Dirichlet labeling and hierarchical processes for clustering functional data

XuanLong Nguyen

Department of Statistics
University of Michigan

Joint work with Alan Gelfand, Duke University
Functional data clustering

- Many applications involve data that can be viewed as functions (e.g., curves, surfaces, multivariate functions)
  - borrowing statistical information across spatial domain
  - borrowing across replicates in the collection
Functional data clustering

- Many applications involve data that can be viewed as functions (e.g., curves, surfaces, multivariate functions)
  - borrowing statistical information across spatial domain
  - borrowing across replicates in the collection

- Model-based clustering using nonparametric labeling processes
  - global/local clustering and partitioning of curves
  - application to progesterone hormone analysis
  - application to image analysis
Functional data clustering

• Many applications involve data that can be viewed as functions (e.g., curves, surfaces, multivariate functions)
  – borrowing statistical information across spatial domain
  – borrowing across replicates in the collection

• Model-based clustering using nonparametric labeling processes
  – global/local clustering and partitioning of curves
  – application to progesterone hormone analysis
  – application to image analysis

• Variational Bayesian inference for labeling processes, i.e., implicit optimization over the space of posterior distributions
Example 1: Clustering curves

Progesterone data

- To characterize hormone levels in terms of typical behaviors
  - global clustering of entire curves; local clustering on a temporal segment
  - effects of contraception, detection of anomalous behavior
Example 2: Natural images

- image segmentation
- image ranking (retrieval from databases), image clustering
Example 3 – a little twist: Learning functional clusters from non-functional data

- data are daily hormone levels from a number of un-identified women (not necessarily realizations of curves)
  - due to confidentiality or lack of information, hormone levels collected different days may belong to different (groups of) women
- local clusters = typical daily hormone behaviors in a menstrual cycle
- global (functional) clusters = typical monthly hormone behaviors
Talk outline

- Global clusters and local clusters in functional data
  - Dirichlet labeling process for label allocation
  - variational inference
  - clustering trajectories and image segmentation
Talk outline

• Global clusters and local clusters in functional data
  – Dirichlet labeling process for label allocation
  – variational inference
  – clustering trajectories and image segmentation

• Functional clustering from (possibly) non-functional data
  – applications to data with partial information (e.g., confidentiality constraints)
  – nested hierarchy of Dirichlet processes mixture
Global clustering via mixture modeling

- Curves are functions defined on domain $D$
- Specify a mixing distribution over the space of entire curves

\[
\theta \sim G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j^*}
\]

\[
\theta_j^* \sim G_0
\]

- proportions $\{\pi_j\}$ is drawn from a “stick-breaking” prior (Gelfand, Kottas, MacEachern, JASA 2005)
- this allows “global” clustering, but not local clustering!
Local clustering

\[ G_x = \sum_{j=1}^{\infty} \pi_j(x) \delta_{\theta_j^*}(x) \]

- At each location \( x \in D \), specify a mixing distribution for the marginal \( G_x \) of curve value \( Y(x) \):

- spatial coupling between \( G_x \)  
  (MacEachern, 2000; Teh, Jordan, Blei, Beal, 2006; Dunson & Park, 2006; Griffin & Steel, 2006; etc)

- there is no global clustering!
Our modeling approach

- entire collection described by a number of “canonical” curves

- each curve can be viewed as a “hybrid” species
  (Duan et al, 2005; Petrone et al, 2007)

- canonical curve allocation via a labeling process

- “hybrid” view also popular in other contexts, e.g., population genetics and text analysis (Prichard et al, 2001; Blei et al, 2003)
$Y(x), \ x \in D \quad L(x), \ x \in D$

- $n$ replicates $Y_1, \ldots, Y_n$ observed at $m$ locations $x_1, \ldots, x_m$
- $k$ canonical curves $\theta^*_j \sim iid \ G_0, \ j = 1, \ldots, k$
\[ Y(x), \ x \in D \quad \text{and} \quad L(x), \ x \in D \]

- \( n \) replicates \( Y_1, \ldots, Y_n \) observed at \( m \) locations \( x_1, \ldots, x_m \)
- \( k \) canonical curves \( \theta_j^{*} \overset{iid}{\sim} G_0, \ j = 1, \ldots, k \)
- random label functions \( L_i : D \rightarrow \{1, \ldots, k\} \)
  \[ L_i | p \overset{iid}{\sim} p, \ i = 1, \ldots, n \]
\( Y(x), \ x \in D \) \quad \quad \quad \quad  \( L(x), \ x \in D \)

