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

Andriy Norets and Justinas Pelenis

November, 2017

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 $k = (k_1, \ldots, k_d)$, $\sum_{i=1}^{d} k_i/\beta_i < 1$,

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

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 $|\Delta z_j|^{\beta_j (1 - \sum k_i/\beta_i)}$
- 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 $J^c = \{1, \ldots, d\} \setminus J$,

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

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

$$\min_{J \in \mathcal{A}} \left[\frac{N_J}{n}\right]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}} = \left[\frac{N_{J_*}}{n}\right]^{\frac{\beta_{J_*^c}}{2\beta_{J_*^c}+1}}$$

Special Case of Lower Bound: $d_x = 0$, $d_y = 1$

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

- Parametric rate if N_1 is constant
- If N_1 is sufficiently fast then standard rate for estimation of β_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: $d_x = 1$, $d_y = 1$

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

- Approximately, n/N_1 observations should be available for estimating β_2-smooth conditional densities of $x|y$.
- 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)
- 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 β_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)):

$$\inf_{\hat{p}} \sup_{p_0 \in \mathcal{P}} P(d(\hat{p}, p_0) \geq \Gamma_n) \geq const > 0, \quad if$$

- $\exists q_j, q_k \in \mathcal{P}$, $0 \leq j < k \leq M$
- $d(q_j, q_k) \geq 2\Gamma_n$,
- $\sum_{j=1}^{M} KL(q_j, q_0)/M < \log(M)/8$

Proof Sketch for Lower Bound

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

- $w_i^0 = 0$, $w_i^j \in \{0, 1\}$, $i \in \{1, \ldots, \prod_{r=1}^{d} m_r\}$,
- c_i - center of rectangle with sides $(1/m_1, ..., 1/m_d)$
- g is infinitely smooth on $[-1/2, 1/2]$, 0 elsewhere, $\int g = 0$
- $m_r = s^{-1/\beta_r^*}$
- $\beta_r^* = \beta_r$ for $r \notin J_*$ (x_r - (treated as) continuous)
- $\beta_r^* = -\log(s)/\log N_r$ 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 ϕ - 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

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

Assume that for any y_J

- $0 < \frac{\underline{\pi}}{N_J} \leq \pi_0(y_J) \leq \frac{\overline{\pi}}{N_J} < \infty$
- $f_{0|J}(\cdot | y_J) \in C^{L, \beta_1, \ldots, \beta_{d_x + d'}}$ $(\Leftarrow f_0 \in C^{L, \beta_1, \ldots, \beta_d})$

Upper Bound on Posterior Contraction Rate

ϵ_n is an upper bound on the posterior contraction rate if

$$\Pi\left(p: \ d(p_0, p) > const \cdot \epsilon_n \middle| Y^n, X^n\right) \stackrel{Pr}{\to} 0.$$

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

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

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

(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 μ^x_{jr} is bounded below by

  $$\exp(-b_2 \mu^{\tau_2})$$

  and

  $$1 - \Pi(\mu^x_j \in [-x, x]^{d_x}) \leq \exp(-b_3 x^{\tau_3})$$
- Prior density for locations μ^y_j is bounded away from zero on $[0, 1]^{d_y}$.
- Prior for σ_r is inverse Gamma (not a standard conditionally conjugate prior).

Proof: Sufficient Conditions

Ghosal, Ghosh, and Vaart (2000): posterior contracts at rate ϵ_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 σ-finite measure
- d - Hellinger or total variation distance
- $\mathcal{P}_n$ is a sieve satisfying

$$\log J(\epsilon_n, \mathcal{P}_n, d) \le c_1 n \epsilon_n^2 \quad (J \text{ - metric entropy})$$

$$\Pi(\mathcal{P}_n^c) \le c_3 \exp\{-(c_2 + 4) n \epsilon_n^2\}$$

- Prior thickness condition for Kullback-Leibler neighborhoods

$$\mathcal{K}(p_0, \epsilon_n) = \left\{ p : \int p_0 \log(p_0/p) < \epsilon_n^2, \int p_0 [\log(p_0/p)]^2 < \epsilon_n^2 \right\}$$

$$\Pi(\mathcal{K}(p_0, \epsilon_n)) \ge c_4 \exp\{-c_2 n \epsilon_n^2\}$$

Proof: Approximation Idea

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

Consider first $J = \{d_x + 1, \dots, d\}$.

