Hierarchical Double Dirichlet Process Mixture of Gaussian Processes

Aditya Tayal and **Pascal Poupart** and **Yuying Li**
{amtayal, ppoupart, yuying}@uwaterloo.ca
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada

Abstract

We consider an infinite mixture model of Gaussian processes that share mixture components between non-local clusters in data. Meeds and Osindero (2006) use a single Dirichlet process prior to specify a mixture of Gaussian processes using an infinite number of experts. In this paper, we extend this approach to allow for experts to be shared non-locally across the input domain. This is accomplished with a hierarchical *double* Dirichlet process prior, which builds upon a standard hierarchical Dirichlet process by incorporating local parameters that are unique to each cluster while sharing mixture components between them. We evaluate the model on simulated and real data, showing that sharing Gaussian process components non-locally can yield effective and useful models for richly clustered non-stationary, non-linear data.

Introduction

Gaussian processes (GPs) have been successfully used in regression, classification, function approximation and density estimation (Rasmussen and Williams 2006). They provide a flexible approach to modeling data by assuming a prior directly on functions without explicitly parameterizing the unknown function. The prior specifies general properties of the function like smoothness and characteristic length-scale, which are encoded by the choice of the kernel covariance function and its parameters (GP hyper-parameters). However, the covariance function is commonly assumed to be stationary and consequently the function is unable to adapt to varying levels of smoothness or noise. This can be problematic in some data-sets such as geophysical data, where for example flat regions will be more smooth than mountainous areas, or financial time-series, where distinct volatility regimes can govern the variability of prices.

One way to address nonstationary functions is to consider complex covariance specifications. For example, Paciorek and Schervish (2003) present a class of kernel functions where the smoothness itself is allowed to vary smoothly. However, this does not account for the possibility of nonstationary noise or sharp changes in the process, and in general, it can be difficult to decide on a covariance function with sufficient flexibility while still ensuring positive definiteness. Another line of work looks at combining several Gaussian processes. Gramacy and Lee (2008) use treed partitioning to divide up the input space and fit different GPs independently in each region. Rasmussen and Ghahramani (2002), Meeds and Osindero (2006) similarly use a divide and conquer strategy inspired by Mixture of Experts architectures. Rasmussen and Ghahramani (2002) divide the input space probabilistically by a gating network into regions corresponding to separate GPs, while Meeds and Osindero (2006) use a posterior from a mixture distribution as the gating network, corresponding to a fully generative model instead of a conditional one. Both employ a single Dirichlet process (DP) to allow for the number of mixture components to grow with data complexity. By allowing separate GPs to operate on different input areas, non-stationary covariance and noise levels can be modeled.

However, in these approaches individual GP experts operate locally in the input space. This can lead to unnecessary experts when a single GP may be more appropriate to model data across distant areas in the input space. We extend the Mixture of Experts model proposed by Meeds and Osindero (2006) (hereon, DP-MoGP) to allow GPs to be used non-locally. This is accomplished by using a hierarchical *double* Dirichlet process (HDDP) prior, which incorporates cluster specific parameters in a standard hierarchical Dirichlet process. We use the HDDP to generate local Gaussian parameters that cluster the input space, while sharing GP mixture components between clusters. We call this the hierarchical double Dirichlet process mixture of Gaussian Processes (HDDP-MoGP).

In DP-MoGP each GP roughly dominates an elliptical region of the input space (single Gaussian), whereas in HDDP-MoGP a GP covers an infinite mixture of Gaussians over the input space, allowing much richer shapes while sharing statistical strength between clusters. For instance, if some areas of the input have sparse observations, but there is similar behavior elsewhere in the input region, it can more reliably identify GP hyperparameters for these sets. An example arises in financial time series: as a new regime starts, we are usually interested in adjusting the model as quickly and accurately as possible. In addition, it provides an inherent clustering algorithm. For example, in geophysical data, we

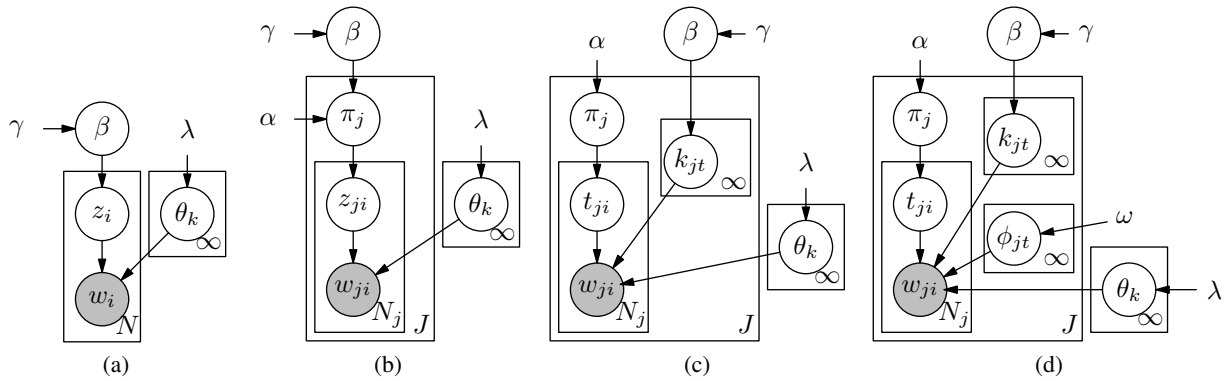


Figure 1: Bayesian-Net plates for (a) Single DP with indicators z_i, (b) HDP with indicators z_{ji}, (c) HDP with indicators t_{ji}, k_{jt}, and (d) Hierarchical double Dirichlet process (HDDP) with indicators t_{ji}, k_{jt}. Parameters $\theta_k \sim H(\lambda)$ are drawn from the top-level base distribution (indexed by k). HDDP extends the HDP model by allowing additional table specific parameters, $\phi_t \sim L(\omega)$ to be sampled separately from the bottom-level base distribution (indexed by t).

can identify regions belonging to similar elevations, and in financial series, we can establish recurring market regimes.

Sec. 2 presents background material, Sec. 3 and 4 develop the HDDP-MoGP model. Sec. 5 illustrates experiment results and we conclude with a discussion in Sec. 6.