- \( n \) replicates \( Y_1, \ldots, Y_n \) observed at \( m \) locations \( x_1, \ldots, x_m \)
- \( k \) canonical curves \( \theta^*_j \overset{iid}{\sim} G_0, \ j = 1, \ldots, k \)
- random label functions \( L_i : D \to \{1, \ldots, k\} \)
  \[
  L_i | \ p \overset{iid}{\sim} p, \ i = 1, \ldots, n
  \]
- given \( \theta^* \) and \( L \), hybrid curve \( \theta_i \) satisfies: \( \theta_i(x) = \theta^*_{L_i}(x) \)
\(Y(x), \ x \in D\)

\(L(x), \ x \in D\)

- \(n\) replicates \(Y_1, \ldots, Y_n\) observed at \(m\) locations \(x_1, \ldots, x_m\)
- \(k\) canonical curves \(\theta^*_j \sim G_0, \ j = 1, \ldots, k\)
- random label functions \(L_i : D \rightarrow \{1, \ldots, k\}\)
  \(L_i | p \sim p, \ i = 1, \ldots, n\)
- given \(\theta^*\) and \(L\), hybrid curve \(\theta_i\) satisfies: \(\theta_i(x) = \theta^*_{L_i}(x)\)
- mixing with a pure error process
  \(Y_i(x) | \theta_i(x) \sim N(\theta_i(x), \tau^2)\)
Labeling process $L$

- $q$ is probability measure on $L \in \{1, \ldots, k\}^D$
  - Desideratum: $q$ allows spatial dependency of labels within the same curve
Labeling process $L$

- $q$ is probability measure on $L \in \{1, \ldots, k\}^D$
  - Desideratum: $q$ allows spatial dependency of labels within the same curve

- Draw $\eta$ from a mean zero Gaussian process with covariance function
  \[ \rho(x_1, x_2) = \exp\{-\phi \|x_1 - x_2\|\}, \]
  where $\phi$ is the scale parameter

- Thresholding $\eta$ to obtain $L \sim q$
Random measure $p$ on labeling functions $L$

- Want $p$ to be a random probability measure, whose mean measure is the measure $q$ defined above
  - Desideratum: random $p$ facilitates labeling sharing across curves

![Diagram showing C.I. for $q$ and a $p$ realization]
Random measure $p$ on labeling functions $L$

- Want $p$ to be a random probability measure, whose mean measure is the measure $q$ defined above
  - Desideratum: random $p$ facilitates labeling sharing across curves

- **Definition:** $p$ is a random measure generated by a Dirichlet process with base measure $q$ and concentration parameter $\alpha$

  $$p \sim DP(\alpha q)$$

  $$L|p \sim p$$
**Definition:** \( p \) is a random measure generated by a Dirichlet process with base measure \( q \) and concentration parameter \( \alpha \)

\[
p \sim DP(\alpha q)
\]

\[
L|p \sim p
\]

- i.e., for arbitrary \( x_1, \ldots, x_m \), \( L \) takes \( k^m \) possible values, whose associated probability is given by \( k^m \)-dim vector \( p_{x_1,\ldots,x_m} \)
  - \( p_{x_1,\ldots,x_m} \) has a \( k^m \)-dim Dirichlet distribution:

\[
(p_{x_1,\ldots,x_m}(\cdot)) \sim Dir(\alpha q_{x_1,\ldots,x_m}(\cdot)),
\]

where \( q \) is the probability measure on \( \{1, \ldots, k\}^D \).
Illustration of label properties

- *Spatial dependence of labels* is driven by $\phi$

  Posterior distribution mode of the labels for the leftmost image

  $$\phi = 0.5, 0.05, 0.01$$
Illustration of label properties

- **Spatial dependence of labels** is driven by $\phi$
  
  Posterior distribution mode of the labels for the leftmost image
  
  $\phi = 0.5, 0.05, 0.01$

- **Local clustering**: Let $L_1, L_2 | p \sim p$. Then, unconditionally,

  $$P(L_1(x) = L_2(x)) = \frac{1}{\alpha + 1} + \frac{1}{k} \cdot \frac{\alpha}{\alpha + 1}.$$

