Inference of functional clustering patterns from non-functional data

Long Nguyen

Department of Statistics
University of Michigan

Madison, April 2012
Talk outline

• Learning functional relationships, but functional data are unavailable
  – functional clustering
  – differs from (non-linear) regression
  – “co-clustering”, involving co-varying mixture distributions

• Hierarchical and nonparametric Bayesian method

• Intuitive computational algorithms for statistical inference
  – Markov Chain Monte Carlo sampling for co-clustering

• Asymptotic results for identifiability and consistency of latent mixing measures
• data are (temp, depth) samples collected at 4 different locations at different times in span of few days

• heterogeneous functional clustering patterns within each location
  – extracting functional clusters
  – interpolation
  – comparisons between groups associated with different locations
Simpler example:
Problem of tracking (connecting the dots)

- data are positions $Y \in \mathbb{R}^d$ of multiple objects moving in a geographical area (positions $Y$ co-vary with time $u$)
- objects move in local clusters (might switch over time)
  - we are not interested in the movement of each individual object; rather we are interested in the paths over which the local clusters evolve
- moving paths are functions of time
Example: Functional clustering without functional data

- data are daily hormone levels from a population sample
- hormone levels from different individuals for different days $u$
- interested in global/functional clusters for a typical individual in the population
A simple ad hoc computational heuristic

- this is viewed as a “co-clustering” problem
- collection of co-varying mixture distributions indexed by covariate $u$
- a heuristic:
  - solve each clustering problem individually
  - mix-match clusters from different mixture distributions
Our approach

- proposed a hierarchical nonparametric model that links “functional/global clusters” to “non-functional/local” data

- several modeling ingredients
  - assume smooth functional clusters using Gaussian process
  - use Dirichlet process mixtures to handle unknown number of clusters
  - probabilistic linkage achieved via conditional hierarchy
Background I: Dirichlet process mixtures

• Dirichlet process (DP) mixtures are natural for handling unknown number of mixing components
  – mixing distribution $G$ is random and distributed according to a DP
Background I: Dirichlet process mixtures

• Dirichlet process (DP) mixtures are natural for handling unknown number of mixing components
  – mixing distribution $G$ is random and distributed according to a DP

• A Dirichlet process $\text{DP}(\alpha_0, G_0)$ defines a distribution on (random) probability measures
  – $\alpha_0$ concentration parameter, $G_0$ centering distribution
Background I: Dirichlet process mixtures

• Dirichlet process (DP) mixtures are natural for handling unknown number of mixing components
  – mixing distribution $G$ is random and distributed according to a DP

• A Dirichlet process $\text{DP}(\alpha_0, G_0)$ defines a distribution on (random) probability measures
  – $\alpha_0$ concentration parameter, $G_0$ centering distribution

• A random draw $G \sim \text{DP}(\alpha_0, G_0)$ admits the “stick-breaking” representation w.p.1:
  
  \[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \]

  – $\delta_{\phi_k}$ denotes an atomic distribution concentrated at $\phi_k$, $\phi_k \overset{iid}{\sim} G_0$
  – stick-breaking weights $\pi_k$ are random and depend only on $\alpha_0$
Background II: Dependent Dirichlet processes

- DDPs modeling framework advocated by MacEachern (1999)
- modeling a collection of Dirichlet processes \{G_u\}: via stick-breaking representation:
  \[ G_u = \sum_{k=1}^{\infty} \pi_{uk} \delta_{\phi_{uk}} \]
  - for each \( u \in V \): \( \pi_{uk} \)'s are called “stick” variables; \( \phi_{uk} \) are “atoms”
  - for each \( k \): \( \pi_k = (\pi_{uk})_{u \in V} \) and \( \phi_k = (\phi_{uk})_{u \in V} \) are stochastic processes indexed by \( u \in V \)

- our problem presents some modeling challenges: nonparametric functional patterns without functional data
Background III: Hierarchical Dirichlet Processes

• HDPs modeling framework due to Teh, Jordan, Blei, Beal (JASA, 2006)

• hierarchy of \textit{recursively} specified Dirichlet processes:

\[
G_u | \alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0) \\
G_0 | \gamma, H \sim \text{DP}(\gamma, H)
\]

• note that $G_u$, $G_0$ and $H$ are probability measures on the same space of atoms

• but they are specified in different levels in the model hierarchy
Proposed approach

- A multi-level nonparametric Bayesian modeling approach:
  - we need a collection of dependent DP’s (as in DDPs)
  - also different Dirichlet processes in different levels (as in HDPs)

- key features:
  - a Dirichlet process for modeling functional atoms
  - Dirichlet processes for modeling local atoms (for each $u$)
  - global and local atoms are related as different levels in the conditional probability hierarchy
  - a *nested* hierarchy of Dirichlet processes (generalizing the HDP)
Some notations

• Data are $(y_{ui})$, indexed by $u \in V$, and $i = 1, \ldots, n_u$

• For each $u \in V$, observations $(y_{ui})_{i=1}^{n_u}$ are draws from a mixture distribution with mixing measure $G_u$ supported by $\theta_u$'s, where $\theta_u \in \Theta_u$
  – e.g., for mixture of gaussians, $\theta_u$'s are the means
Some notations

