Introduction to Bayesian nonparametrics

Andriy Norets  
Department of Economics, Princeton University  
September, 2010

Outline

- Dirichlet Process (DP) - prior on the space of probability measures.  
- Some applications of DP.  
- DP mixtures.  
- Role of consistency in Bayesian nonparametrics.  
- Schwartz posterior consistency theorem and extensions.  
- Applications to conditional density estimation (consistency, MCMC estimation methods, comparison with classical methods on real data).

Books on Bayesian nonparametrics

- Dey et al. (1998) - applications and MCMC estimation methods.  
- Hjort (2010) - collection of up to date reviews of theoretical and applied work in different subfields.

Notation and discrete example

- $(X, B, P)$ - probability space for observations.  
- $M(X)$ - space of probability measures on $X$ ($P \in M(X)$).  
- For Bayesian nonparametric inference we need to put a prior $\pi$ on $M(X)$.  

Dirichlet process prior for $X = \{1, \ldots, n\}$ and $M(X) = \{p = (p_1, \ldots, p_n) : p_i \geq 0, \sum p_i = 1\},$

$$\pi(p_1, \ldots, p_n) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} p_1^{\alpha_1-1} \cdots p_n^{\alpha_n-1}$$

known as the Dirichlet distribution with parameter $\alpha$, $D(\alpha)$.
Discrete example continued

The likelihood has the same functional form as the prior:

\[
l(y_1, \ldots, y_K | p) = p_1^{\sum_k 1(y_k = 1)} \cdots p_n^{\sum_k 1(y_k = n)}
\]

Thus, the posterior is also a Dirichlet distribution

\[
\pi(p | y) = D(\alpha_1 + \sum_k 1\{y_k = 1\}, \ldots, \alpha_1 + \sum_k 1\{y_k = n\})
\]

If we treat vector \( \alpha \) as a measure on \( X \) \((\alpha(i) = \alpha_i)\),

\[
\pi(p | y) = D(\alpha + \sum_{i=1}^{K} \delta_{y_i}), \text{ where } \delta_{y_i} \text{ is a point mass at } y_i.
\]

Ferguson (1973) construction

Kolmogorov’s extension theorem for general case:

- Define finite dimensional distributions first: for a finite partition \( X = \bigcup_{i=1}^{n} A_i \), let

\[
(P(A_1), \ldots, P(A_n)) \sim \text{Dirichlet}(\alpha(A_1), \ldots, \alpha(A_n))
\]

where \( \alpha \) is a measure on \( X \) (it is a parameter for the DP)

- Next, using properties of the Dirichlet distribution, show that an extension of these finite dimensional distributions to \( M(X), \text{DP}(\alpha) \), is unique and that, as in the discrete case, the posterior is also \( \text{DP}(\alpha + \sum_{i=1}^{K} \delta_{y_i}) \).

Sethuraman (1994) construction

Sethuraman showed that DP can be represented as

\[
\text{DP}(\alpha) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}
\]

where

\[
\theta_i \sim \alpha(\cdot)/\alpha(X), \text{ iid.}
\]

\[
p_1 = q_1, \quad p_i = q_i \prod_{j=1}^{i-1} (1 - q_j) \quad \text{(stick breaking)}
\]

\[
q_i \sim \text{Beta}(1, \alpha(X)), \text{ iid.}
\]

Thus, DP puts probability 1 on discrete probability measures.

Applications of DP prior

Ferguson (1973) showed that

- mean
- variance and covariance
- CDF

all converge to sample analogs.
Application of DP prior to estimation of CDF

Under squared loss, the prior point estimate is

$$\hat{F}(t) = \int P(-\infty, t]DP_\alpha(dP) = \frac{\alpha(-\infty, t]}{\alpha(R)}$$

because $(-\infty, t]$ and $(t, \infty)$ is a partition of $R$ and under $DP_\alpha$

$$P(-\infty, t] \sim \text{Beta}(\alpha(-\infty, t], \alpha(t, \infty)).$$

To get posterior point estimate, use $\alpha + \sum_{i=1}^{K} \delta_{y_i}$ instead of $\alpha$:

$$\hat{F}(t|y_1, \ldots, y_K) = \frac{\alpha(R)}{\alpha(R) + K} \cdot \frac{\alpha(-\infty, t]}{\alpha(R)} + \frac{K}{\alpha(R) + K} \cdot \text{Empirical CDF}$$