- **Global clustering**:

  $$P(L_1 = L_2) = \frac{1}{\alpha + 1}.$$
Hierarchical model behavior and identifiability

- interplay of three distributions
  - canonical curve distribution $G_0$, labeling process $p$ (driven by $\phi$), and pure error process (driven by $\tau$)

- additional constraints/priors available depending on modeling goals
  - inference of canonical curves (e.g., hormone analysis):
    $\Rightarrow$ roles of $\phi$ and $\tau$
  - inference of labels (e.g., image segmentation):
    $\Rightarrow$ roles of $G_0$
  - predictive posterior inference (e.g., on new locations/curves)
Model fitting via MCMC

• Gibbs updates of canonical curves are standard

• Updates of precision parameters $\alpha, \tau$ are also standard
  (Escobar & West, 1995)

• The most computationally intensive part is the update of labels $L_1, \ldots, L_n$
  and label switching parameter $\phi$
  – the sampler cannot handle this step exactly
  – the challenge lies in the (latent) high dimensional thresholded Gaussian process $q$
    nested within the Dirichlet process

• We propose a variational inference method to obtain approximate posterior distribution for labels $L$ and label parameter $\phi$
Variational Bayesian inference

- The core of Bayesian inference is the posterior computation:

\[
\text{posterior} \propto \text{likelihood} \times \text{prior}
\]

\[
w(\theta|X_1, \ldots, X_n) \propto \prod_{i=1}^{n} P(X_i|\theta)\pi(\theta)
\]

- \(w(\cdot|X)\) is the solution to the following optimization problem:

\[
w(\cdot|X_1, \ldots, X_n) = \arg\min_{w(\cdot)} - \int \sum_{i=1}^{n} \log P(X_i|\theta)dw(\theta) + D(w||\pi),
\]

where \(D\) denotes the Kullback-Leibler (KL) divergence

- If inference is intractable, consider modifying either loss function or the space of distributions
Variational approximation for $q(L(x)|\phi, \text{Data})$

- Let $E$ be a subset of edges connecting pairs of locations $x_1, \ldots, x_m$.
- $\tilde{q}_E$ is the best approximation (in the sense of KL divergence) among all Markov random fields (MRF) $Q_E$ using pairwise potential functions

$$\tilde{q}_E = \arg\min_{Q \in Q_E} D(q||Q).$$
Variational approximation for $q(L(x)|\phi, \text{Data})$

Let $E$ be a subset of edges connecting pairs of locations $x_1, \ldots, x_m$

$\tilde{q}_E$ is the best approximation (in the sense of KL divergence) among all Markov random fields (MRF) $Q_E$ using pairwise potential functions

$$\tilde{q}_E = \arg\min_{Q \in Q_E} D(q||Q).$$

If $E_1 \supset E_2$ then $D(q||\tilde{q}_{E_1}) \leq D(q||\tilde{q}_{E_2})$

How to choose collection of edges $E$?

- e.g., shortest spanning tree for $x_1, \ldots, x_m$, resulting in linear computational complexity of inference (in $m$)
Variational approximation for $\phi|L, \text{Data}$

- Resultant posterior distribution for $\phi$:

$$P(\phi|L) \propto \prod_{(x_i, x_j) \in E} q(L(x_i), L(x_j)|\phi)^\lambda \pi(\phi)$$

$$= \arg\min_{\omega(.) \in W} \int_\phi \lambda \sum_{(x_i, x_j) \in E} -\log q(L(x_i), L(x_j))d\omega(\phi) + D(\omega||\pi)$$
Variational approximation for $\phi|L, \text{Data}$

- Resultant posterior distribution for $\phi$:

$$P(\phi|L) \propto \prod_{(x_i, x_j) \in E} q(L(x_i), L(x_j)|\phi)^{\lambda} \pi(\phi)$$

$$= \arg\min_{w(\cdot) \in W} \int_{\phi} \lambda \sum_{(x_i, x_j) \in E} -\log q(L(x_i), L(x_j)) dw(\phi) + D(w||\pi)$$

- $P(\phi|L)$ shrinks toward the true $\phi^*$ in an appropriately defined metric:

**Proposition 1.** Suppose that $L = (L(x_1), \ldots, L(x_m))$ is drawn from $q$ for some "true" $\phi^* > 0$; and that $\pi(\cdot)$ puts sufficient mass around $\phi^*$ then under the distribution $q$

$$\int \frac{1}{|E|} \left| \log(P(\phi^*|L)/P(\phi|L)) \right| dP(\phi|L) = O_P(1/m).$$

