Notes 4: Dependent Nonparametric Prior Models

Outline

1. Dependent Dirichlet processes
2. Models for finite collections of distributions
3. Spatial Dirichlet process models
1. Dependent Dirichlet processes

- The earlier part of the Bayes nonparametrics literature has focused on problems where a single distribution is assigned a nonparametric prior.

- However, in many applications, the objective is modeling a collection of distributions \( G = \{G_s : s \in S\} \) — for example, \( S \) might be a time interval, a spatial region, or a covariate space.

- Obvious options:
  - assume that the distribution is the same everywhere, e.g., \( G_s \equiv G \sim \text{DP}(\alpha, G_0) \) for all \( s \) — this is too restrictive
  - assume that the distributions are independent and identically distributed, e.g., \( G_s \sim \text{DP}(\alpha, G_0) \) independently for each \( s \) — this is wasteful

- We would like something in between ...
**Dependent Dirichlet processes**

- A similar dilemma arises in parametric models. Consider the hierarchical model:

\[
y_{ij} = \theta_i + \epsilon_{ij} \quad \epsilon_{ij} \sim \text{i.i.d. } N(0, \sigma^2)
\]

\[
\theta_i = \eta + \nu_i \quad \nu_i \sim \text{i.i.d. } N(0, \tau^2)
\]

with \(\eta \sim N(\eta_0, \kappa^2)\)

- If \(\tau^2 \to 0\) we have \(\theta_i = \eta\) for all \(i\), i.e., all means are the same ("maximum" borrowing of information across groups)

- If \(\tau^2 \to \infty\) all the means are different and independent from each other (no information is borrowed)

- To obtain a setting that is *between* the two extremes above, the hierarchical model can be extended by assigning a prior to \(\tau^2\)

- How can we generalize this idea to distributions?
Dependent Dirichlet processes

- A number of modeling approaches have been considered in the literature:
  - Introducing dependence through the baseline distributions of conditionally independent nonparametric priors, e.g., product of mixtures of DPs (refer to Notes 1, section 5). Simple but restrictive
  - Mixing of independent draws from a Dirichlet process:
    \[ G_s = w_1(s)G_1^* + w_2(s)G_2^* + \ldots + w_p(s)G_p^* \]
    where \( G_i^{* \ ind} \sim \text{DP}(\alpha, G_0) \) and \( \sum_{i=1}^{p} w_i(s) = 1 \) (e.g., Müller, Quintana & Rosner, 2004). Hard to extend to uncountable \( S \)
  - **Dependent Dirichlet process** (DDP): Starting with the stick-breaking construction of the DP, and replacing the weights and/or atoms with appropriate stochastic processes on \( S \) (MacEachern, 1999; 2000). Very general procedure; most of the models discussed here can be framed as DDPs
Dependent Dirichlet processes

- Recall the constructive definition of the Dirichlet process: $G \sim \text{DP}(\alpha, G_0)$ if and only if $G = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\theta_\ell}$ where the $\theta_\ell$ are i.i.d. from $G_0$; and $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1} (1 - z_r)$, $\ell = 2, 3, \ldots$, with $z_r$ i.i.d. Beta($1, \alpha$)

- To construct a DDP prior (MacEachern, 2000) for the collection of random distributions, $\mathcal{G} = \{G_s : s \in S\}$, define $G_s$ as

$$G_s = \sum_{\ell=1}^{\infty} \omega_\ell(s) \delta_{\theta_\ell(s)}$$

$\rightarrow$ with $\theta_{\ell,S} = \{\theta_\ell(s) : s \in S\}$, for $\ell = 1, 2, \ldots$, independent realizations from a (centering) stochastic process $G_{0,S}$ defined on $S$

$\rightarrow$ and stick-breaking weights defined through independent realizations $z_{\ell,S} = \{z_\ell(s) : s \in S\}$, $\ell = 1, 2, \ldots$, from a stochastic process on $S$ with marginals $z_\ell(s) \sim \text{Beta}(1, \alpha(s))$

$\rightarrow$ the stochastic processes that define the dependent atoms and weights of $G_s$ would typically arise through transformations of Gaussian processes
Dependent Dirichlet processes

- Key property: for any fixed $s$, the DDP construction yields a DP prior distribution for $G_s$

- Smoothness (e.g., continuity) properties of the centering process $G_{0,s}$ and the stochastic process that defines the weights drive continuity of DDP realizations — we would typically seek smooth evolution for the distributions $G_s$, with the level of dependence between $G_s$ and $G_{s'}$ driven by the distance between $s$ and $s'$

- DDP models that involve dependent atoms $\theta_\ell(s)$ require some form of replicate responses at the observed index points $s$

- A key inferential objective in many applications is prediction for distributions $G_{s_0}$ at new index points $s_0$ where data have not been collected

- **DDP mixture models:** as with DPs, we usually employ the DDP prior to model the distribution of the parameters in a hierarchical model
Dependent Dirichlet processes

- “Single-\(\rho\)” DDP model: special case of the general construction where the weights do not depend on \(s\), and only the atoms are realizations from the centering stochastic process \(G_{0,s}\):

\[
G_s = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\theta_\ell}(s)
\]

with \(\omega_1 = z_1, \omega_\ell = z_\ell \prod_{r=1}^{\ell-1} (1 - z_r), \ell = 2, 3, \ldots\), with \(z_r\) i.i.d. Beta(1, \(\alpha\))

- Advantage: computation in single-\(\rho\) DDP mixture models is simplified, since, for the observed data, they can be expressed as DP mixtures

- Disadvantage: it can be somewhat restrictive, for example, single-\(\rho\) DDP priors can not produce a collection of independent distributions, not even as a limiting case; also, dependent weights can generate local dependence structure which is desirable in temporal or spatial applications
Dependent Dirichlet processes

- Some applications of DDP mixtures: dynamic density estimation (Rodriguez & ter Horst, 2008); survival regression modeling (De Iorio et al., 2009); quantile regression (Kottas & Krnjajić, 2009); inference for spatial point patterns observed over discrete time (Taddy, 2010); modeling and risk assessment for developmental toxicity studies (Fronczyk & Kottas, 2010)

- Section 2 provides an overview of three classes of priors for finite collections of distributions: ANOVA DDP (De Iorio et al., 2004); hierarchical DPs (Teh. et al., 2006), which are related to the “analysis of densities” model (Tomlinson & Escobar, 1999); and nested DPs (Rodríguez et al., 2008) — Section 3 presents spatial DPs (Gelfand et al., 2005; Kottas et al., 2008)

- Several other approaches: order-dependent DDPs (Griffin & Steel, 2006); generalized spatial DPs (Duan et al., 2007); kernel stick-breaking processes (Dunson & Park, 2007); dependent Polya tree regression models (Trippa et al., 2010); stick-breaking autoregressive processes (Griffin & Steel, 2010); ...
Models for finite collections of distributions

2. Models for finite collections of distributions

ANOVA dependent Dirichlet process models

- Consider an index set $S$ such that $s = (s_1, \ldots, s_p)$ corresponds to a vector of categorical variables — e.g., in a clinical setting, $G_{s_1,s_2}$ might correspond to the random effects distribution for patients treated at levels $s_1$ and $s_2$ of two different drugs.

