group 2. Right panel: Posterior probability that two sample curves share the same functional atom.

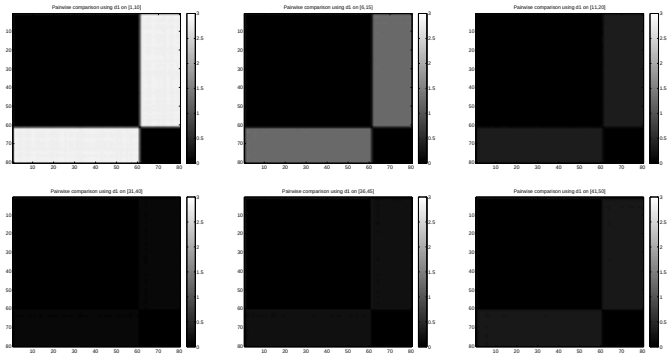


Figure 6: Pairwise comparison using d_1 for varying domains as indicated at the top of each panel.

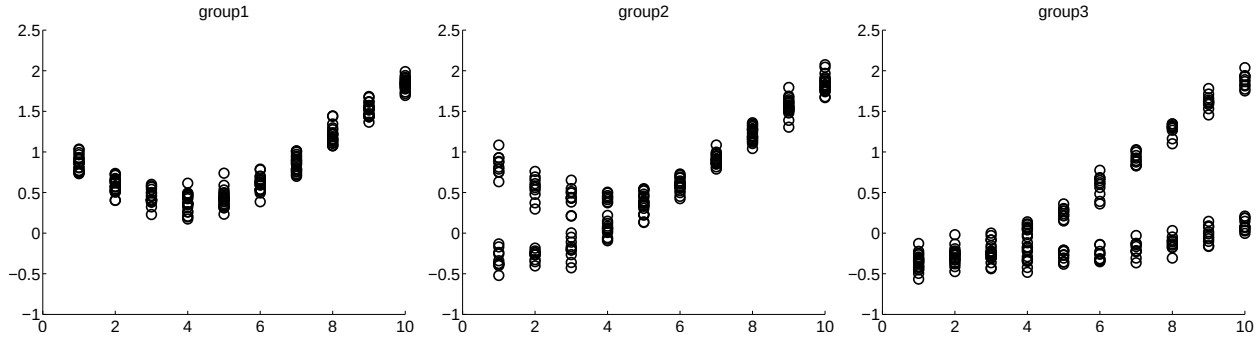


Figure 7: Data set 3. Data given are collection of "dots", not curves, indexed by group membership u, and level x.

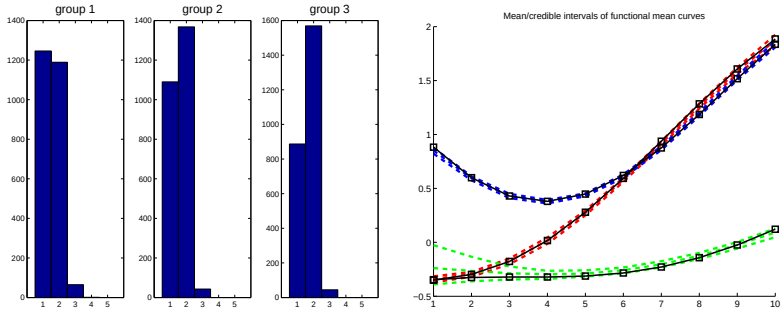


Figure 8: Data set 3. Left panel: Posterior distribution of the functional atoms. Right panel: Mean estimate and credible intervals (in dash) for the functional atoms. The "true" functional atoms are solid plots with square markers.

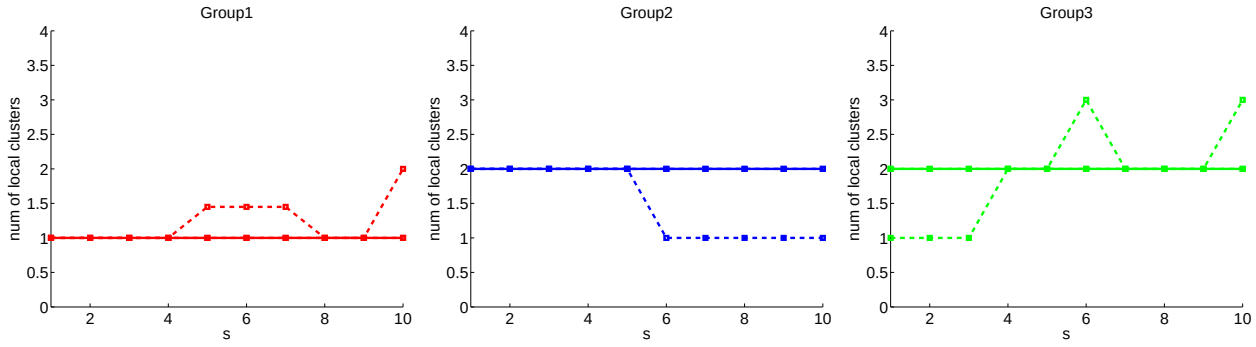


Figure 9: Data set 3. Comparing number of local atoms across u and x.