- Posterior mode converges in probability to $\phi^*$ at the rate $\sqrt{\frac{r_m}{m}}$, where $r_m$ is the number of edges whose lengths are bounded by a fixed constant
Predictive posterior distributions

- a simulated example on 1-dim domain $D = [1, 40]$
Progesterone hormone data

- Inference of canonical curves
  (– contraception grouping is *unknown* to the analysis)

- Left figure: Distinctive canonical behaviors emerge approximately 7 days after ovulation date

- Right figure: $\phi$ is set very small to insist on “global” clusters
  $\Rightarrow$ global clustering not possible!
Pairwise comparison of hormone curves

- Proportion of equal labels for pairs of replicates
  - left: for the whole curve
  - right: for a curve segment [20, 24]
    (last 5 days of the monitored cycle)

- Curves indexed from 67 to 88 belong to the contraception group
Detecting unusual hormone levels

- Curves numbered 12, 27, 30, 50, 74, 82, 85 have large degree of departure from the “mean” behavior
Image segmentation

- canonical curves are random constant functions
- image segments correspond to (random) level sets
Similar image retrieval

- For a given image (leftmost), list 5 most “similar” images

  "happy cows"

  "trees"

  "cars"

  "aircrafts"

  "buildings"

  "faces"
Summary

• Global clusters and local clusters in functional data
  – Dirichlet labeling process for label allocation
  – variational inference
  – clustering trajectory curves and image segmentation

• Dirichlet labeling processes provide a useful alternative to existing modeling frameworks, e.g., Markov random fields
  – label behavior easy to interpret (via latent hierarchies)
  – label behavior quite rich (via latent stochastic processes)
  – computationally tractable (compared to MRFs)

• Functional clustering from non-functional grouped data
  – applications in data with partial information (e.g., confidentiality constraints)
  – nested hierarchical Dirichlet processes
Recall – a little twist: Learning functional clusters from non-functional data

- data are daily hormone levels from a number of un-identified women (not necessarily realizations of curves)
  - due to confidentiality or lack of information, hormone levels collected different days may belong to different (groups of) women
- local clusters = typical daily hormone behaviors in a menstrual cycle
- global (functional) clusters = typical monthly hormone behaviors
Set-up of non-functional grouped data
Set-up of non-functional grouped data

- Non-functional data set-up: For group $u$, and data $y_{ui}$:
  
  $y_{ui} | \theta_{ui} \sim F(\cdot | \theta_{ui})$
  
  $\theta_{ui} \overset{iid}{\sim} G_u$ for $i = 1, 2, \ldots$

- $G_u$ is the mixing distribution for local clusters associated with group $u$
Set-up of non-functional grouped data

- Non-functional data set-up: For group \( u \), and data \( y_{ui} \):

\[
y_{ui} \mid \theta_{ui} \sim F(\cdot \mid \theta_{ui})
\]

\[
\theta_{ui} \overset{iid}{\sim} G_u \quad \text{for} \quad i = 1, 2, \ldots
\]

- \( G_u \) is the mixing distribution for local clusters associated with group \( u \)

- Modeling problem:
  - notion of global (functional) clusters that emerge when all groups aggregate
  - specification of distribution \( Q \) for global clusters
  - linking up \( Q \) to the non-functional data
Formalizing the notion of global clusters

- Let $\Theta_u$ be the space for local atom $\theta_u$ associated with group $u$
- Define product space $\Theta = \prod_{u \in V} \Theta_u$
- Global atoms $\phi := (\phi_u)_{u \in V}$ is an element in $\Theta$
- Modeling approach: hierarchical nonparametric Bayes
  - to specify a (random) distribution $Q$ for “global” $\phi$ (using Dirichlet process)
  - Linking up $Q$ to the mixing distribution of local clusters $G_u$’s via a nested hierarchy of Dirichlet processes
Nested hierarchical Dirichlet processes (nHDP)

- extending the hierarchical Dirichlet process (HDP) framework (Teh, Jordan, Blei, Beal, JASA 2006) to enable functional clustering
  - while HDP is a recursive hierarchy of Dirichlet processes (DPs), nHDP is a nested hierarchy of DPs