Background

Dirichlet Processes

A Dirichlet process (DP), denoted by $G_0 \sim \text{DP}(\gamma, H)$, is a distribution whose domain itself is a random distribution. H is an arbitrary base distribution and γ is the concentration parameter. A draw from a DP returns an output distribution whose support is a set of discrete samples from the base distribution. Weights, β, are sampled from a stick-breaking process (Sethuraman 1994), denoted by $\beta \sim \text{GEM}(\gamma)$:

$$\begin{aligned} G_0(\theta) &= \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k), \qquad \theta_k \sim H(\lambda), \\ \beta_k &= \beta'_k \prod_{l=1}^{k-1} (1-\beta'_l), \qquad \beta'_k \sim \text{ Beta}(1,\gamma). \end{aligned} \tag{1}$$

where $\delta(\theta, \theta_k)$ is the Kronecker delta function. The DP can be used as a prior on the parameters of a mixture model of unknown complexity. To generate observations we sample $\bar{\theta}_i \sim G_0$ and $w_i \sim F(\bar{\theta}_i)$, usually via sampling an indicator variable $z_i \sim \beta$, which corresponds to the component generating $w_i \sim F(\theta_{z_i})$ (Fig. 2a).

The hierarchical Dirichlet process (HDP) (Teh et al. 2006) extends the DP allowing for sharing of mixture components among groups of data. The HDP draws group specific distributions $G_j \sim \text{DP}(\alpha, G_0)$, $j = 1, ..., J$ from a base distribution G_0, which itself is sampled from a global DP prior according to (1). Thus each group is associated with a mixture model, where a group generally represents different entities in a collection, for instance documents in a corpus or individuals in a population. J represents the total number of such entities. Since G_0 is always discrete, there is a strictly positive probability of the group specific distributions having overlapping support, thus allowing parameters to be shared between groups:

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k), \qquad \pi_j \sim \text{ DP}(\alpha, \beta). \tag{2}$$

An observation w_{ji} corresponds to a unique global component, θ_k, via an indicator variable $z_{ji} \sim \pi_j$ and $w_{ji} \sim F(\theta_{z_{ji}})$ (Fig. 2b). The generative process can be described using the Chinese restaurant franchise (CRF) analogy (Teh et al. 2006), where each group corresponds to a restaurant in which customers (observations), w_{ji}, sit at tables (clusters), t_{ji}. Each table shares a single dish (parameter) $\theta_{k_{ji}}$, which is ordered from a global shared menu G_0 (Fig. 1c). We can analytically integrate G_0 and G_j to determine the marginals,

$$p(t_{ji} | t_{j1}, ..., t_{ji-1}, \alpha) \propto \textstyle\sum_t n_{jt} \delta(t_{ji}, t) + \alpha \delta(t_{ji}, t^{new}), \tag{3}$$

$$p(k_{jt} | k_{j1}, ..., t_{jt-1}, \gamma) \propto \textstyle\sum_k m_k \delta(k_{jt}, k) + \gamma \delta(k_{jt}, k^{new}), \tag{4}$$

where n_{jt} is the number of customers seated in table t of restaurant j, and m_k is the number of tables assigned to θ_k.

GPs and Infinite Mixture of GPs

A Gaussian process (GP) describes a distribution over functions (Rasmussen and Williams 2006). For a real process $f(x), x \in \mathbb{R}^D$, we define a mean function $m(x) = \mathbf{E}[f(x)]$ and a covariance function $k(x,x') = \mathbf{E}[(f(x) - m(x))(f(x') - m(x'))]$ and write the Gaussian process as $f(x) \sim \text{GP}(m(x), k(x,x'))$. Thus a GP is a collection of (possibly infinite) random variables. Consistency is ensured, since for any finite subset of inputs the marginal distribution of function values are multivariate normal:

$$f | X, \theta \sim \text{ N}\big(m(X), K(X,X') \big| \theta\big), \tag{5}$$

where $X \in \mathbb{R}^{N \times D}$, consists of N points stacked in rows, $m(X) \in \mathbb{R}^N$, $[m(X)]_i = m(X_{i,:})$ is the mean function,

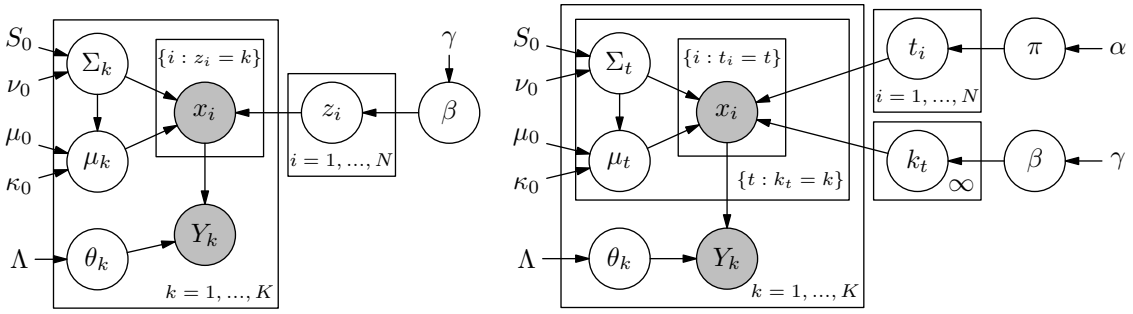


Figure 2: Bayesian-net plates for (a) DP-MoGP, $Y_k = \{y_i : z_i = k\}$, and (b) HDDP-MoGP, $Y_k = \{y_i : k_{t_i} = k\}$. In DP-MoGP, each GP specified by its hyperparameters, θ_k, is associated with a single elliptical Gaussian, (μ_k, Σ_k) in the input space, whereas in HDDP-MoGP, θ_k, is drawn from the top level level DP is associated with several local input Gaussians, (μ_t, Σ_t), which are drawn from the bottom level DP.

$K(X,X') \in \mathbb{R}^{N\times N}$, $[K(X,X')]_{ij} = k(X_{i,:}, X'_{j,:})$ is the kernel function, and θ is the set of hyperparameters used in the mean and covariance functions.