- For example, consider $f(\cdot; G_{s_1,s_2}, \sigma^2) = \int \mathcal{N}(\cdot; \eta, \sigma^2) \, dG_{s_1,s_2}(\eta)$ where

  $$G_{s_1,s_2} = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_{h,s_1,s_2}}$$

with, say, $\theta_{h,s_1,s_2} = m_h + A_{h,s_1} + B_{h,s_2} + AB_{h,s_1,s_2}$, and

  $$m_h \sim G_0^m \quad A_{h,s_1} \sim G_0^A \quad B_{h,s_2} \sim G_0^B \quad AB_{h,s_1,s_2} \sim G_0^{AB}$$

- Typically, $G_0^m$, $G_0^A$, $G_0^B$, $G_0^{AB}$ are normal distributions and identifiability constrains are used (e.g., $A_{h,1} = B_{h,1} = 0$ and $AB_{h,1,s_2} = AB_{h,s_1,1} = 0$)
Models for finite collections of distributions

- Note that the atoms of $G_{s_1,s_2}$ have a structure that resembles a two-way ANOVA
- Indeed, the ANOVA-DDP mixture model can be formulated as a DP mixture of ANOVA models where, at least in principle, there can be up to one different ANOVA for each observation:

$$\int N(\cdot; d_{s_1,s_2} \eta, \sigma^2) dF(\eta), \quad F \sim \text{DP}(\alpha, G_0)$$

where $d_{s_1,s_2}$ is a design vector selecting the appropriate coefficients from $\eta$ and $G_0 = G_0^m G_0^A G_0^B G_0^{AB}$

- In practice, just a small number of ANOVA models: recall that the DP prior clusters observations — also, note that if a single component is used, we recover a parametric ANOVA model
- Expressing the model as a DP mixture simplifies computation: MCMC methods for DP mixtures can be used to fit the ANOVA-DDP model
Models for finite collections of distributions

Hierarchical Dirichlet processes

- A model for exchangeable collections of distributions

- Consider observations $y_{ij} \sim F_j$. For example, $y_{ij}$ might correspond to the SAT score obtained by student $i = 1, \ldots, r_j$ in school $j = 1, \ldots, J$

- Hierarchical Dirichlet process (HDP) mixture models estimate $F_j$ by identifying latent classes of students within each school, and by sharing classes across schools

- Let

$$f(y_{ij}; G_j) = \int k(y_{ij}; \eta) dG_j(\eta), \quad G_j \sim \text{DP}(\alpha, G_0), \quad G_0 \sim \text{DP}(\beta, H)$$

- Conditionally on $G_0$, the mixing distribution for each school is an independent sample from a DP — dependence across schools is introduced, since they all share the same baseline measure $G_0$
Models for finite collections of distributions

- $G_0$, being drawn from a DP, is almost surely discrete $G_0 = \sum_{\ell=1}^{\infty} \omega_{\ell} \delta_{\phi_{\ell}}$

- Therefore, when we draw the atoms for $G_j$ we are forced to choose among $\phi_1, \phi_2, \ldots$, i.e., we can write $G_j$ as:

$$G_j = \sum_{\ell=1}^{\infty} \pi_{\ell,j} \delta_{\phi_{\ell}}$$

- Note that the weights assigned to the atoms are not independent — if $\phi_{\ell}$ has a large associated weight $\omega_{\ell}$ under $G_0$, then the weight $\pi_{\ell,j}$ under $G_j$ will likely be large for every $j$ (in fact, $E(\pi_{\ell,j}) = \omega_{\ell}$)

- Note that the HDP is a DDP, but not a single-$p$ DDP

- In spite of that, an MCMC sampler can be devised for posterior simulation by composing two Pólya urns, one built from $(\alpha, G_0)$ and one from $(\beta, H)$ — the resulting MCMC algorithm is similar to the marginal sampler for DP mixture models, but bookkeeping is harder
Models for finite collections of distributions

Nested Dirichlet Processes

- Also a model for exchangeable distributions — rather than borrowing strength by sharing clusters among all distributions, the nested DP (NDP) borrows information by clustering similar distributions together

