Submitted to Econometrica

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

Andriy Norets and Justinas Pelenis

August 26, 2021

Submitted to Econometrica

ADAPTIVE BAYESIAN ESTIMATION OF DISCRETE-CONTINUOUS DISTRIBUTIONS UNDER SMOOTHNESS AND SPARSITY

ANDRIY NORETS[1]

AND JUSTINAS PELENIS

We consider nonparametric estimation of a mixed discrete-continuous distribution under anisotropic smoothness conditions and a possibly increasing number of support points for the discrete part of the distribution. For these settings, we derive lower bounds on the estimation rates. Next, we consider a nonparametric mixture of normals model that uses continuous latent variables for the discrete part of the observations. We show that the posterior in this model contracts at rates that are equal to the derived lower bounds up to a log factor. Thus, Bayesian mixture of normals models can be used for (up to a log factor) optimal adaptive estimation of mixed discrete-continuous distributions. The proposed model demonstrates excellent performance in simulations mimicking the first stage in the estimation of structural discrete choice models.

KEYWORDS: Bayesian nonparametrics, adaptive rates, minimax rates, anisotropic smoothness, posterior contraction, discrete-continuous distribution, mixed scale, mixtures of normal distributions, latent variables, discrete choice models.

.

1. INTRODUCTION

Nonparametric estimation methods have become more accessible and useful in empirical work due to availability of fast computers and very large datasets. The theory and practical implementation of nonparametric methods for continuous data are very well developed at this point. However, in most economic applications, the data contain both continuous and discrete variables. Nonparametric methods for multivariate discrete and mixed discrete-continuous distributions and their theoretical properties are less well understood and developed. We address this issue in the present paper.

The standard flexible approach to estimation of discrete distributions is to use sample frequencies as estimators of the corresponding probabilities. These estimators do not perform well in the case where the number of values that discrete variables can take is larger

andriy_norets@brown.edu
pelenis@ihs.ac.at

[1] We thank Jesse Shapiro and participants of seminars at Harvard-MIT and Yale and OBayes conferences for helpful discussions. We also thank three anonymous referees for suggestions that improved the manuscript. The first author gratefully acknowledges the support from NSF Grant SES-1851796.

or comparable to the sample size, which we, following Hall and Titterington (1987), refer to as sparsity. The sparsity in the multivariate case is rather a rule than an exception; for example, estimating a joint distribution of 5 discrete variables each taking 10 values would involve estimation of 10^5 probabilities by the corresponding sample frequencies. The presence of continuous variables in addition to the discrete ones further exacerbates the problem. In economics, these issues often arise in the context of estimation of single-agent and game-theoretic static and dynamic discrete choice models. Popular two stage estimation procedures for these models pioneered by Hotz and Miller (1993) deal with discrete dependent variables such as market entry decisions and discrete covariates such as the number of entrants currently in the market. A natural solution to this problem that appears to work well in practice (Aitchison and Aitken (1976), Li and Racine (2007)) is to smooth discrete data, hoping that probabilities at nearby discrete values are close or smooth in some sense and that one could learn about a probability of a certain value from the observations at nearby values. Of course, smoothing can only be beneficial if the underlying data have certain smoothness properties. Ideally, a procedure for estimation of discrete distributions should be able to optimally take advantage of smoothness in the data generating process if it is present and at the same time perform no worse than the standard frequency estimators if the data generating process is not (sufficiently) smooth.

In this paper, we formalize these ideas for multivariate mixed discrete-continuous distributions by setting up an asymptotic framework where the multivariate discrete part of the data generating distribution can have either a large or a small number of support points and it can be either very smooth or not, and these characteristics can differ from one discrete coordinate to another. In these settings, we derive optimal minimax rates for estimation of discrete-continuous distributions. We show that smoothing is beneficial only for a subset of discrete variables with a quickly growing number of support points and/or sufficiently high level of smoothness.

We propose an estimation procedure that adaptively (without a priori knowledge of smoothness levels of the data generating process) achieve the derived optimal convergence rates. The procedure is based on a Bayesian mixture of multivariate normal distributions. Mixture models have proven to be very useful for Bayesian nonparametric modeling of univariate and multivariate distributions of continuous variables. These models possess outstanding asymptotic frequentist properties: in Bayesian nonparametric estimation of smooth densities the posterior in these models contracts at optimal adaptive rates up to a log factor (Rousseau (2010), Kruijer et al. (2010), Shen, Tokdar, and Ghosal (2013)). Tractable Markov chain Monte Carlo (MCMC) algorithms for exploring posterior distributions of these models are available and they are widely used in empirical work (see Dey, Muller, and Sinha (1998)).

From the computational perspective, discrete variables can be easily accommodated through the use of continuous latent variables in Bayesian MCMC estimation (Albert and Chib (1993)). In nonparametric modelling of discrete-continuous data by mixtures, latent variables were used by Canale and Dunson (2011) and Norets and Pelenis (2012) among others. