Adaptive Bayesian Estimation of Mixed Discrete-Continuous Distributions under Smoothness and Sparsity

Andriy Norets and Justinas Pelenis

2018
Motivation and Questions

- Bayesian models based on mixtures:
  - convenient computationally
  - posterior contracts at optimal minimax rate (up to log) for smooth true densities (Shen, Tokdar, and Ghosal (2013), STG)
  - can be used for modeling discrete data through continuous latent variables

- Is it a good idea to use mixture models for discrete data?
- Will posterior contract at an optimal rate?
- Appropriate settings for asymptotics? Optimal rates?
Summary of Results

▶ Data Generating Process
  ▶ support of discrete variables can become finer with $n$
    (sparse multinomials as in Hall and Titterington (1987))
  ▶ probability mass function is “smooth”

▶ We establish lower bounds on estimation rates for multivariate
  discrete-continuous anisotropic distributions

▶ For mixture models, posterior contraction rates are equal to
  the derived lower bounds up to a log factor

▶ Excellent finite sample performance in simulations
DGP

- Continuous $x \in \mathbb{R}^{d_x}$
- Discrete $y = (y_1, \ldots, y_{d_y})$, $y_k \in \left\{ \frac{1-1/2}{N_k}, \frac{2-1/2}{N_k}, \ldots, \frac{N_k-1/2}{N_k} \right\}$.
- $A_y$ - rectangle with center $y$ and side lengths $(\frac{1}{N_1}, \ldots, \frac{1}{N_{d_y}})$,
  $[0, 1]^{d_y} = \bigcup_y A_y$
- DGP density-probability mass function

$$p_0(x, y) = \int_{A_y} f_0(x, \tilde{y}) d\tilde{y},$$

where $f_0$ is a density on $\mathbb{R}^{d_x} \times [0, 1]^{d_y}$

- So far, without loss of generality.
Anisotropic \((\beta_1, \ldots, \beta_d)\)-Holder Class \(C^{L, \beta_1, \ldots, \beta_d}\)

\(f \in C^{L, \beta_1, \ldots, \beta_d}\) if for any \(k = (k_1, \ldots, k_d)\), \(\sum_{i=1}^{d} k_i/\beta_i < 1\),

\[|D^k f(z + \Delta z) - D^k f(z)| \leq L \sum_{j=1}^{d} |\Delta z_j|^\beta_j(1 - \sum k_i/\beta_i)\]

where \(\Delta z_j = 0\) when \(\sum_{i=1}^{d} k_i/\beta_i + 1/\beta_j < 1\).

- Is this definition standard?
- Ibragimov and Hasminskii (1984) did not restrict mixed derivatives
- When \(\beta_j = \beta, \forall j, \beta_j(1 - \sum k_i/\beta_i) = \beta - \lfloor \beta \rfloor\), get standard definition for isotropic case.
Anisotropic \((\beta_1, \ldots, \beta_d)\)-Holder Class \(C^{L,\beta_1,\ldots,\beta_d}\)

\(f \in C^{L,\beta_1,\ldots,\beta_d}\) if for any \(k = (k_1, \ldots, k_d)\), \(\sum_{i=1}^{d} k_i / \beta_i < 1\),

\[
|D^k f(z + \Delta z) - D^k f(z)| \leq L \sum_{j=1}^{d} |\Delta z_j|^{\beta_j(1 - \sum_{i=1}^{d} k_i / \beta_i)}
\]

where \(\Delta z_j = 0\) when \(\sum_{i=1}^{d} k_i / \beta_i + 1 / \beta_j < 1\).

- STG use \(|\Delta z_j|^{\min(\beta_j - k_j, 1)}\) instead of \(|\Delta z_j|^{\beta_j(1 - \sum_{i=1}^{d} k_i / \beta_i)}\)
- It would not work in our proof of lower bounds
- Anisotropic Taylor expansion used in the proof of upper bounds can be obtained under our assumption
Theorem 1: Lower Bound on Estimation Rate

- $\mathcal{A}$ - collection of all subsets of $\{d_x + 1, \ldots, d\}$, $d = d_x + d_y$
- For $J \in \mathcal{A}$, define $J^c = \{1, \ldots, d\} \setminus J$,

$$N_J = \prod_{k \in J} N_k, \quad \beta_{J^c} = \left[ \sum_{k \in J^c} \beta_k^{-1} \right]^{-1},$$

$\beta_{\emptyset} = \infty$, and $N_{\emptyset} = 1$.
- Lower bound in TVD and Hellinger is

$$\min_{J \in \mathcal{A}} \left[ \frac{N_J}{n} \right]^{\frac{\beta_{J^c}}{2\beta_{J^c} + 1}} = \left[ \frac{N_{J^*_*}}{n} \right]^{\frac{\beta_{J^*_*}}{2\beta_{J^*_*} + 1}}$$
Special Case of Lower Bound: $d_x = 0, d_y = 1$

$$\min \left\{ \frac{N_1}{n}^{1/2}, n^{-\frac{\beta_1}{2\beta_1+1}} \right\}$$

- Parametric rate if $N_1$ is constant
- If $N_1$ is sufficiently fast then standard rate for estimation of $\beta_1$-smooth functions
- Hall and Titterington (1987) obtained this lower bound for mean summed square error under slightly different smoothness definition
Special Case of Lower Bound: \( d_x = 1, d_y = 1 \)

\[
\min \left\{ \left[ \frac{N_1}{n} \right]^{\frac{\beta_2}{2\beta_2 + 1}}, n^{-\frac{\beta_{1,2}}{2\beta_{1,2} + 1}} \right\}
\]

- Approximately, \( n/N_1 \) observations should be available for estimating \( \beta_2 \)-smooth conditional densities of \( x \mid y \).
- If \( N_1 \) is sufficiently fast then we can exploit smoothness in \( y \).
Frequentist Literature on Smoothing Sparse Discrete Data

- $N_k$’s are constant: Aitchison and Aitken (1976), Hall, Racine, and Li (2004) (cross-validation), Efromovich (2011)
- Burman (1987): lower bounds and discrete kernels for $d_y = 1$, $d_x = 0$, $\beta_1 = 2$.
- Hall and Titterington (1987): lower bounds and discrete kernels for $d_y = 1$, $d_x = 0$, general $\beta_1$; cross-validated bandwidth selection for $\beta_1 = 2$.
- Dong and Simonoff (1995): upper bounds for $d_x = 0$, $d_y \leq 4$, $\beta_1 = \ldots = \beta_{d_y} = 4$, fast $N_k$’s
- Aerts et al. (1997): local polynomial smoothers for $d_x = 0$, general $d_y$, $\beta_1 = \ldots = \beta_{d_y}$, fast $N_k$’s.
Proof Sketch for Lower Bound

Th. 2.5 in Tsybakov (2008) (Ibragimov and Hasminskii (1977)):

\[
\inf_{\hat{p}} \sup_{p_0 \in \mathcal{P}} P(d(\hat{p}, p_0) \geq \Gamma_n) \geq \text{const} > 0, \quad \text{if} \\
\exists q_j, q_k \in \mathcal{P}, \ 0 \leq j < k \leq M \\
d(q_j, q_k) \geq 2\Gamma_n, \\
\sum_{j=1}^{M} KL(q_j, q_0)/M < \log(M)/8
\]
Proof Sketch for Lower Bound

\[ q_j(x, y) = \int_{A_y} \left[ 1 + \Gamma_n \cdot \sum_{i} w_i^j \prod_{r=1}^{d} g(m_r(x_r - c_i^r)) \right] dx_{d+1:d} \]

- \( w_i^0 = 0, \ w_i^j \in \{0, 1\}, \ i \in \{1, \ldots, \prod_{r=1}^{d} m_r\} \),
- \( c_i \) - center of rectangle with sides \((1/m_1, \ldots, 1/m_d)\)
- \( g \) is infinitely smooth on \([-1/2, 1/2]\), 0 elsewhere, \( \int g = 0 \)
- \( m_r = \Gamma_n^{-1/\beta_r^*} \)
- \( \beta_r^* = \beta_r \) for \( r \notin J_* \) (\( x_r \) - (treated as) continuous)
- \( \beta_r^* = -\log(\Gamma_n)/\log N_r \) for \( r \in J_* \) (smoothness at which we would be indifferent to treating \( x_r \) as continuous, \( \beta_r^* \geq \beta_r \))
- The rest of the proof is similar to the continuous case.
Model