- Data are \((y_{ui})\), indexed by \(u \in V\), and \(i = 1, \ldots, n_u\)

- For each \(u \in V\), observations \((y_{ui})_{i=1}^{n_u}\) are draws from a mixture distribution with mixing measure \(G_u\) supported by \(\theta_u\)'s, where \(\theta_u \in \Theta_u\)
  - e.g., for mixture of gaussians, \(\theta_u\)'s are the means

- Define product space \(\Theta = \prod_{u \in V} \Theta_u\)

- A global (functional) atom \(\phi := (\phi_u)_{u \in V}\) is an element in \(\Theta\)

- \(\phi\) is random and distributed by mixing measure \(Q\), which varies around a smooth stochastic process \(H\) (e.g., Gaussian process)
Full model specification (nested HDP)  

(Nguyen, 2010)

• observations from each group indexed by $u$ are drawn independently from a mixture model:

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

for any $u \in V; \ i = 1, \ldots, n_u$
Full model specification (nested HDP)  

(Nguyen, 2010)

• observations from each group indexed by \( u \) are drawn independently from a mixture model:

\[
\begin{align*}
y_{ui} \mid \theta_{ui} & \sim iid \quad F(\cdot \mid \theta_{ui}) \\
\theta_{ui} \mid G_u & \sim iid \quad G_u
\end{align*}
\]

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

• probability distribution \( H \), which specifies centering distribution for global clusters, is taken to be a Gaussian process on \( \Theta \)

• mixing measures \( G_u \) are given a hierarchy of DPs:

\[
\begin{align*}
Q \mid \gamma, H & \sim \quad DP(\gamma, H), \\
G_u \mid \alpha_u, Q & \sim \quad DP(\alpha_u, Q_u), \text{ for all } u \in V
\end{align*}
\]
Nested hierarchy of Dirichlet processes
Statistical dependence among $G_u$’s

• the dependence conferred by centering distribution $H$ entails the 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

• for any measurable sets $A$ and $B$: 
Statistical dependence among $G_u$’s

- the dependence conferred by centering distribution $H$ entails the 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

- for any measurable sets $A$ and $B$:

$$\text{Corr}(G_u(A), G_v(B)) \to \begin{cases} 
0 & \text{as } \|u - v\| \to \infty \\
1 & \text{if } A = B, \|u - v\| \to 0 
\end{cases}$$
Statistical dependence among $G_u$’s

- the dependence conferred by centering distribution $H$ entails the 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

- for any measurable sets $A$ and $B$:

$$
\operatorname{Corr}(G_u(A), G_v(B)) \rightarrow \begin{cases} 
0 & \text{as } ||u - v|| \rightarrow \infty \\
1 & \text{if } A = B, ||u - v|| \rightarrow 0
\end{cases}
$$

- relations between the two levels in the Bayesian hierarchy: the correlation ratio

$$
\frac{\operatorname{Corr}(G_u(A), G_v(B))}{\operatorname{Corr}(Q_u(A), Q_v(B))}
$$

decreases from 1 to 0 as $\gamma$ ranges from 0 to $\infty$
Stick-breaking representation

- Mixing measure $Q$ 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 mixing measure $G_u$ varies 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):

\[
\psi_t | \{\psi_l\}_{l \neq t}, \gamma, H \sim \sum_{k=1}^{K} \frac{q_k}{q. + \gamma} \delta_{\phi_k} + \frac{\gamma}{q. + \gamma} H.
\]
Posterior inference

• Nested HDP is amenable to Gibbs sampling
  – sampling local atoms by integrating out $G_u$’s
  – sampling global atoms by integrating out $Q$, and centering 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 covariate levels $u$
Recall: simple computational heuristic

- viewed as a “co-clustering” problem, one for each \( u \)
- collection of co-varying mixture distributions indexed by covariate \( u \)
- a heuristic:
  - solve each clustering problem individually (allowing for sampling of number of clusters)
  - mix-match clusters from mixture distributions across different \( u \)’s
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 cluster labels \(z\):

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

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 global/functional clusters \(\phi\):

\[
p(\phi_k|z, \text{Data}) \propto H(\phi_k) \prod_{ui:z_{ui}=k} F(y_{ui}|\phi_{uk}) \text{ for each } k = 1, \ldots, K.
\]
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 $(\sigma, \omega) = (1, 0.01)$
  - smoothness specification is same as ground truth
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 $H$

Global (functional) clusters cannot be identified unless sufficiently smooth, 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 \textit{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)
Modeling of temperature/depth in Atlantic ocean

- data are (temp, depth) samples collected at 4 different locations at different times
- functional clustering within each location
- functional comparisons (ANOVA) between locations
Posterior distribution of global atoms

Mean/credible intervals of functional mean curves

\(\phi_1\)  \(\phi_2\)  \(\phi_3\)  \(\phi_4\)  \(\phi_5\)  \(\phi_6\)
Number of functional clusters

All groups

0 500 1000 1500 2000

0 100 500 1000

2 4 6 8 10
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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>
Varying number of local clusters with depth

Posterior mean (solid) and (.05,.95) credible intervals (dash)
Identifiability and posterior consistency

• motivation: under what conditions can we ensure identifiability, posterior consistency, and convergence rates of (latent) functional clusters on basis of non-functional data?