- $p_0(x,y) = \pi_0(y) p_0(x|y)$
- 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

$$
\begin{aligned}
p_0(x,y) &\approx \int_{A_y} \sum_{y'} \sum_{j=1}^{m_x} \alpha_{j|y'} \pi_0(y') \phi(x; \mu_{j|y'}, \sigma^x) \phi(\tilde{y}, y', \sigma^y) d\tilde{y} \\
&= p(x, y|\theta^*, m)
\end{aligned}
$$

Next, need to find a neighborhood of θ^*, S_{θ^*}, for which approximation error is still below ϵ_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 $\Gamma_n = [N_J/n]^{\frac{\beta}{2\beta+d_x}}$, from STG:

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

- Total $\#$ of mixture components: $m = N_J \cdot c_1 \Gamma_n^{-d_x/\beta} (\log n)^{c_2}$
- $\Pi(m) = \exp(-m) \geq \exp(-n[\Gamma_n \log(n)^t]^2) \Leftrightarrow$

$$c_1 N_J \Gamma_n^{-d_x/\beta} (\log n)^{c_2} \leq n[\Gamma_n \log(n)^t]^2 \quad \Leftrightarrow$$

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

$$t > c_2/2$$

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

Model	Log Score		% Correct pred-ns	
	Mean	Median	Mean	Median
Probit	-137.23	-136.69	66.08%	66.37%
Kernel	-138.21	-135.99	65.91%	65.77%
FMMN(m=1)	-137.27	-136.81	66.02%	65.77%
FMMN(m=2)	-132.30	-131.86	67.95%	68.02%
FMMN(m=3)	-133.32	-132.60	67.76%	67.57%
FMMN(m=4)	-133.13	-131.86	68.21%	68.02%

Future Work

- Check that results go though 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

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

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

- β - smoothness level
- d - dimension of (y, x)
- τ: $f_0(z) \leq c \exp(-b||z||^\tau)$
- τ_1: $\Pi(m = i) \propto \exp(-b_1 i (\log i)^{\tau_1})$
- τ_2: $\exp(-b_2 \mu^{\tau_2}) \leq$ prior density for μ^y_{jk}.

References I

Aerts, M., I. Augustyns, and P. Janssen (1997): "Local Polynomial Estimation of Contingency Table Cell Probabilities," *Statistics*, 30, 127–148.

Aitchison, J. and C. G. G. Aitken (1976): "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Burman, P. (1987): "Smoothing Sparse Contingency Tables," *Sankhy?: The Indian Journal of Statistics, Series A (1961-2002)*, 49, 24–36.

Canale, A. and D. B. Dunson (2015): "Bayesian multivariate mixed-scale density estimation," *Statistics and its Interface*, 8, 195–201.

DeYoreo, M. and A. Kottas (2017): "Bayesian Nonparametric Modeling for Multivariate Ordinal Regression," *Journal of Computational and Graphical Statistics*, 0, 1–14.

Dong, J. and J. S. Simonoff (1995): "A Geometric Combination Estimator for d-Dimensional Ordinal Sparse Contingency Tables," *Ann. Statist.*, 23, 1143–1159.

Gerfin, M. (1996): "Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation," *Journal of Applied Econometrics*, 11, 321–339.

Ghosal, S., J. K. Ghosh, and A. W. v. d. Vaart (2000): "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531.

Hall, P., J. Racine, and Q. Li (2004): "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hall, P. and D. M. Titterington (1987): "On Smoothing Sparse Multinomial Data," *Australian Journal of Statistics*, 29, 19–37.

Ibragimov, I. and R. Hasminskii (1977): "Estimation of infinite-dimensional parameter in Gaussian white noise," *Doklady Akademii Nauk SSSR*, 236, 1053–1055.

Ibragimov, I. A. and R. Z. Hasminskii (1984): "More on the estimation of distribution densities," *Journal of Soviet Mathematics*, 25, 1155–1165.

Norets, A. (2017): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models," Unpublished manuscript, Brown University.

References II

Norets, A. and D. Pati (2017): "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 9801012.

Norets, A. and J. Pelenis (2012): "Bayesian modeling of joint and conditional distributions," *Journal of Econometrics*, 168, 332–346.

Shen, W., S. T. Tokdar, and S. Ghosal (2013): "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures," *Biometrika*, 100, 623–640.

Tsybakov, A. B. (2008): *Introduction to Nonparametric Estimation (Springer Series in Statistics)*, Springer, New York, USA.