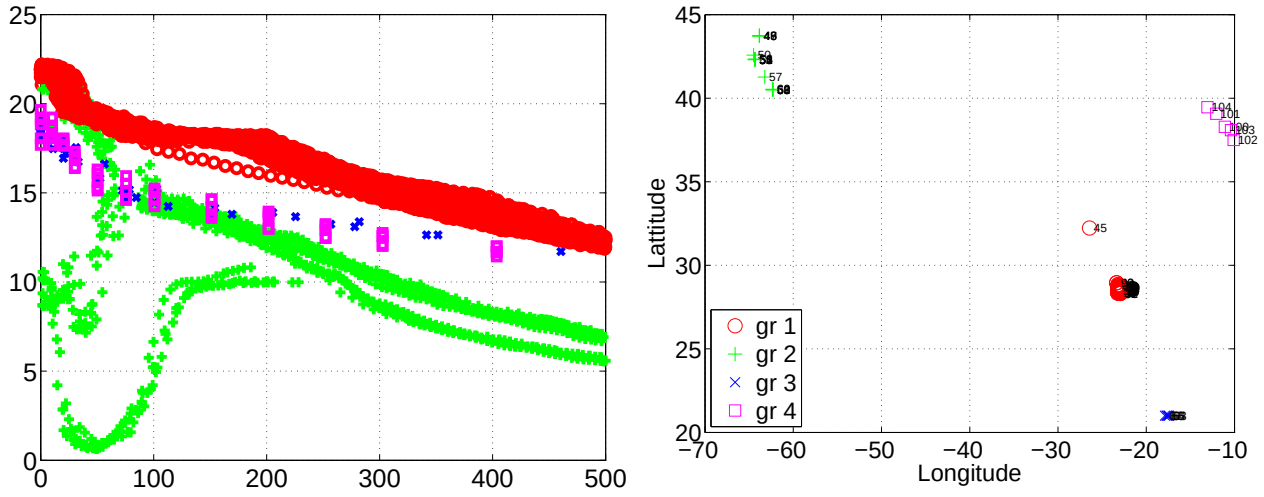


Figure 10: Data set 4. Ocean temperature and depth data, collected in 4 groups in the Atlantic Ocean. Left panel: Y axis represents temperature (Celsius), X axis represents depth (in meters). Measurements from group 1 are illustrated in circles, Group 2's are '+'s, Group 3's are 'x"s, Group 4's are squares. The geographical locations of the 4 groups are depicted in the right panel.

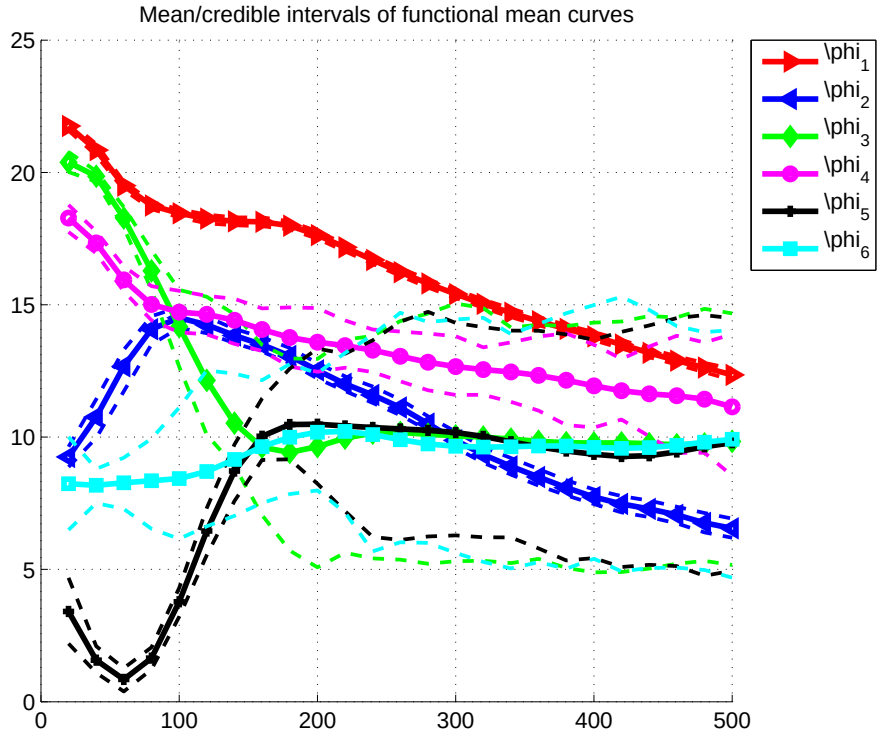


Figure 11: Data set 4. Posterior means and (.05,.95) credible intervals of the functional atoms.