Applications of multinomial-Dirichlet model

- Rubin (1981) (Bayesian bootstrap): multinomial-Dirichlet model delivers a posterior distribution for certain parameters (mean, variance) which is almost identical to classical bootstrap distribution.
- Lancaster (2003): OLS with White heteroscedasticity robust variance covariance matrix can be justified by Bayesian bootstrap if the parameter of interest is the minimizer of $E(y - \beta x)^2$. Poirier (2008) obtain different versions of White's estimator used in classical literature under different prior specifications.
- Chamberlain and Imbens (2003) suggest multinomial-Dirichlet model with added prior moment equality restrictions as a Bayesian way to handle moment equality models.

Multinomial-Dirichlet model

- Several authors use a multinomial likelihood with the support given by observations in the sample (or all possible values for observations) and an improper Dirichlet prior on the multinomial probabilities.
- This can be seen as a limit of DP when $\alpha(X) \to 0$ (the prior becomes “uninformative”, $DP_{\alpha + \sum_{i=1}^{K} \delta_{y_i}} \to D_{\sum_{i=1}^{K} \delta_{y_i}}$).
- A parameter of interest is defined as a function of $(p_1, \ldots, p_K)$ (parameters of the multinomial likelihood) and $(y_1, \ldots, y_K)$ (observations), e.g., the mean is $\sum p_i y_i$.

DP mixtures (most common use of DP)

$$y_i|m_i, s_i \sim N(m_i, s_i^2)$$
$$m_i, s_i|F \sim F$$
$$F \sim DP(\alpha)$$

By Sethuraman construction, this is equivalent to a countable mixture of normals:

$$y_i \sim \sum_{i=1}^{\infty} p_i N(m_i, s_i^2)$$

Practical MCMC algorithms are available (Escobar and West (1995), Neal (2000)).

The model is similar to a finite mixture of normals, which we'll consider below.
Applications of DP mixtures

- Dey et al. (1998) - a collection of applications.
- Conley et al. (2008) - IV with flexible error term distributions.
- Burda et al. (2008) - discrete choice model with heterogeneous coefficients (heterogeneity distribution is modeled by DP mixtures).

Motivation for studying consistency in the Bayesian framework

- Priors in infinite (high) dimensional problems are hard to construct/elicit.
- There are known examples of seemingly reasonable priors that lead to unreasonable posteriors (Diaconis and Freedman (1986)).
- Posterior consistency (posterior concentrates around true parameter values asymptotically) provides an additional check on the prior distributions.

(A version of) Schwartz posterior consistency theorem

For a measure \( \mu \) on \( X \), let \( L_\mu \) be the space of probability densities w.r.t. \( \mu \). Let \( \pi \) be a prior on \( L_\mu \).

If any Kullback-Liebler neighborhood of the “true” density, \( f_0 \in L_\mu \), has positive prior \( \pi \) probability, then the posterior is consistent in weak topology:

\[
\pi(U|y_1, \ldots, y_n) \rightarrow 1, \text{ a.s.} f_0^n
\]

for any weak topology neighborhood \( U \).

Generalizations of Schwartz theorem

- Ghosal et al. (2000) - posterior convergence rates.
- Ghosal and van der Vaart (2007) - some Bayesian nonparametric procedures achieve minimax rates in estimation of univariate density.
Two approaches based on mixtures

- Model the joint distribution by finite mixtures of multivariate normals (FMMN) and extract the conditional distributions of interest from the model (West, Mueller, Escobar (1994), Wu and Ghosal (2009), Norets and Pelenis (2009)).
- Model the conditional distribution directly by smooth mixtures of regressions (SMR) – mixing probabilities are functions of covariates (Jacobs et al. (1991), Jiang and Tanner (1999), Geweke and Keane (2007), Villani et al. (2009), Norets (2010)).