For $y \in \mathbb{R}^{d_y}$ and $x \in \mathbb{R}^{d_x}$, 

$$p(x, y|\theta, m) = \int_{A_y} \sum_{j=1}^{m} \alpha_j \phi(x, \tilde{y}; \mu_j, \sigma) d\tilde{y},$$

where $\theta = (\alpha_1, \mu_1, \ldots, \alpha_m, \mu_m, \sigma)$, $\mu_j \in \mathbb{R}^d$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)$ and $\phi$ - normal density with diagonal covariance.

- $\Pi(\theta|m)$
- $\Pi(m)$
- (Dirichlet process mixture should also work)
MCMC Estimation through Data Augmentation

- $X^n = (x_1, \ldots, x_n), \ Y^n = (y_1, \ldots, y_n)$
- Explicitly use latent variables $\tilde{y} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$
- Introduce mixture allocation latent variables: $s = (s_1, \ldots, s_n)$,

$$x_i, \tilde{y}_i | s_i = j, \theta, m \sim N(\mu_j, \sigma)$$

- Gibbs sampler for $\theta, \tilde{y}, s | m, X^n, Y^n$
- $\theta | m, \tilde{y}, s, X^n, Y^n$, same as in simple Normal model.
- $\tilde{y}_i | \ldots \sim N(\mu_{s_i}, \sigma) \cdot 1_{A_y}$
- $P(s_i = j | \ldots) \propto \alpha_j \phi(x_i, \tilde{y}_i; \mu_j, \sigma)$
- Reversible jump for $m$ (or Dirichlet process mixture)
Assumptions on DGP for Upper Bound

For $J = \{d'+1, \ldots, d_y\}$, $y = (y_{J^c}, y_J)$, define marginal pmf

$$\pi_0(y_J) = \int \int_{A_{y_J}} f_0(x, \tilde{y}) d\tilde{y} dx$$

and conditional pdf

$$f_{0|J}(x, \tilde{y}_{J^c}|y_J) = \int_{A_{y_J}} f_0(x, \tilde{y}) d\tilde{y} / \pi_0(y_J)$$

Assume that for any $y_J$

- $0 < \frac{\pi}{\bar{N}_J} \leq \pi_0(y_J) \leq \frac{\bar{\pi}}{\bar{N}_J} < \infty$
- $f_{0|J}(\cdot|y_J) \in C^{L,\beta_1,\ldots,\beta_{d_x}+d'} \iff f_0 \in C^{L,\beta_1,\ldots,\beta_d}$
**Upper Bound on Posterior Contraction Rate**

$\epsilon_n$ is an upper bound on the posterior contraction rate if

$$\prod \left( p : d(p_0, p) > \text{const} \cdot \epsilon_n \left| Y^n, X^n \right. \right) \overset{Pr}{\rightarrow} 0.$$  

**Theorem 2**: under standard assumptions on priors and smoothness assumptions on $f_0$ from the previous slide

$$\epsilon_n = \left[ \frac{N_J}{n} \right]^{\frac{\beta Jc}{2\beta Jc + 1}} \cdot (\log n)^t,$$

which coincides with the lower bound up to $(\log n)^t$ when $J = J_\ast$.