- An example: assessment for quality of care in hospitals across the nation → we want to cluster states with similar distributions, and simultaneously cluster hospitals with similar outcomes
  → let $y_{ij}$ be the percentage of patients in hospital $i = 1, \ldots, n_j$ within state $j = 1, \ldots, J$ who received the appropriate antibiotic on admission
  → NDP mixture model: $f(y_{ij}; G_j) = \int k(y_{ij}; \eta)dG_j(\eta)$, where

$$G_j \sim \sum_{k=1}^{K} \omega_k \delta_{G^*_k} \quad \text{and} \quad G^*_k = \sum_{\ell=1}^{\infty} \pi_{\ell k} \delta_{\theta_{\ell k}}$$

where $\theta_{\ell k} \sim H$, $\pi_{\ell k} = u_{\ell k} \prod_{r<\ell} (1 - u_{rk})$ with $u_{\ell k} \sim \text{Beta}(1, \beta)$, and $\omega_k = v_k \prod_{r<k} (1 - v_r)$ with $v_k \sim \text{Beta}(1, \alpha)$
In this case, we write \( \{G_1, \ldots, G_J\} \sim DP(\alpha, DP(\beta, H)) \).

- Notationwise, the NDP resembles the HDP, but it is quite different
- The NDP is not a single-\( p \) DDP model
- Note that the NDP generates two layers of clustering: states, and hospitals within groups of states. However, groups of states are conditionally independent from each other
- A standard marginal sampler is not feasible in this problem — computation can be carried out using an extension of the blocked Gibbs sampler
Models for finite collections of distributions
3. Spatial Dirichlet process models

- Spatial data modelling: based on **Gaussian processes** (distributional assumption) and **stationarity** (assumption on the dependence structure)

- Basic modelling for a spatial random field $Y_D = \{Y(s) : s \in D\}$, $D \subseteq R^d$:

  $$Y(s) = \mu(s) + \theta(s) + \epsilon(s)$$

  - $\mu(s)$ mean process, e.g., $\mu(s) = x'(s)\beta$

  - $\theta(s)$ a spatial process, typically, a mean 0 isotropic Gaussian process, i.e., $\text{Cov}(\theta(s_i), \theta(s_j) \mid \sigma^2, \phi) = \sigma^2 \rho_\phi(||s_i - s_j||) = \sigma^2 (H(\phi))_{i,j}$

  - a pure error (nugget) process, e.g., $\epsilon(s)$ i.i.d. $N(0, \tau^2)$

- Induced model for observed sample (**point referenced spatial data**), $Y = (Y(s_1), ..., Y(s_n))$, at sites $s^{(n)} = (s_1, ..., s_n)$ in $D$

  $$Y \mid \beta, \sigma^2, \phi, \tau^2 \sim N(X'\beta, \sigma^2 H(\phi) + \tau^2 I_n)$$
Spatial Dirichlet process models

- **Objective of Bayesian nonparametric modelling:** develop prior models for the distribution of $\theta_D = \{\theta(s) : s \in D\}$, and thus for the distribution of $Y_D = \{Y(s) : s \in D\}$, that relax the Gaussian and stationarity assumptions.

- In general, a fully nonparametric approach requires replicate observations at each site, $Y_t = (Y_t(s_1), ..., Y_t(s_n))'$, $t = 1, ..., T$, though imbalance or missingness in the $Y_t(s_i)$ can be handled.

- Temporal replications available in various applications, e.g., in epidemiology, environmental contamination, and weather modeling → direct application of the methodology for spatial processes (when replications can be assumed approximately independent) → more generally, extension to **spatio-temporal modelling**, e.g., through dynamic spatial process modelling viewing $Y(s, t) \equiv Y_t(s)$ as a temporally evolving spatial process (Kottas, Duan & Gelfand, 2008).
Spatial Dirichlet process models

- **Spatial Dirichlet process**: arises as a dependent DP where $G_0$ is extended to $G_{0D}$, a random field over $D$, e.g., a stationary Gaussian process — thus, in the DP constructive definition, each $\theta_\ell$ is extended to $\theta_{\ell,D} = \{\theta_\ell(s) : s \in D\}$ a realization from $G_{0D}$, i.e., a random surface over $D$

- Hence, the spatial DP:

$$G_D = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\theta_{\ell,D}}$$

random process over $D$ centered at $G_{0D}$ (notation: $G_D \sim \text{SDP}(\alpha, G_{0D})$)