FMMN and specifications of SMR

(Conditional) observables density:

\[ p(y|x, \theta, M) = \sum_{j=1}^{m} \alpha_j^m(x; \theta) \phi(y, \mu_j^m(x; \theta), \sigma_j^m(x; \theta)), \]

\( \phi \) is normal pdf, \( \mu_j^m(x; \theta) \) - mean, \( \sigma_j^m(x; \theta)^2 \) - variance
(non-diagonal variance will be denoted by \( H^{-1} \).)

How to specify mixing probabilities \( \alpha_j^m(x; \theta) \) and \( \mu_j^m(x; \theta) \) and \( \sigma_j^m(x; \theta)^2 \)?

- do not depend on \( x \),
- depend on linear functions of \( x \),
- or flexible, e.g., (transformations of) polynomials in \( x \)?

Consistency results

If \( f(y|x) \) is in the KL closure of \( p(y|x, \theta, M_m) \) then we can put a prior on the number of mixture components \( m \) and apply Schwartz posterior consistency theorem to obtain consistency of posterior in weak topology.

Results can be generalized from i.i.d to Markov process settings as in Ghosal and Tang (2006).

Summary of results from Norets (2010)

- Simplest flexible specification: \( \alpha_j^m \)'s – multinomial logit with linear indices in \( x \), \( (\mu_j^m, \sigma_j^m) \) are independent of \( x \).
- Polynomial terms in logit reduce the number of mixture components \( m \).
- Alternatively, \( (\alpha_j^m, \sigma_j^m) \) independent of \( x \) and \( \mu_j^m \) polynomial in \( x \) also work when \( y \) is univariate.
- However, making \( \alpha_j^m \) rather than \( \mu_j^m \) flexible in \( x \) improves the convergence rate.
- Making \( \sigma_j^m \) a function of \( x \) weakens some restrictions on \( F \).
- Modeling conditional distributions directly is better.
Assumptions on target distribution $F$

1. $f(y|x)$ is continuous in $y$ on $Y$ for all $x \in X$.
2. The second moments of $y$ are finite.
3. Unbounded support ($f(y|x) > 0$). For a cube, $C_r(y)$, with center $y$ and side length $r > 0$
   \[
   \int \log \frac{f(y|x)}{\inf_{z \in C_r(y)} f(z|x)} F(dy, dx) < \infty,
   \]
   or
   \[
   \int \sup_{z \in C_r(y)} \frac{d \log f(z|x)}{dz} ||F(dy, dx)|| < \infty,
   \]
4. Bounded support can be accommodated by using some other shapes instead of cubes $C_r(y)$.

Examples satisfying assumptions

- **Exponential:** $f(y|x) = \gamma(x) \exp\{-\gamma(x)y\}$, \(\gamma(x) > 0\), and $\int \gamma^{-2} dF < \infty$
- **Student t,** $f(y|x) \propto \nu + ((y - b(x))/c(x))^2)^{-\nu + 1/2}$, $\nu > 2$, $b(x)^2$ and $c(x)^2$ are integrable w.r.t. $f(x)$.
- Uniformly bounded support, $f(y|x)$ is continuous in $y$, $\infty > f(y|x) \geq f > 0$.
Bounded away from zero condition can be substituted with monotonicity at the boundary.

Estimation for FMMN

Latent states (Diebolt and Robert (1994)): $s_i \in \{1, \ldots, m\}$
- $y_i|s_i, \theta, M \sim \phi(\cdot; \mu_{s_i}, H_{s_i}^{-1})$
- $P(s_i = j|\theta, M) = \alpha_j$