($x$ can have unbounded support but with sub-exponential tails and envelope function $L$ that behaves like $f_0$ in the tails)
Previous Posterior Asymptotics Results for Constant $N_k$’s

- Norets and Pelenis (2012) - weak consistency for mixtures with a variable number of components
- DeYoreo and Kottas (2017) - weak consistency for Dirichlet process mixtures
- Canale and Dunson (2015) - contraction rates for Dirichlet process mixtures (dimension in their rate is $d_y + d_x$, which is non-optimal for constant $N_k$’s)
Assumptions on Prior

- \( \Pi(m = i) \propto \exp(-b_1 i (\log i)^{\tau_1}) \)
- \( \Pi(\alpha_1, \ldots, \alpha_m | m) \) is Dirichlet\((a/m, \ldots, a/m)\), \( a > 0 \)
- Prior density for locations \( \mu^X_{jr} \) is bounded below by
  \[
  \exp(-b_2 \mu^{T_2})
  \]
  and
  \[
  1 - \Pi(\mu^X_j \in [-x, x]^{dx}) \leq \exp(-b_3 x^{T_3})
  \]
- Prior density for locations \( \mu^Y_j \) is bounded away from zero on \([0, 1]^{dy}\).
- Prior for \( \sigma_r \) is inverse Gamma (not a standard conditionally conjugate prior).
Proof: Sufficient Conditions

Ghosal, Ghosh, and Vaart (2000): posterior contracts at rate $\epsilon_n$ if

- $Z_i \overset{iid}{\sim} p_0, Z^n = (Z_1, \ldots, Z_n)$
- $p_0 \in \mathcal{P}$ - space of densities w.r.t. a $\sigma$-finite measure
- $d$ - Hellinger or total variation distance
- $\mathcal{P}_n$ is a sieve satisfying
  \[
  \log J(\epsilon_n, \mathcal{P}_n, d) \leq c_1 n\epsilon_n^2 \quad (J \text{- metric entropy})
  \]
  \[
  \Pi(\mathcal{P}_n^c) \leq c_3 \exp\{- (c_2 + 4) n\epsilon_n^2\}
  \]
- Prior thickness condition for Kullback-Leibler neighborhoods
  \[
  \mathcal{K}(p_0, \epsilon_n) = \left\{ p : \int p_0 \log(p_0/p) < \epsilon_n^2, \int p_0 [\log(p_0/p)]^2 < \epsilon_n^2 \right\}
  \]
  \[
  \Pi(\mathcal{K}(p_0, \epsilon_n)) \geq c_4 \exp\{- c_2 n\epsilon_n^2\}
  \]
Proof: Approximation Idea

Approximation results are key, e.g., need to find \((\theta^*, m)\) s.t.
\[
KL(p_0(x, y), p(x, y|\theta^*, m)) \leq \epsilon_n^2.
\]

Consider first \(J = \{d_x + 1, \ldots, d\}\).

- \(p_0(x, y) = \pi_0(y)p_0(x|y)\)
- For \(\sigma^y \to 0\), \(\int_{A_y} \phi(\tilde{y}, y', \sigma^y) \approx 1\) when \(y = y', 0\) otherwise.

\[
\pi_0(y) \approx \int_{A_y} \sum_{y'} \pi_0(y')\phi(\tilde{y}, y', \sigma^y) d\tilde{y}
\]

- From STG: \(\forall y'\),

\[
p_0(x|y') \approx \sum_{j=1}^{m_x} \alpha_{j|y'} \phi(x; \mu_{j|y'}, \sigma^x)
\]
Proof: Approximation Idea

Combine the approximations from the previous slide into

\[ p_0(x, y) \approx \int_{A_y} \sum_{y'} \sum_{j=1}^{m_x} \alpha_{j\mid y'} \pi_0(y') \phi(x; \mu_{j\mid y'}, \sigma^x) \phi(\tilde{y}, y', \sigma^y) d\tilde{y} \]

\[ = p(x, y \mid \theta^*, m) \]

Next, need to find a neighborhood of \( \theta^* \), \( S_{\theta^*} \), for which approximation error is still below \( \epsilon_n \) and its prior probability

\[ \geq \exp\{-c_2 n \epsilon_n^2\} \].
Proof: Prior Probability of KL neighborhoods

For example, consider $m$ (isotropic case, $\beta_j = \beta$).