- Key property: if

$$\theta_D = \{\theta(s) : s \in D\} \mid G_D \sim G_D, \text{ and } G_D \sim \text{SDP}(\alpha, G_{0D})$$

then for any $s^{(n)} = (s_1, ..., s_n)$, $G_D$ induces $G^{(s^{(n)})} \equiv G^{(n)}$, a random distribution for $(\theta(s_1), ..., \theta(s_n))$, and $G^{(n)} \sim \text{DP}(\alpha, G_0^{(n)})$, where $G_0^{(n)} \equiv G_0^{(s^{(n)})}$ is $n$-variate normal (if $G_{0D}$ is a Gaussian process)
Spatial Dirichlet process models

- For stationary $G_{0D}$, the smoothness of realizations from $\text{SDP}(\alpha, G_{0D})$ is determined by the choice of the covariance function of $G_{0D}$
  → for instance, if $G_{0D}$ produces a.s. continuous realizations, then
  \[ G^{(s)} - G^{(s')} \to 0 \text{ a.s. as } \|s - s'\| \to 0 \]
  → we can learn about $G^{(s)}$ more from data at neighboring locations than from data at locations further away (as in usual spatial prediction)

- Random process $G_D$ is centered at a stationary Gaussian process, but it is nonstationary, it has nonconstant variance, and it yields non-Gaussian finite dimensional distributions

- More general spatial DP models?
  → allow weights to change with spatial location, i.e., allow realization at location $s$ to come from a different surface than that for the realization at location $s'$ (Duan, Guindani & Gelfand, 2007)
Spatial Dirichlet process models

- Almost sure discreteness of realizations from $G_D$?
  $\rightarrow$ mix $G_D$ against a pure error process $\mathcal{K}$ (i.i.d. variables $\epsilon(s)$ with mean 0 and variance $\tau^2$) to create random process over $D$ with continuous support

- **Spatial DP mixture model:** If $G_D \sim \text{SDP}(\alpha, G_{0_D}), \theta_D \mid G_D \sim G_D,$ and $Y_D - \theta_D \mid \tau^2 \sim \mathcal{K}$

\[
F \left( Y_D \mid G_D, \tau^2 \right) = \int \mathcal{K} \left( Y_D - \theta_D \mid \tau^2 \right) dG_D \left( \theta_D \right)
\]

i.e., $Y(s) = \theta(s) + \epsilon(s); \theta(s)$ from a spatial DP; $\epsilon(s)$, say, i.i.d. $\text{N}(0, \tau^2)$
(again, random process $F$ is **non-Gaussian** and **nonstationary**)

- Adding covariates, the induced model at locations $s^{(n)} = (s_1, ..., s_n)$,

\[
f \left( \mathbf{Y} \mid G^{(n)}, \beta, \tau^2 \right) = \int f_{N_n} \left( \mathbf{Y} \mid X'\beta + \theta, \tau^2 I_n \right) dG^{(n)} \left( \theta \right)
\]

where $\mathbf{Y} = (Y(s_1), ..., Y(s_n))', \theta = (\theta(s_1), ..., \theta(s_n))',$ and $X$ is a $p \times n$ matrix with $X_{ij} = \text{value of the } i\text{-th covariate at the } j\text{-th location}$
Spatial Dirichlet process models

- Data: for \( t = 1, \ldots, T \), response \( \mathbf{Y}_t = (Y_t(s_1), \ldots, Y_t(s_n))' \) (with latent vector \( \mathbf{\theta}_t = (\theta_t(s_1), \ldots, \theta_t(s_n))' \)), and matrix of covariate values \( \mathbf{X}_t \)

- \( G_0^{(n)}(\cdot \mid \sigma^2, \phi) = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi)) \) where \( (H_n(\phi))_{i,j} = \rho_\phi(s_i - s_j) \) (or \( \rho_\phi(||s_i - s_j||) \)), induced by a mean 0 stationary (or isotropic) Gaussian process \( \rho_\phi(|| \cdot ||) = \exp(-\phi || \cdot ||), \phi > 0 \), for the data examples