The joint distribution of observables and unobservables

\[
p(\{y_i, s_i\}_{i=1}^T; \{\alpha_j, \mu_j, H_j\}_{j=1}^m | M) \propto \prod_{i=1}^T \alpha_{s_i} |H_{s_i}|^{-0.5} \exp\{-0.5(y_i - \mu_{s_i})'H_{s_i}(y_i - \mu_{s_i})\}
\]
\[
\alpha_1^{\alpha_1 - 1} \cdots \alpha_m^{\alpha_m - 1}
\]
(Dirichlet prior)
\[
\prod_j |H_j|^{0.5} \exp\{-0.5(\mu_j - \mu)'H_j(\mu_j - \mu)\}
\]
(Normal-Wishart prior)
\[
\prod_j |H_j|^{(\nu - d - 1)/2} \exp\{-0.5 \text{tr } S H_j\}
\]
((\nu, S, \mu, \lambda, a) - hyperparameters)
Gibbs sampler

\[ p(\{y_i, s_i\}_{i=1}^T; \{\alpha_j, \mu_j, H_j\}_{j=1}^m | \mathcal{M}) \propto \]
\[ \prod_{i=1}^T \alpha_{s_i} | H_{s_i}|^{0.5} \exp[-0.5(y_i - \mu_{s_i})' H_{s_i} (y_i - \mu_{s_i})] \]
\[ \alpha_1^{a-1} \cdots \alpha_m^{a-1} \quad \text{(Dirichlet prior)} \]
\[ \prod_j |H_j|^{0.5} \exp\{-0.5(\mu_j - \mu)' \Lambda H_j (\mu_j - \mu)\} \quad \text{(Normal-Wishart)} \]
\[ \prod_j |H_j|^{(\nu-d-1)/2} \exp\{-0.5 \operatorname{tr} S H_j\} \quad \text{prior} \]

Gibbs sampler blocks:
\[ p(s_i = j | \ldots) \propto \alpha_j | H_j|^{0.5} \exp[-0.5(y_i - \mu_j)' H_j (y_i - \mu_j)] \]

Gibbs sampler

\[ p(\{y_i, s_i\}_{i=1}^T; \{\alpha_j, \mu_j, H_j\}_{j=1}^m | \mathcal{M}) \propto \]
\[ \prod_{i=1}^T \alpha_{s_i} | H_{s_i}|^{0.5} \exp[-0.5(y_i - \mu_{s_i})' H_{s_i} (y_i - \mu_{s_i})] \]
\[ \alpha_1^{a-1} \cdots \alpha_m^{a-1} \quad \text{(Dirichlet prior)} \]
\[ \prod_j |H_j|^{0.5} \exp\{-0.5(\mu_j - \mu)' \Lambda H_j (\mu_j - \mu)\} \quad \text{(Normal-Wishart)} \]
\[ \prod_j |H_j|^{(\nu-d-1)/2} \exp\{-0.5 \operatorname{tr} S H_j\} \quad \text{prior} \]

Gibbs sampler blocks:
\[ p(\alpha | \ldots) \propto \alpha_1^{\sum_i 1\{s_i=1\}+a-1} \cdots \alpha_m^{\sum_i 1\{s_i=m\}+a-1} \]

Gibbs sampler, discrete variables

Map discrete variables into intervals of \( R \) and model continuous latent variables, \( y_{i,k}^* \in y_{i,k} = [a_{i,k}, b_{i,k}] \), by the mixture.

Add a block for the latent variables:
\[ p(y_{i,k}^* | \ldots) \propto \exp[-0.5(y_{i,k}^* - \mu_{s_i})' H_{s_i} (y_{i,k}^* - \mu_{s_i})] \cdot 1_{[a_{i,k}, b_{i,k}]}(y_{i,k}^*) \]

It is truncated normal.
Gibbs sampler, SMR

- If $\mu_j^m$ is linear or polynomial in $x$, Gibbs sampler is almost the same.
- If $\sigma_j^m$ depends on $x$ then Gibbs sampler blocks for parameters of $\sigma_j^m$ will have non-standard form. One needs to use more sophisticated MCMC algorithms, e.g., Metropolis-Hastings.
- If $\alpha_j^m$ is defined by multinomial logit then Gibbs sampler block for the logit parameters is similar to maximum likelihood estimation of logit model.