- If we need approximation error $\Gamma_n \cdot \log(n)^t$ for the conditionals, where $\Gamma_n = [NJ/n]^{2\beta + d_x}$, from STG:

$$m_x = c_1 \Gamma_n^{-\frac{d_x}{\beta}} \cdot (\log n)^{c_2}$$

- Total # of mixture components: $m = NJ \cdot c_1 \Gamma_n^{-\frac{d_x}{\beta}} \cdot (\log n)^{c_2}$

- $\Pi(m) = \exp(-m) \geq \exp(-n[\Gamma_n \log(n)^t]^2) \iff$

$$c_1 NJ \Gamma_n^{-\frac{d_x}{\beta}} \cdot (\log n)^{c_2} \leq n[\Gamma_n \log(n)^t]^2 \iff$$

$$c_1 NJ/n(\log n)^{c_2-2t} \leq \Gamma_n^{2+d_x/\beta} \iff$$

$$t > c_2/2$$
Proof: \( J \neq \{d_x + 1, \ldots, d\} \)

- Approximation argument above is easy to adapt
- (Hellinger, TVD, KL) distances and ratios for mixed discrete-continuous distributions are bounded by distances and ratios for the corresponding latent variable densities.
- Bounds on entropy for mixture of multivariate normals from previous literature also apply for the same reason (to \( J = \{d_x + 1, \ldots, d\} \) case as well).
Evaluating Model Quality

- Cross Validated Log Scoring Rule
\[
\sum_{i=1}^{n} \log p(z_i | Z^{n/i}) \approx \sum_{i=1}^{n} \log \frac{1}{K} \sum_{k=1}^{K} p(z_i | Z^{n/i}, \theta^k)
\]

- We use: Modified Cross Validated Log Scoring Rule:
  Randomly order sample observations and use the first \( n_1 \) observations for inference and the rest for evaluation. Repeat this process several times and compare means or medians.
\[
\sum_{i=n_1+1}^{n} \log p(z_i | Z^{n_1})
\]
Labor Market Participation

- Source: Norets and Pelenis (2012)
- 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 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:** Modified cross-validated log scores and classification rates

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Score Mean</th>
<th>Log Score Median</th>
<th>% Correct pred-ns Mean</th>
<th>% Correct pred-ns Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probit</td>
<td>-137.23</td>
<td>-136.69</td>
<td>66.08%</td>
<td>66.37%</td>
</tr>
<tr>
<td>Kernel</td>
<td>-138.21</td>
<td>-135.99</td>
<td>65.91%</td>
<td>65.77%</td>
</tr>
<tr>
<td>FMMN(m=1)</td>
<td>-137.27</td>
<td>-136.81</td>
<td>66.02%</td>
<td>65.77%</td>
</tr>
<tr>
<td>FMMN(m=2)</td>
<td>-132.30</td>
<td>-131.86</td>
<td>67.95%</td>
<td>68.02%</td>
</tr>
<tr>
<td>FMMN(m=3)</td>
<td>-133.32</td>
<td>-132.60</td>
<td>67.76%</td>
<td>67.57%</td>
</tr>
<tr>
<td>FMMN(m=4)</td>
<td>-133.13</td>
<td>-131.86</td>
<td>68.21%</td>
<td>68.02%</td>
</tr>
</tbody>
</table>
Future Work

- Check that results go through for Dirichlet process mixtures.
- Simulations/applications for variable $m$ or Dirichlet process mixtures.
- Extend results from Norets and Pati (2017) for continuous conditional densities to mixed discrete-continuous case.
- Implement MCMC for direct estimation of conditional distributions (extend Norets (2017)).
Power of log in the rate for continuous case

\[ \epsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t \]

\[ t > \frac{d(1 + 1/\beta + 1/\tau)}{2 + d/\beta} + \max\left\{ \tau_1, 1, \tau_2/\tau \right\} + \max\left\{ 0, \frac{1 - \tau_1}{2} \right\} \]

- \( \beta \) - smoothness level
- \( d \) - dimension of \((y, x)\)
- \( \tau \): \( f_0(z) \leq c \exp(-b\|z\|^\tau) \)
- \( \tau_1 \): \( \prod(m = i) \propto \exp(-b_1 i (\log i)^{\tau_1}) \)
- \( \tau_2 \): \( \exp(-b_2 \mu^{\tau_2}) \leq \) prior density for \( \mu_{jk}^y \).
References I