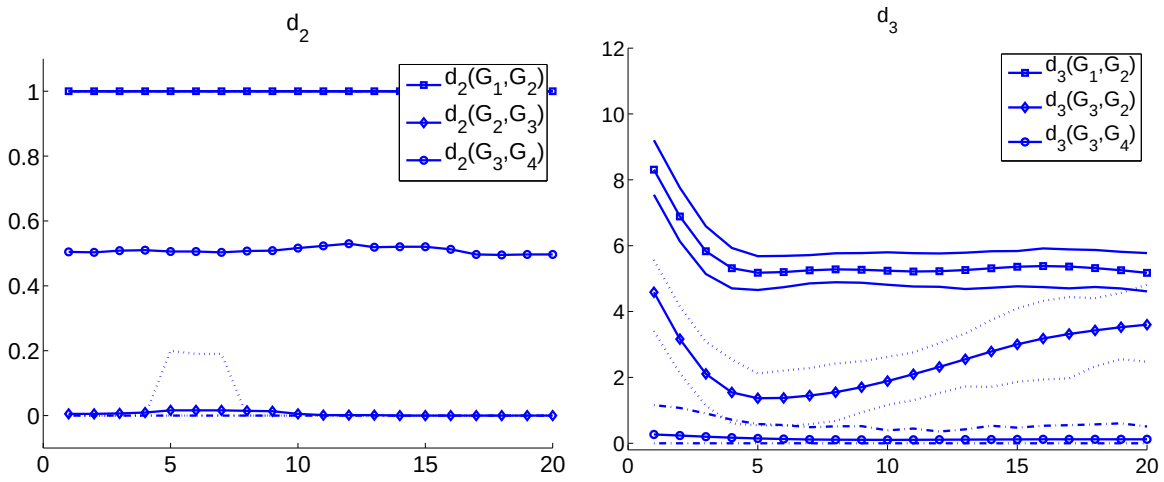


Figure 12: Data set 4. Posterior distributions of distance measures d_2 and d_3, applied to windows of depth interval $[x, x + 4] \times 20$ meters.

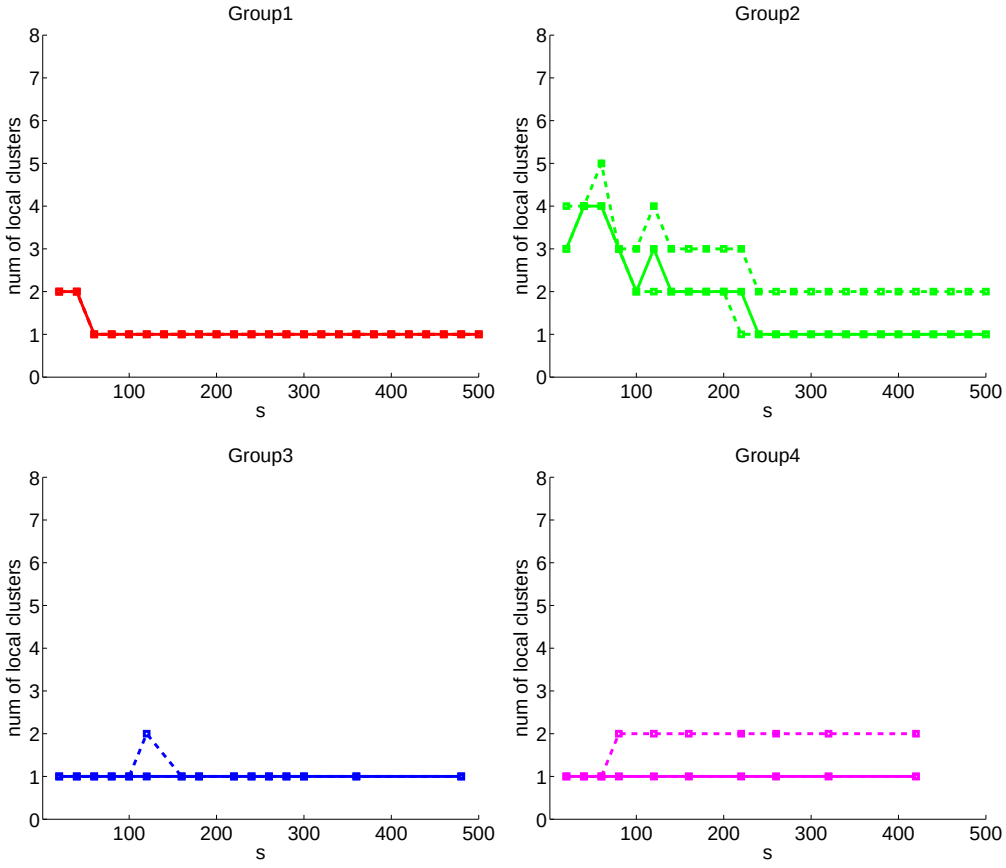


Figure 13: Data set 4. Comparing number of local clusters that vary with depth level x. The plots show posterior mean (solid) and (.05,.95) credible intervals.