A popular choice is a constant mean and a squared exponential covariance function,

$$m(x) = c,$$
$$k(x,x') = \sigma^s \exp\left\{\frac{-||x-x'||^2}{l^2}\right\} + \sigma^\varepsilon \delta(x,x'), \tag{6}$$

which sets a prior in terms of average value (c), signal variance (σ^s), length-scale (l) and noise (σ^ε) of the function. The set $\theta = \{c, \sigma^s, l, \sigma^\varepsilon\}$ constitutes the (hyper-)parameters of the GP.

The specification of the prior is important, because it fixes the properties of the functions considered for inference. As discussed earlier, for some datasets, using a stationary set of hyperparameters may be too restrictive over the entire domain. Meeds and Osindero (2006) present an infinite mixture of GP experts to overcome this issue using a generative probabilistic model and a single Dirichlet process (DP-MoGP) shown in Fig. **??**.

The generative process does not produce i.i.d. data points. Therefore the generative process is formulated as a joint distribution over a dataset of a given size.

To generate the data, we first construct a partition of the N observations into at most N clusters using a Dirichlet process. This assignment of observations is denoted by the indicator variables $\{z_i\}$. For each cluster of points, $\{z_i : z_i = k\}$, $k = 1, ..., K$, we sample the input Gaussian (μ_k, Σ_k) from a normal-inverse-Wishart prior with hyperparameters $\{S_0, \nu_0, \mu_0, \kappa_0\}$. We can then sample locations of the input points, $X_k = \{x_i : z_i = k\}$. For each cluster we also sample the set of GP hyperparameters denoted by θ_k, where $\theta_k \sim H(\Lambda)$. Finally, using the input locations, X_k and the set of GP hyperparameters, θ_k, for individual clusters we formulate the GP mean vector and output covariance matrix, and sample the set of output variables $Y_k = \{y_i : z_i = k\}$ from the joint Gaussian distribution given by (5).

Input data are conditionally i.i.d. given cluster k, while output data are conditionally i.i.d. given the corresponding cluster of input data. Thus we can write the full joint distribution for N observations, assuming at most N clusters, as follows:

$$P(\{x_i\},\{y_i\},\{z_i\},\{\mu_k,\Sigma_k\},\{\theta_k\}) = P(\beta|\gamma) \times \prod_{i=1}^{N} P(z_i|\beta) \cdots$$
$$\times \prod_{k=1}^{N} \big[(1 - \mathbf{I}_{X_k=\emptyset}) P(Y_k|X_k,\theta_k) P(\theta_k|\Lambda) P(X_k|\mu_k,\Sigma_k) \cdots$$
$$P(\mu_k,\Sigma_k|S_0,\nu_0,\mu_0,\kappa_0) + \mathbf{I}_{X_k=\emptyset} D_0(\mu_k,\Sigma_k,\theta_k)\big], \tag{7}$$

where $\mathbf{I}_{X_k=\emptyset}$ is the indicator function which has a value of one when the set X_k is empty and zero otherwise. $D_0(\cdot)$ is a delta function on a dummy set of parameters to ensure proper normalization. The individual probability distributions in (7) are given by

$$\begin{aligned} \beta \mid \gamma &\sim \text{GEM}(\gamma), & z_i \mid \beta &\sim \beta, \\ \Sigma_k &\sim \text{IW}(S_0, \nu_0), & \mu_k \mid \Sigma_k &\sim \text{N}(\mu_0, \Sigma_k/\kappa_0), \\ X_k \mid \mu_k, \Sigma_k &\sim \text{N}(\mu_k, \Sigma_k), & \theta_k &\sim H(\Lambda), \\ Y_k \mid X_k, \theta_k &\sim \text{N}(m(X_k), K(X_k, X'_k)|\theta_k). \end{aligned}$$

The input space is partitioned into separate regions using a DP prior, where each region is dominated by a GP expert with unique hyperparameter specifications. As a result, data that require non-stationary covariance functions, multimodal outputs, or discontinuities can be modeled.

HDDP-MoGP

A drawback with the DP-MoGP approach is that each GP expert for cluster k acts *locally* over an area in the input space defined by the Gaussian, (μ_k, Σ_k). Each cluster will almost surely have a distinct GP associated to it. Consequently, strong clustering in the input space can lead to several GP expert components even if a single GP would do a good job of modeling the data. In addition, local GP experts prevent information from being shared across disjoint regions in the input space.

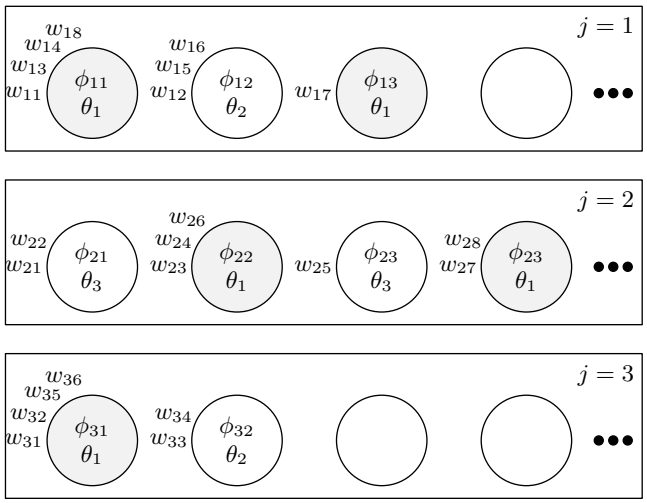


Figure 3: A depiction of a Chinese restaurant franchise with preferred seating. Each restaurant, j, is denoted by a rectangle. Customers (w_{ji}'s) are seated at tables (circles) in the restaurants. At each table two dishes are serverd. One of the dishes is served from a global menu (θ_k) which can be served on other tables within the restaurant or in other restaurants (ex. shaded gray circles all share the same global dish, θ_1). The other dish is custom made for the table (ϕ_{jt}) and is unique to that table-restaurant.

To address these shortcomings, we extend the DP-MoGP model by using a separate infinite mixture distribution for the input density of each expert. This allows GP experts to own complex non-local regions in the input space. In order to share GP experts over an infinite mixture input distribution, we incorporate an augmented version of a hierarchical DP prior, which we describe below.