- Bayesian model: (conjugate DP mixture model)

\[
\mathbf{Y}_t \mid \mathbf{\theta}_t, \mathbf{\beta}, \tau^2 \quad \text{ind.} \quad \sim \quad N_n(\mathbf{Y}_t \mid X_t' \mathbf{\beta} + \mathbf{\theta}_t, \tau^2 I_n), \quad t = 1, \ldots, T
\]

\[
\mathbf{\theta}_t \mid G^{(n)} \quad \text{i.i.d.} \quad \sim \quad G^{(n)}, \quad t = 1, \ldots, T
\]

\[
G^{(n)} \mid \alpha, \sigma^2, \phi \quad \sim \quad \text{DP}(\alpha, G_0^{(n)}); G_0^{(n)} = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi))
\]

with hyperpriors for \( \mathbf{\beta}, \tau^2, \alpha, \sigma^2, \) and \( \phi \)

- Posterior inference using standard MCMC techniques for DP mixtures (refer to the second set of notes) — extensions to accommodate missing data — methods for prediction at new spatial locations
Data examples

• *Simulated data set*: generate data from a non-Gaussian process $Z_D$ arising from a two-component mixture of independent Gaussian processes with constant means $\mu_k$ and covariance functions $\sigma^2 \exp(-\phi_k \|s - s'\|)$, $k = 1, 2$ (processes 1 and 2 are sampled with probabilities $q$ and $1 - q$)

• Generate $T = 75$ replications (at the $n = 39$ sites given in Figure 1) from the process $Z(s) + e(s)$, where $\sigma = 0.5$, $\phi_1 = \phi_2 = 0.0025$, $\mu_1 = -2$, $\mu_2 = 2$, $q = 0.75$, and $e(s)$ is a pure error process with variance $\tau^2 = 0.5$

• New sites for spatial prediction (denoted by “*” in Figure 1)

• Comparison with a Gaussian process mixture model,

\[
\theta_t \mid \sigma^2, \phi \overset{i.i.d.}{\sim} G^{(n)}_0 = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi)), \quad t = 1, \ldots, T
\]

(limiting case, as $\alpha \to \infty$, of the spatial DP mixture model)
Figure 1: Geographic map of the Languedoc-Roussillon region in southern France
Figure 2: Simulated data. Posterior predictive densities under spatial DP mixture model (solid lines) and GP mixture model (dotted lines)
Spatial Dirichlet process models

- *Precipitation data from the Languedoc-Rousillon region in southern France*

- Original version of the dataset includes 108 altitude-adjusted 10-day aggregated precipitation records for the 39 sites in Figure 1

- We work with a subset of the data based on the 39 sites but only 75 replicates (to avoid records with too many 0-s), which have been log-transformed with site specific means removed

- Preliminary exploration of the data suggests that spatial association is higher in the northeast than in the southwest

- In the interest of validation for spatial prediction, we removed two sites from each of the three subregions in Figure 1, specifically, sites $s_4$, $s_{35}$, $s_{29}$, $s_{30}$, $s_{13}$, $s_{37}$, and refitted the model using only the data from the remaining 33 sites
Spatial Dirichlet process models

Figure 3: French precipitation data. Image plots based on functionals of posterior predictive distributions at observed sites and a number of new sites (darker colors correspond to smaller values)
Figure 4: French precipitation data. Bivariate posterior predictive densities for pairs of sites \((s_4, s_{35}), (s_{29}, s_{30}), (s_{13}, s_{37})\) and \((s_4, s_{13})\) based on model fitted to data after removing sites \(s_4, s_{35}, s_{29}, s_{30}, s_{13}\) and \(s_{37}\) (overlaid on data observed at the corresponding pairs of sites in the full dataset)


References