• two layers of complexity:
  – use of Gaussian process to introduce smoothness of functional clusters
  – use of Dirichlet process to capture heterogeneity via multiple clusters

• recent work on posterior consistency: Barron, Schervish & Wasserman; Shen & Wasserman; Ghosal & van der Vaart, Walker; Ghosal, Ghosh, & R. V. Ramamoorthi; Lijoi, Walker & Prunster;
Posterior consistency and identifiability in infinite mixture

• suppose that $G$ is a discrete mixing measure on space $\Theta$

• combining $G$ with density of likelihood $f(\cdot | \theta)$ to obtain a mixture distribution:

$$p_G(x) = \int f(x | \theta) dG(\theta).$$

• data $X_1, \ldots, X_n$ are iid from $p_{G^*}(\cdot)$ for some “true” mixing measure $G^*$

• endow $G$ with a prior $\Pi$ (such as Dirichlet process)

• question: how fast does the posterior distribution of $G$:

$$\Pi(G | X_1, \ldots, X_n)$$

shrink in the neighborhood of true $G^*$, as $n$ tends to infinity?
Wasserstein metric for discrete measures

- let $\rho$ be a metric of space $\Theta$
- $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$
- Wasserstein metric $d_\rho(G, G')$ is defined as:

$$d_\rho(G, G') = \inf_q \sum_{i,j} q_{ij} \rho(\theta_i, \theta'_j),$$

where $q$ is matrix of joint probabilities on $(i, j)$ such that $\sum_j q_{ij} = p_i$ and $\sum_i q_{ij} = p'_j$. 
Wasserstein metric for discrete measures

- let $\rho$ be a metric of space $\Theta$
- $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$
- Wasserstein metric $d_\rho(G, G')$ is defined as:

$$d_\rho(G, G') = \inf_q \sum_{i,j} q_{ij} \rho(\theta_i, \theta'_j),$$

where $q$ is matrix of joint probabilities on $(i, j)$ such that $\sum_j q_{ij} = p_i$ and $\sum_i q_{ij} = p'_j$.

- if $\Theta = \mathbb{R}^d$, $\rho$ is usual Euclidean metric
- if $\Theta = l_\infty[0, 1]$ a Banach space of bounded functions on $[0, 1]$, $\rho$ is the uniform norm
Theorem 1: Finite mixtures

(Nguyen, 2011)

• If $\Theta = \mathbb{R}^d$ and $f(\cdot|\theta)$ belongs to a family satisfying suitable identifiability conditions. Assume there are $k < \infty$ mixture components, $k$ known. Then, there is a constant $M > 0$ such that:

$$\Pi(d_\rho(G, G^*) > M n^{-1/4} | X_1, \ldots, X_n) \to 0$$

in $P_{G^*}$-probability.

– this generalizes a result of Chen (1995)

• If $\Theta = l_\infty([0,1])$, $G$ is distributed by mixture of $k$ Gaussian sample paths with smoothness $\gamma$, while true $G^*$ is supported by elements of $\Theta$ with smoothness $\gamma^*$. Then,

$$\Pi(d_\rho(G, G^*) > M n^{-\frac{\gamma \wedge \gamma^*}{2(2\gamma \wedge \gamma^* + 1)}} | X_1, \ldots, X_n) \to 0$$

in $P_{G^*}$-probability.
Theorem 2: Infinite mixtures with Dirichlet prior

(Nguyen, 2011)

Assume that the number of mixture components is unknown.

• If $\Theta = \mathbb{R}^d$ and $f(\cdot | \theta)$ belongs to a family of ordinary smooth density functions with smoothness $\beta > 0$. Then, for any $\delta > 0$, there is a constant $M > 0$ such that:

$$
\Pi(d_\rho(G, G^*) > M(\log n/n)^{(d+2)(4+(2\beta+1)d)+\delta} | X_1, \ldots, X_n) \rightarrow 0
$$

in $P_{G^*}$-probability.

• If $\Theta = \mathbb{R}^d$ and $f(\cdot | \theta)$ belongs to a family of supersmooth density functions with smoothness $\beta > 0$. Then, there is a constant $M > 0$ such that:

$$
\Pi(d_\rho(G, G^*) > M(\log n)^{-1/\beta} | X_1, \ldots, X_n) \rightarrow 0
$$

in $P_{G^*}$-probability.
Open questions remain ...

- Posterior consistency for Dirichlet process mixture using Gaussian process as centering measure

- Posterior consistency for our nested HDP model (for which functional data are not available)
Summary

• inference of global/functional clusters from local/non-functional data

• the framework of nested hierarchy of Dirichlet processes

• applicability to a range of problems and data sets

• initial results towards full theoretical analysis (i.e., posterior consistency) for nonparametric Bayesian models of this type

• relevant papers