Hierarchical Double Dirichlet Process

We extend a standard HDP to have local parameters specific to each cluster in addition to global parameters that are shared. We call this a hierarchical double DP (HDDP) prior and is shown in Fig. 1d. Here, each table draws a set of parameters, $\theta_{k_{jt}} \sim G_j$ from atoms of the top level base prior, allowing identical values θ_k, as in a standard HDP (see equation 2). In addition, table specific parameters $\phi_{t_{ji}} \sim L(\omega)$ are sampled from another base prior at the bottom level DP. The local parameters, $\phi_{t_{ji}}$, are sampled directly from a continuous prior, thus almost surely having unique values. An observation w_{ji} corresponds to a unique global component θ_k *and* a local component ϕ_{jt}, so $w_{ji} \sim F(\theta_k, \phi_{jt})$.

The generative process can be described as a Chinese restaurant franchise *with preferred seating* (see Fig. 3). In this case, the metaphor of the Chinese restaurant franchise is extended to allow table specific preferences *within* a restaurant. In addition to ordering from a global menu, which is shared across restaurants in a franchise, each table requests a custom dish that is unique to that table-restaurant. A key difference from a standard CRF description is the ability to distinguish tables within a restaurant that are served the same global dish. For instance, in the example depicted in Fig. 3, restaurant $j = 1$ has two tables that serve the global dish θ_1, however, because they have unique table specific parameters, ϕ_{11} and ϕ_{13}, these tables can be differentiated. Thus, in the case of a *single* restaurant, the HDP prior can be used to serve the *same* global dish to *different* tables. The prior probability customer w_{ji} sits at a table is given by (3) and the posterior probability of sitting at a table is weighted by the joint global and table specific likelihood (for example, see equations 11 and 12 in the inference procedure).

HDDP-MoGP

We use an HDDP prior for a mixture of Gaussian processes, which allows GP experts to be shared non-locally across the input space. We call this the HDDP-MoGP. In the HDDP prior, GP hyperparameters correspond to globally shared parameters and input Gaussians correspond to local parameters specific to each cluster.

Fig. 5 illustrates an example of HDDP-MoGP using the *CRF with preferred seating* metaphor. For notational simplicity, we consider the case of a single restaurant, i.e. $J = 1$, and therefore drop the j index. One could consider $J > 1$ when it is desired to detect identical GPs both within a restaurant and across a collection of restaurants, for example volatility regimes in different stocks or identical terrains in different images. Note, even though we have only one restaurant, we still require the HDP aspect of the HDDP prior to allow sharing of global menu items (GP experts, θ_k) among different tables (clusters) in the restaurant. In the example of Fig. 5, we have three tables occupied, which correspond to three unique input Gaussian distributions, ϕ_t, $t = 1,2,3$. Tables $t = 1$ and $t = 3$ share GP expert $k = 1$, specified by a set of GP hyperparameters, θ_1. Fig. 4 depicts a corresponding division of the input-space in two-dimensions, where the shaded regions own the same GP expert, θ_1.

More specifically, observations (points) belong to a table (cluster). Each table, t, is associated with a local Gaussian in the input space, $\phi_t = (\mu_t, \Sigma_t)$ and a globally shared

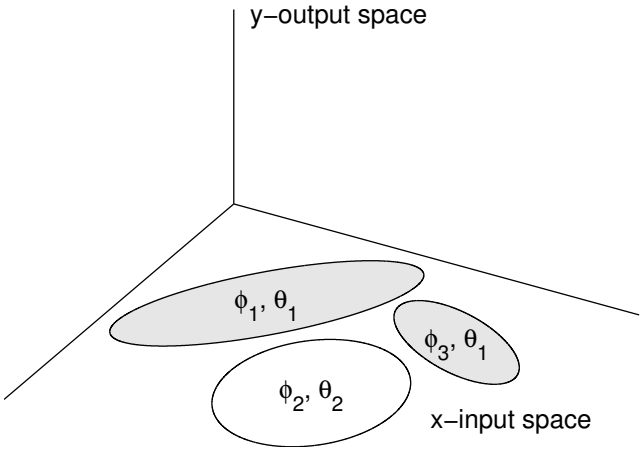


Figure 4: An example allotment of input space in the HDDP-MoGP model. Each elliptical region represents a two-dimensional Gaussian distribution. The shaded regions are owned by the same GP expert, illustrating a mixture distribution as the input density.

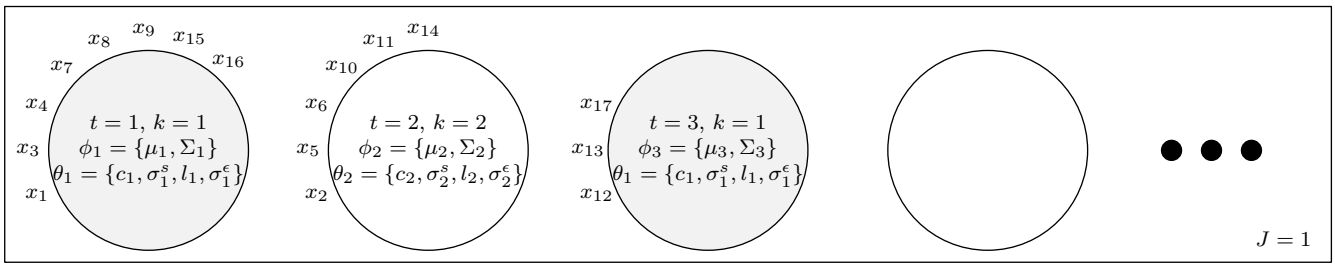


Figure 5: An example configuration of input data for HDDP-MoGP model using the CRF with preferred seating metaphor. Points (x_i's) are seated at tables (circles), which represent input Gaussian clusters. Each table is served a table specific dish, ϕ_t, which corresponds to a local input Gaussian, and a globally shared dish, θ_k, which corresponds to a GP expert specification. We consider only one restaurant (rectangle). The HDP mechanism is necessary to ensure that distinct tables can be served the same global GP expert, θ_k, from the top-level base prior.