- key features of the nHDP approach:
  - global and local atoms are related as different levels in the Bayesian hierarchy
  - fully nonparametric specification
Hierarchical model specification

- Data: Let $y_{u1}, y_{u2}, \ldots, y_{un_u}$ be the observations obtained within group $u$, and assume that each observation is drawn independently from a mixture model:

$$y_{ui} | \theta_{ui} \sim F(\cdot | \theta_{ui})$$
$$\theta_{ui} | G_u \sim G_u$$

for any $u \in V; \ i = 1, \ldots, n_u$

- The $G_u$ are given by nested hierarchy of DPs:

$$G_u | \alpha_u, Q \sim \text{DP}(\alpha_u, Q_u), \text{ for all } u \in V$$
$$Q | \gamma, H \sim \text{DP}(\gamma, H),$$

- Base measure $H$ is taken to be a Gaussian process indexed by $u \in V$
Nested hierarchy of Dirichlet processes

\[ H \]

\[ Q \]

\[ G \]

\[ \theta \]
Stick-breaking representation

• $Q$ provides the support for global clusters:

$$Q = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where $\phi_k = (\phi_{uk} : u \in V)$ are independent draws from $H$, and $\beta = (\beta_k)_{k=1}^{\infty} \sim \text{GEM} (\gamma)$.

• $Q_u$ is the induced marginal of $Q$ at $u$, while $G_u$ centers around the $Q_u$, and provides the support for local clusters:

$$Q_u = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_{uk}}$$

$$G_u = \sum_{k=1}^{\infty} \pi_{uk} \delta_{\phi_{uk}}$$
Pólya-urn characterization

- Sampling of *local atoms* distributed by $G_u$ (which is integrated out):

$$\theta_{ui}|\theta_{u1}, \ldots, \theta_{u,i-1}, \alpha_u, Q \sim \sum_{t=1}^{m_u} \frac{n_{ut}}{i - 1 + \alpha_u} \delta_{\psi_{ut}} + \frac{\alpha_u}{i - 1 + \alpha_u} Q_u.$$  

- $Q_u$ is the induced marginal of distribution $Q$. 
Pólya-urn characterization

• Sampling of local atoms distributed by $G_u$ (which is integrated out):

\[
\theta_{ui} | \theta_{u1}, \ldots, \theta_{u,i-1}, \alpha_u, Q \sim \sum_{t=1}^{m_u} \frac{n_{ut}}{i - 1 + \alpha_u} \delta_{\psi_{ut}} + \frac{\alpha_u}{i - 1 + \alpha_u} Q_u.
\]

- $Q_u$ is the induced marginal of distribution $Q$

• Sampling of global atoms distributed by $Q$ (which is integrated out):

\[
\psi_t | \{\psi_l\}_{l \neq t}, \gamma, H \sim \sum_{k=1}^{K} \frac{q_k}{q} \delta_{\phi_k} + \frac{\gamma}{q + \gamma} H.
\]
Sharing lunch in a communal eatery

global atoms = lunch boxes = $\phi = (\phi_u)_{u \in V}$
local atoms = lunch items = $\phi_u$

$u$ indexes categories, e.g., appetizer or main entree

- lunch items are consumed based on popularity within its category
- lunch boxes are bought based on its popularity, and are to be shared by all

Statistical problem: Inference about distribution of lunch boxes bought based on distributions of lunch items consumed
Spatial dependence among $G_u$’s

- spatial dependence conferred by base measure $H$ entails spatial dependence among local distributions $G_u$’s
- suppose that $H$ is a Gaussian process, $\phi = (\phi_u : u \in V) \sim N(\mu, \Sigma)$, where $\Sigma$ takes standard exponential form

- $Corr(G_u(A), G_v(B)) \to 1$ as $u \to v$, and and vanishes as $\|u - v\| \to \infty$
- $Corr(G_u(A), G_u(B))$ increases as either $\alpha_u$ or $\alpha_v$ increases

Relations between the two levels in the Bayesian hierarchy:

- $Corr(G_u(A), G_v(B)) \leq Corr(Q_u(A), Q_v(B))$
- $\gamma$ ranges from 0 to $\infty$, correlation measure ratio
  
  $$Corr(G_u(A), G_v(B))/Corr(Q_u(A), Q_v(B))$$

  decreases from 1 to 0.
Posterior consistency and identifiability

• The nested Hierarchical Dirichlet process prior distribution placed on the space of joint density of \((y_u)_{u \in V}\) yields consistent posterior distributions