Evaluating model quality

- Marginal Likelihood, $p(Y_T|M)$, is difficult to compute.
- Cross Validated Log Scoring Rule

$$
\sum_{t=1}^{T} \log p(y_t|Y_{T/t},M) = \sum_{t=1}^{T} \log \frac{1}{K} \sum_{k=1}^{K} p(y_t|Y_{T/t},\theta^k,M)
$$

- We use: Modified Cross Validated Log Scoring Rule

Randomly order sample observations and use the first $T_1$ observations for inference and the rest for evaluation. Repeat this process several times and compare means or medians.

$$
\sum_{t=T_1+1}^{T} \log p(y_t|Y_{T_1},M)
$$

Labor Market Participation

- Gerfin (1996) cross-section dataset. Compare probit, kernel (Hall et al. (2004)) and FMMN.
- Binary dependent variable - Labor force participation dummy.
- Independent variables: Log of non-labor income, Age, Education, Number of young children, Number of old children, Foreign dummy.
- Number of observations: $T = 872$. Split into two random samples of $T_1 = 650$ and $T_2 = 222$ observations. Use $T_1$ as an estimation sample, and $T_2$ as a prediction sample for 50 different random splits.

Comparison of different models

Table: Cross-validated log scores and classification rates for Labor Force Participation

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Score</th>
<th>%Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>Probit</td>
<td>-136.42</td>
<td>-136.79</td>
</tr>
<tr>
<td>Kernel</td>
<td>-138.02</td>
<td>-135.63</td>
</tr>
<tr>
<td>FMMN(m=1)</td>
<td>-136.47</td>
<td>-135.62</td>
</tr>
<tr>
<td>FMMN(m=2)</td>
<td>-132.26</td>
<td>-132.13</td>
</tr>
<tr>
<td>FMMN(m=3)</td>
<td>-133.20</td>
<td>-132.88</td>
</tr>
</tbody>
</table>

Mean and median log scores and classification rates for probit, kernel and FMMN models evaluated at 50 random evaluation samples of $T_2 = 222$ observations.
Boston Housing Data

- Data were analysed by kernel methods by Li and Racine (2008).
- Compare kernel methods from Hall et al. (2004) (cross-validation) with FMMN and SMR.
- Dependent variable - price of the house in a given area.
- Independent variables: average number of rooms in the area, percentage of the population having lower economic status in the area, weighted distance to five Boston employment centers.
- Number of observations: \( T = 506 \). Split into two random samples of \( T_1 = 400 \) and \( T_2 = 106 \) observations. Use \( T_1 \) as an estimation sample, and \( T_2 \) as a prediction sample for 100 different random splits.

Comparison of different models

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Score Mean</th>
<th>Log Score Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonparametric</td>
<td>-293.17</td>
<td>-289.44</td>
</tr>
<tr>
<td>FMMN(m=3)</td>
<td>-292.50</td>
<td>-291.93</td>
</tr>
<tr>
<td>FMMN(m=4)</td>
<td>-284.55</td>
<td>-284.58</td>
</tr>
<tr>
<td>FMMN(m=5)</td>
<td>-283.57</td>
<td>-283.27</td>
</tr>
<tr>
<td>FMMN(m=6)</td>
<td>-283.21</td>
<td>-283.18</td>
</tr>
<tr>
<td>FMMN(m=ML)</td>
<td>-283.50</td>
<td>-283.15</td>
</tr>
<tr>
<td>SMR(m=4)</td>
<td>-280.29</td>
<td>-280.31</td>
</tr>
<tr>
<td>SMR(m=7)</td>
<td>-280.23</td>
<td>-279.19</td>
</tr>
</tbody>
</table>

Daily returns on S&P 500

Results from Geweke and Keane (2007)

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictive likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMR</td>
<td>-1602.0</td>
</tr>
<tr>
<td>t-GARCH(1,1)</td>
<td>-1625.5</td>
</tr>
<tr>
<td>Threshold EGARCH(1,1)</td>
<td>-1637.5</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>-1660.5</td>
</tr>
<tr>
<td>Normal iid</td>
<td>-1848.5</td>
</tr>
</tbody>
</table>

Summary

- Bayesian framework provides a rich collection of non- and semi-parametric methods.
- Some of these methods deliver familiar classical estimators. This can help understand assumptions under which the estimators are appropriate in finite samples.
- In small samples, mixture based methods compare favorably with classical parametric and nonparametric alternatives.
- We still have a lot to learn: posterior convergence rates, flexible priors in structural models, moment (in)equality models, ...
References I


References II


References III