GP expert, k, specified by a set of GP hyperparameters, θ_k. We sample GP hyperparameters from the top level DP and Gausssian distributions from the bottom level DP, in an HDDP prior. Thus, GP experts are shared among tables, while each table corresponds to a local input Gaussian. As a result, we have potentially infinite tables serving the same GP expert, representing an infinite Gaussian mixture distribution as the input density for each GP expert.

The generative process does not produce i.i.d. data points, and therefore we describe the process using a joint distribution over a dataset of a given size (as done for DP-MoGP). Fig. **??** shows the generative model for HDDP-MoGP. To construct a complete set of N sample points from the prior we would perform the following operations:

1. Construct a partition of the N observations into at most N tables (clusters) using Eq. (3). This assigns a table, t_i, $i = 1,...,N$, to each point in the set of observations. Once all table assignments are determined for the finite set of samples, we set $T = \max(\{t_i\})$, as the number of tables used.
2. For each table, sample a dish from the global shared menu using Eq. (4). This assigns an index corresponding to a GP expert, k_t, $t = 1,...T$, for each table. We set $K = \max(\{k_t\})$ as the number of GP experts.
3. For each table, sample an input Gaussian distribution (μ_t, Σ_t), $t = 1,...,T$ from a normal-inverse-Wishart prior with hyperparameters $\{S_0, \nu_0, \mu_0, \kappa_0\}$.
4. Given the input Gaussian distribution for each table (μ_t, Σ_t), sample the locations of the input points $X^t = \{x_i : t_i = t\}$.
5. For each GP expert, sample hyperparamters for the GP expert, $\theta_k \sim H(\Lambda)$.
6. For each GP expert, use the set of GP hyperparameters, θ_k, and input locations, $X_k = \{x_i : k_{t_i} = k\}$, which potentially spans multiple tables, to formulate the GP output mean vector and covariance matrix. Sample the set of output variables $Y_k = \{y_i : k_{t_i} = k\}$ according to joint Gaussian distribution given by (5).

Input data are conditionally i.i.d. given cluster t, while sets of output data are conditionally i.i.d. given respective input data corresponding to clusters that belong to GP expert, k.

The full joint distribution for N points can be expressed as,

$$
\begin{aligned}
&P(\{x_i\},\{y_i\},\{t_i\},\{k_t\},\{\mu_t,\Sigma_t\},\{\theta_k\},\pi,\beta) \\
&= P(\beta|\gamma)P(\pi|\alpha) \times \prod_{i=1}^{N} P(t_i|\pi) \cdots \\
&\times \prod_{t=1}^{N} \big[(1-\mathbf{I}_{X^t=\emptyset})P(X^t|\mu_t,\Sigma_t)P(\mu_t,\Sigma_t|S_0,\nu_0,\mu_0,\kappa_0) \cdots \\
&\qquad P(k_t|\beta) + \mathbf{I}_{X^t=\emptyset} D_0(\mu_t,\Sigma_t)\big] \cdots \\
&\times \prod_{k=1}^{N} \big[(1-\mathbf{I}_{Y_k=\emptyset})P(Y_k|X_k,\theta_k)P(\theta_k|\Lambda) + \mathbf{I}_{Y_k=\emptyset} D_0(\theta_k)\big],
\end{aligned}
\tag{8}
$$

where we have assumed at most N GP experts, $k = 1,...,N$, and at most N input clusters, $t = 1,...,N$, since we do not know K (total GPs) or T (total tables) beforehand, but at maximum they can be N of them. $\mathbf{I}_{X^t=\emptyset}$ and $\mathbf{I}_{Y_k=\emptyset}$ are indicator functions and $D_0(\cdot)$ is a delta function on a dummy set of parameters to ensure proper normalization. The individual distributions in (8) are given by:

$$
\begin{aligned}
\beta \mid \gamma &\sim \text{GEM}(\gamma), & k_t \mid \beta &\sim \beta, \\
\pi \mid \alpha &\sim \text{GEM}(\alpha), & t_i \mid \pi &\sim \pi, \\
\Sigma_t &\sim \text{IW}(S_0, \nu_0), & \mu_t \mid \Sigma_t &\sim \text{N}(\mu_0, \Sigma_t/\kappa_0), \\
X^t \mid \mu_t, \Sigma_t &\sim \text{N}(\mu_t, \Sigma_t), & \theta_k \mid \Lambda &\sim H(\Lambda), \\
Y_k \mid X_k, \theta_k &\sim \text{N}(m(X_k), K(X_k), X_k)')|\theta_k).
\end{aligned}
$$

The HDDP-MoGP model uses an HDDP prior with both local and shared parameters to define GP experts over an infinite Gaussian mixture in the input space. We remark, in the case of a single restaurant, i.e. $J = 1$, a nested or coupled set of urns (Beal, Ghahramani, and Rasmussen 2002) can functionally accomplish the role of HDP in the HDDP prior, though it would not be hierarchical in the Bayesian sense and the resulting inference procedure would become awkward (refer to Teh et al. 2006, for further discussion).

Gibbs sampler

We propose a Gibbs sampling algorithm for inference. To compute the likelihood of x_i, given $t_i = t$, and all other inputs, $X^{-i} = \{x_\imath : \imath \neq i\}$, and table assignments, $t = \{t_i\}$, we

form the set $X^{t^{-i}} = \{x_\iota : t_\iota = t, \iota \neq i\}$, and obtain a multivariate Student-t conditional posterior,

$$P(x_i|X^{t^{-i}}, S_0, \nu_0, \mu_0, \kappa_0) = t_{\hat{\nu}_t}(x_i; \hat{\mu}_t, \hat{\Sigma}_t) \triangleq g_t^{-i}(x_i), \tag{9}$$

$$n = |X^{t^{-i}}|, \quad \hat{\nu}_t = \nu_0 + n - d + 1,$$

$$\hat{\mu}_t = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\overline{X^{t^{-i}}},$$

$$S = \textstyle\sum_\iota (\{X^{t^{-i}}\}_\iota - \overline{X^{t^{-i}}})(\{X^{t^{-i}}\}_\iota - \overline{X^{t^{-i}}})^T,$$

$$\hat{\Sigma}_t = \frac{1}{\kappa_n \hat{\nu}_t}\left[S_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{X^{t^{-i}}} - \mu_0)(\overline{X^{t^{-i}}} - \mu_0)^T\right].$$

Here, we analytically marginalized $\{\mu_t, \Sigma_t\}$, since $x_i \sim \mathrm{N}(\mu_t, \Sigma_t)$ has unknown mean and covariance and we are using the normal-inverse-Wishart conjugate prior, $(\mu_t, \Sigma_t) \sim \mathrm{NIW}(S_0, \nu_0, \mu_0, \kappa_0)$ (Gelman et al. 2003).