• Under additional assumptions on base measure \(H\), it can be shown that global atomss (i.e., distributed by \(Q\)) can be identified given collection of data locally grouped by index \(u\)
Posterior inference

• Hierarchical model is amenable to Gibbs sampling
  – sampling local atoms by integrating out $G_u$’s
  – sampling global atoms by integrating out $Q$, and possible base measure $H$

• Conditional distribution of DP-distributed measure is again a Dirichlet process

• Computational speedup is achieved by replacing the spatial process $H$ by a graphical model
  – inference for tree-structured or chain-structure model requires time linear in number of locations $u$
Exploiting stick-breaking representation

Construct a Markov chain on space of stick-breaking representations \((z, q, \beta, \phi)\).

Sampling \(\beta\): \(\beta|q \sim \text{Dir}(q_1, \ldots, q_K, \gamma)\).

Sampling \(z\):

\[
p(z_{ui} = k|z^{-ui}, q, \beta, \phi_k, \text{Data}) = \begin{cases} 
(n_{u,k}^- + \alpha_u \beta_k)F(y_{ui}|\phi_k) & \text{if } k \text{ previously used} \\
\alpha_u \beta_{\text{new}} f_{yk_{\text{new}}}^j(y_{ui}) & \text{if } k = k_{\text{new}}.
\end{cases}
\]

Sampling \(q\): \(q_k = \sum_{u \in V} m_{uk}\) where:

\[
p(m_{uk} = m|z, m^{-uk}, \beta) = \frac{\Gamma(\alpha_u \beta_k)}{\Gamma(\alpha_u \beta_k + n_{u,k})} s(n_{u,k}, m)(\alpha_u \beta_k)^m.
\]

Sampling \(\phi\):

\[
p(\phi_k|z, \text{Data}) \propto H(\phi_k) \prod_{ui: z_{ui} = k} F(y_{ui}|\phi_k) \text{ for each } k = 1, \ldots, K.
\]
Tracking example

Prior specification:

- concentration parameters $\gamma \sim \text{Gamma}(5, 1)$ and $\alpha \sim \text{Gamma}(20, 20)$
- variance $\sigma^2_\epsilon$ of $F(\cdot)$ is given prior $\text{InvGamma}(5, 1)$
- prior for global atoms $H$ is a mean-0 Gaussian Process using $(\sigma, \omega) = (1, 0.01)$
Clustering bifurcation behavior

- prior for global atoms $H$ is a mean-0 Gaussian Process using $(\sigma, \omega) = (1, 0.05)$
- other prior specifications are the same as previous data example
Inference of global clusters (tracks)

Left: Number of global clusters is 5 with > 90%
Right: (.05,.95) credible intervals of global cluster estimates
Global clusters of bifurcating behavior

Left: Number of global clusters is 3 with > 90%
Right: (.05,.95) credible intervals of global cluster estimates
Evolution of local clusters

Posterior distribution of the number of local clusters associating with different group index (location) \( u \).
Effects of vague prior for $H$

Very vague prior for $H$ results in weak identifiability of global clusters, even as the local clusters are identified reasonably well.
Clustering progesterone hormone

- Hormone levels collected from a number of women
- Subject ids are withheld, so hormone trajectories are *not* given
- Comparison to hybrid DP approach (Petrone et al, 2009), which does use the trajectorial information
Temporally varying number of local clusters

Number of global clusters:
Estimates of global clusters

Left: Clustering results using the nHDP mixture model
Right: The hybrid-DP approach of Petrone, Guindani and Gelfand (2009)

Black solids are sample mean curves of the contraceptive group and no-contraceptive group
Pairwise comparison of hormone curves

Each entry in the heatmap depicts the posterior probability that the two curves share the same local clusters, averaged over the last 4 days in the menstrual cycle.

nHDP approach (Left panel) provides sharper clusterings than the hybrid DP approach (Right panel)
data are (temp, depth) samples collected at 4 different locations at different times

- functional clustering within each location
- functional comparisons (ANOVA) between locations
Model formalization