For the likelihood of the output y_i, given $t_i = t$, all other inputs, X^{-i}, and outputs, $Y^{-i} = \{y_\iota : \iota \neq i\}$, table assignments, t, and table-dish assignments, $k = \{k_t\}$, we form the sets $X_{k_t}^{-i} = \{x_\iota : k_{t_\iota} = k_t, \iota \neq i\}$, $Y_{k_t}^{-i} = \{y_\iota : k_{t_\iota} = k_t, \iota \neq i\}$, and obtain a Normal conditional posterior,

$$P(y_i|x_i, X_{k_t}^{-i}, Y_{k_t}^{-i}, \theta_{k_t}) = \mathrm{N}\big(y_i|\bar{f}_*, \mathrm{cov}(f_*)\big) \triangleq f_{k_t}^{-i}(y_i), \tag{10}$$

$$\bar{f}_* = m(x_i) + \cdots K(x_i, X_{k_t}^{-i})[K(X_{k_t}^{-i}, X_{k_t}^{-i})]^{-1}(Y_{k_t}^{-i} - m(X_{k_t}^{-i})),$$

$$cov(f_*) = K(x_i, x_i) - \cdots K(x_i, X_{k_t}^{-i})[K(X_{k_t}^{-i}, X_{k_t}^{-i})]^{-1}K(X_{k_t}^{-i}, x_i).$$

This corresponds to the operation of conditioning the joint Gaussian prior for the GP, see Eq. (5), on the observations, i.e. the GP prediction formula (Rasmussen and Williams 2006). Note, all $m(\cdot)$ and $K(\cdot,\cdot)$ evaluations are conditioned on θ_{k_t}, which we have omitted for notational clarity.

Now, we can obtain the conditional posterior for t_i, given the remainder of the variables, by combining the conditional prior (3), with the likelihood of generating (x_i, y_i),

$$p(t_i|t^{-i}, k) \propto \begin{cases} n_t^{-i} g_t^{-i}(x_i) \cdot f_{k_t}^{-i}(y_i) & \text{if } t_i \text{ exists,} \\ \alpha p(x_i, y_i | t_i = t^{new}, t, k) & \text{if } t_i = t^{new}, \end{cases} \tag{11}$$

where $t = \{t_i\}$, $k = \{k_t\}$ and the superscript indicates that we exclude the variable with that index. Note, the posterior of a table assignment is weighted by table specific likelihood $g_t^{-i}(x_i)$. The likelihood of $t_i = t^{new}$ is obtained by summing out $k_{t^{new}}$,

$$p(x_i, y_i | t_i = t^{new}, t^{-i}, k) = g_{t^{new}}^{-i}(x_i)\left[\sum_{k=1}^{K} \frac{m_k}{m+\gamma} f_k^{-i}(y_i) + \frac{\gamma}{m+\gamma} f_{k^{new}}^{-i}(y_i)\right]. \tag{12}$$

Thus for a new table (input Gaussian), $t_i = t^{new}$, obtained from (11), we determine its dish (GP hyperparameters), $k_{t^{new}}$, from (12) according to,

$$p(k_{t^{new}} | t, k^{-t^{new}}) \propto \begin{cases} m_k f_k^{-i}(y_i) & \text{if } k \text{ exists,} \\ \gamma f_{k^{new}}^{-i}(y_i) & \text{if } k = k^{new}. \end{cases}$$

Similarly, the conditional posterior for k_t is obtained from (4) with likelihood $f_k^{-t}(Y_t)$, $Y_t = \{y_i : t_i = t\}$,

$$p(k_t = k | t, k^{-t}) \propto \begin{cases} m_k f_k^{-t}(Y_t) & \text{if } k_t \text{ exists,} \\ \gamma f_{k^{new}}^{-t}(Y_t) & \text{if } k_t = k^{new}. \end{cases}$$

For GP hyperparameters, θ_k, we do not have an analytical posterior, and thus resort to a hybrid MCMC algorithm as described in Rasmussen (1996) and Neal (1997). Alternatively, we can optimize the marginal likelihood of the GP using gradient descent as described in Rasmussen and Williams (2006) and use the obtained estimate as a sample from the posterior, generally leading to faster mixing rates.

Experiments

Simulated Data

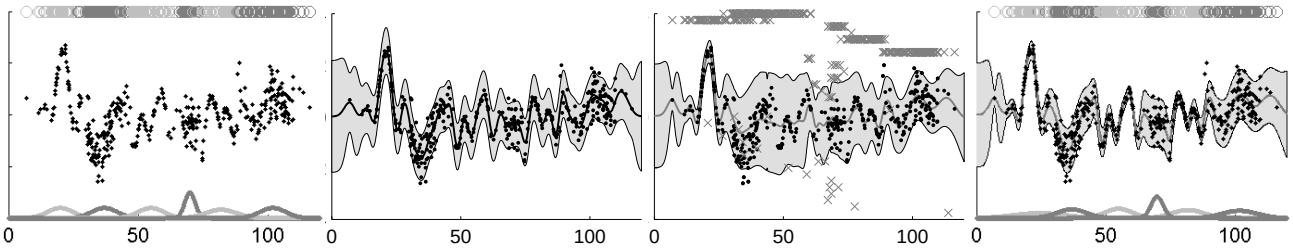
We test the HDDP-MoGP model on simulated data shown in Fig. 6a. It is a two GP mixture with six Gaussian input regions, described in Fig. 6e. GP mean and covariance functions are given by (6), with $c = 0$, $\sigma^s = 1$ fixed. The input Gaussians are shown at the bottom of Fig. 6a, where the shading indicates the GP. At the top, we also indicate which GP is used to sample each point.

Fig. 6b-d show results using a single GP, DP-MoGP, and HDDP-MoGP models, respectively. We use the following priors for the GP hyperparameters (see Rasmussen 1996, for discussion),

$$l^{-1} \sim \mathrm{Ga}\left(\frac{a}{2}, \frac{a}{2\mu_l}\right), \quad \log(\sigma^{s2}) \sim \mathrm{N}(-1, 1), \quad \log(\sigma^{\varepsilon 2}) \sim \mathrm{N}(-3, 3^2), \quad c \sim \mathrm{N}(0,\ \mathrm{std}(Y)). \tag{13}$$

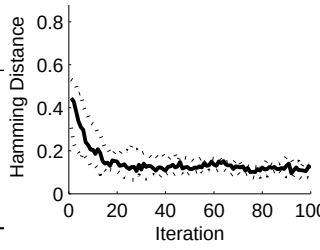
Here, length-scale, l, has an inverse Gamma distribution with $E(l^{-1}) = \mu_l$ and small a produces vague priors. We use $a = 1$ and $\mu_l = 1$ as recommended.

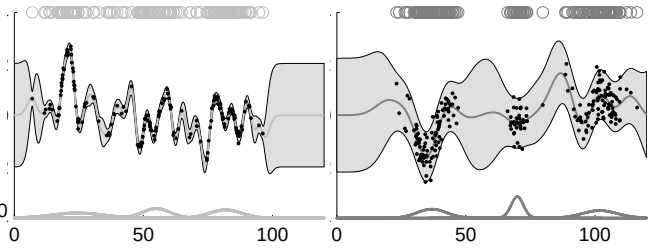
Fig. 6f plots the median (black), 10th and 90th percentiles (gray) of the Hamming distance between the estimated and true labels at each iteration of Gibbs sampling. These are obtained from ten different initializations during HDDP-MoGP training. We observe the Hamming distance stabilizes fairly quickly at about the 30th iteration. Fig. 6g-h illustrate the two GP components learned. The estimated parameters are shown in Fig. 6e, indicating that HDDP-MoGP successfully recovers the generating model. In comparison, we observe the single GP does not adjust for varying length-scale or noise, as expected. DP-MoGP, on the other hand, discovers too many GPs—five main clusters and several smaller ones. DP-MoGP cannot identify common GP components that are non-local, and in addition, as some input areas have fewer points, it struggles to reliably estimate the GP specification, for example, missing the fourth input Gaussian cluster altogether. We find that HDDP-MoGP improves identifiability by sharing statistical strength between non-local regions in the input space. Specifically, it is the hyperparameters of the GPs that are learned and shared by several regions. Generally evidence in separate regions have little effect on the belief about the placement of the underlying curve in other regions because of their distance. However, here the goal is to discover regions of similar smoothness



(a) Ground truth (2 GPs) (b) Single GP (c) DP-MoGP (d) HDDP-MoGP

	l_k	σ^{ε}_k	$\hat{l}_k$	$\hat{\sigma}^{\varepsilon}_k$
GP 1	2	0.1	1.95	0.10
GP 2	5	0.5	5.20	0.47
	μ_t	Σ_t	$\hat{\mu}_t$	$\hat{\Sigma}_t$
GP 1	20	5	24.5	10.1
GP 2	37	5	36.8	5.7
GP 1	55	5	55.1	5.4
GP 2	70	2	69.9	2.4
GP 1	82	6	81.8	6.2
GP 2	102	5	101.9	6.8





(e) Model parameters: True vs. HDDP-MoDP (f) HDDP-MoGP Gibbs iterations (g) HDDP-MoGP $k = 1$ (h) HDDP-MoGP $k = 2$

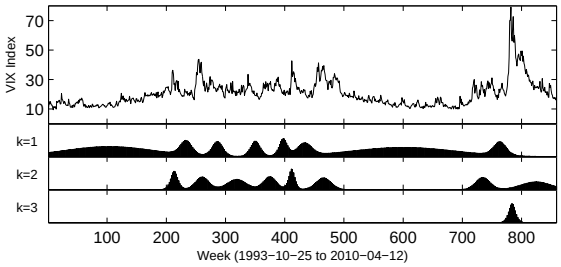
Figure 6: Comparing single GP, DP-MoGP and HDDP-MoGP on (a) simulated data generated using two GPs and six input Gaussians. (b),(c),(d) show the learned models. Where applicable, input Gaussians are shown at the bottom and labels at top, differentiated by shade/displacement. (b) Single GP does not adapt to non-stationarity, (c) DP-MoGP fails to identify GPs correctly, while (d) HDDP-MoGP benefits from sharing GP hyperparameters across distant regions to recover the model. (e) shows the parameters learned using HDDP-MoGP and (f) shows convergence by measuring Hamming distance between true and estimated labels. (g), (h) show the individual GPs obtained from HDDP-MoGP covering non-local input regions.

or noise regimes, so even if regions are far from each other, evidence about the hyperparameters of the GP in one region will have impact in other regions where the same GP dominates.

VIX Index

VIX index is a widely followed index, calculated by the Chicago Board Options Exchange from S&P 500 index option prices. It is considered a proxy of expected future volatility of the stock market, which is ascribed to have a mean reverting behaviour. However, it has proven to be challenging to model because both the mean and rate of reversion are non-stationary. Thus we consider the use of HDDP-MoGP to model non-stationary characteristics of the VIX Index.