- Data are collection of $Y_{ux}(i)$, for $i = 1, \ldots, n_{ux}$
- $u$ indexes group (location), $x$ indexes depth
- At each location, there are global clusters representing typical functional relationship between depth vs temp
- Modeling data associated with each location using a nHDP: for each $u = 1, \ldots, 4$:

$$\{\theta_{ux}(i)\} | \gamma, \alpha, H \overset{iid}{\sim} \text{nHDP}(\gamma, \alpha, H)$$

$$Y_{ux}(i) | \theta_{ux}(i) \overset{indep}{\sim} \text{N}(\theta_{ux}(i), \tau_u^2)$$

- Base probability measure $H$ is a draw from a Dirichlet process:

$$H \sim \text{DP}(\alpha_0, H_0),$$

$$H_0 \equiv \text{GP}(\mu_0, \Sigma_0)$$
Fully nonparametric hierarchical specification

\[ H_0 \equiv \text{GP}(\mu_0, \Sigma_0) \]
\[ H|H_0 \sim \text{DP}(\gamma, H_0), \]
\[ Q_u|H \sim^{iid} \text{DP}(\alpha_0, H), \text{ for all } u \in V \]
\[ G_{u;x}|Q_u \sim^{indep} \text{DP}(\alpha_u, Q_{u;x}). \]

As before,

\[ \theta_{ux}(i)|Q_{u;x} \sim Q_{u;x} \text{ for all } i = 1, \ldots, n_u; u \in V \]
\[ Y_{ux}(i)|\theta_{ux}(i) \sim N(\theta_{ux}(i), \tau_u^2) \text{ for all } i = 1, \ldots, n_u; u \in V. \]
Posterior distribution of global atoms

Mean/credible intervals of functional mean curves

\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6
Posterior mean/std of mixing proportions
of the dominant functional clusters for each group of data

<table>
<thead>
<tr>
<th>group ($u$)</th>
<th>$\pi_{u1}$</th>
<th>$\pi_{u2}$</th>
<th>$\pi_{u3}$</th>
<th>$\pi_{u4}$</th>
<th>$\pi_{u5}$</th>
<th>$\pi_{u6}$</th>
<th>$\pi_{u7}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98 (0.01)</td>
<td>0.00 (0)</td>
<td>0.00 (0)</td>
<td>0.0022 (0)</td>
<td>0.00 (0)</td>
<td>0.00 (0)</td>
<td>0.00 (0)</td>
</tr>
<tr>
<td>2</td>
<td>0.07 (0.20)</td>
<td>0.70 (0.16)</td>
<td>0.08 (0.05)</td>
<td>0.06 (0.03)</td>
<td>0.01 (0.02)</td>
<td>0.02 (0.04)</td>
<td>0.00 (0)</td>
</tr>
<tr>
<td>3</td>
<td>0.08 (0.24)</td>
<td>0.01 (0.02)</td>
<td>0.01 (0.02)</td>
<td>0.01 (0.02)</td>
<td>0.86 (0.24)</td>
<td>0.00 (0.01)</td>
<td>0.00 (0)</td>
</tr>
<tr>
<td>4</td>
<td>0.07 (0.23)</td>
<td>0.01 (0.02)</td>
<td>0.01 (0.03)</td>
<td>0.01 (0.02)</td>
<td>0.86 (0.22)</td>
<td>0.00 (0.01)</td>
<td>0.00 (0)</td>
</tr>
</tbody>
</table>

![Graph showing the posterior mean and std of mixing proportions for each group.](image-url)
Posterior mean of noise variance

<table>
<thead>
<tr>
<th>group ((u))</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean/std for (\tau_u)</td>
<td>0.50 (0.22)</td>
<td>1.15 (2.13)</td>
<td>2.13 (0.29)</td>
<td>1.88 (0.25)</td>
</tr>
</tbody>
</table>
Varying number of local clusters with depth

Posterior mean (solid) and (.05,.95) credible intervals (dash)
Summary

- Global clusters and local clusters in functional data
  - Dirichlet labeling process for label allocation
  - variational inference
  - clustering trajectory curves and image segmentation
  - appears in Nguyen & Gelfand (Statistica Sinica, 2011)

- Learning functional clusters from non-functional grouped data
  - applications in data with partial information (e.g., confidentiality constraints)
  - nested hierarchical Dirichlet processes
  - appears in Nguyen (Bayesian Analysis, 2010)

- This talk focuses mainly on modeling and computation
  - the identifiability and posterior consistency of nonparametric mixing distributions in hierarchical models are currently under investigation