We use (6) for GP mean and covariance functions, attempting to capture a smooth varying latent value of volatility. Fig. 7a shows the input Gaussians obtained using HDDP-MoGP with (13) as priors. The model identifies three GP regimes listed in Fig. 7b. The first GP, $k = 1$, corresponds to a regular market environment, with average volatility of 18.2%. The second GP, $k = 2$, represents a more uncertain environment, with average volatility at 28.3% and larger noise. This regime covers periods that include the Asian Cri-



(a)

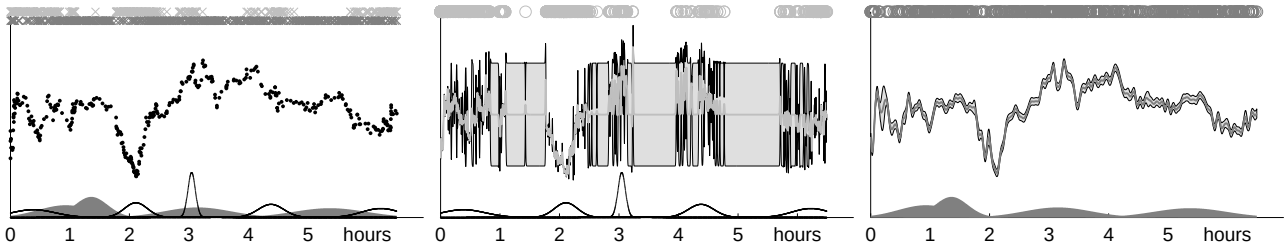
	GP		
k	1	2	3
c	18.2	28.3	42.1
σ^s	5.2	6.1	17.1
l	5.7	3.5	2.1
σ^{ε}	1.3	2.7	6.4

(b)

Model	*MSE*	$\overline{LL}$	$\overline{\sigma}^2$
Single GP	18.6	-2.1	3.6
DP-MoGP	18.9	-1.0	11.2
HDDP-MoGP	16.5	-0.8	10.4

(c)

Figure 7: (a) Top: VIX Index, Bottom: Input Gaussians for each GP learned using HDDP-MoGP. (b) Show hyperparameters for each GP. (c) Compares prediction quality using a rolling window. HDDP-MoGP results in the lowest mean square error and highest log likelihood.



(a) Tick data for RIM stock on 2-May-2007 (b) GP $k = 1$ (c) GP $k = 2$

Figure 8: Two recurring market behaviors in asynchronous high-frequency tick data learned using the HDDP-MoGP model. (a) shows tick data superimposed with HDDP-MoGP model. (b), (c) show individual GPs. (b) GP 1 captures sharp price movements reflecting quick buy-in/sell-off periods, while (c) GP 2 corresponds to business-as-usual with a relatively smooth price evolution.

sis of 1997, Dot-com bust and 9/11 of 2001 and the Euro-debt crisis of early 2010. Finally, the third GP, $k = 3$, reflects extremely unusual market conditions, with large average volatility and noise, occurring for a short while only in the Fall of 2008 as the collapse of Lehman Brothers nearly brought markets to a halt.

We also compare the predictive performance of HDDP-MoGP, DP-MoGP and single GP, using a rolling window approach. Mean squared error (*MSE*), average log likelihood ($\overline{LL}$) and average prediction uncertainty ($\bar{\sigma}^2$) for each model are shown in Fig. 7c. The single GP tends to be overly cautious in regular market environments, while not sufficiently accounting for highly volatile periods, resulting in higher error and lower likelihood. DP-MoGP improves likelihood, but still has a high MSE, since it cannot determine shifting regimes quickly and accurately enough. HDDP-MoGP obtains the highest likelihood with the lowest MSE as it can reuse regimes seen in the past.

High-Frequency Tick

We use HDDP-MoGP to model high-frequency tick data for RIM stock over a period of a day. Tick data arrives asynchronously with unequal time lapses. High-frequency data consists of short bursts of heightened activity followed by periods of relative inactivity. We consider using the HDDP-MoGP to identify these behaviors from price action data directly. Standard econometric approaches require preprocessing the data, for example by sampling at a fixed frequency, which can result in loss of information.

We use (6) for the GP functions and normalize prices to zero mean and unit standard deviation ($c = 0$, $\sigma^s = 1$). Results using HDDP-MoGP are shown in Fig. 8. The model discovers two GP components. In the first one ($k = 1$), length scale is $l_1 = 23.2$ and noise $\sigma^{\varepsilon}{}_1 = 0.02$. The second ($k = 2$), has length scale $l_2 = 183.5$ and noise $\sigma^{\varepsilon}{}_2 = 0.06$. The second GP is relatively smooth and distributed widely, reflecting business-as-usual price behavior. We occasionally see the emergence of the first GP, which marks much more sharp and directed price movements, suggesting quick sell-off or buy-in activity. We see that HDDP-MoGP can identify recurring classes of market behavior, unlike single GP or DP-MoGP, and without any preprocessing of the data.

Discussion

We have presented an infinite mixture model of GPs that share mixture components non-locally across the input domain through a hierarchical double Dirichlet process (HDDP-MoGP). An hierarchical double Dirichlet process (HDDP) extends a standard HDP by incorporating local parameters for each cluster in addition to globally shared parameters. The HDDP-MoGP model inherits the strengths of a Dirichlet process mixture of GPs (Meeds and Osindero 2006), but defines GPs over complex clusters in the input space formed by an infinite mixture of Gaussians. Experiments show the model improves identifiability by sharing data between non-local input regions and can be useful for clustering similar regions in the data. An interesting direction for future work is to consider other forms of input distributions, for instance hidden Markov models.

References

Beal, M. J.; Ghahramani, Z.; and Rasmussen, C. E. 2002. The infinite hidden Markov model. In *NIPS*.

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 2003. *Bayesian Data Analysis*. 2nd edition.

Gramacy, R. B., and Lee, H. K. H. 2008. Bayesian treed Gaussian process models with an application to computer modeling. *J. Amer. Statistical Assoc.*

Meeds, E., and Osindero, S. 2006. An alternative infinite mixture of Gaussian process experts. In *NIPS*.

Neal, R. M. 1997. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Dept. of Statistics, U. of Toronto.

Paciorek, C., and Schervish, M. 2003. Nonstationary covariance functions for Gaussian process regression. In *NIPS*.

Rasmussen, C. E., and Ghahramani, Z. 2002. Infinite mixtures of Gaussian process experts. In *NIPS*.

Rasmussen, C. E., and Williams, C. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Rasmussen, C. E. 1996. *Evaluations of Gaussian Processes and Other Methods for Non-Linear Regression*. Ph.D. Dissertation, U. of Toronto.

Sethuraman, J. 1994. A constructive definition of Dirichlet priors. *Statist. Sinica*.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *J. Amer. Statistical Assoc.*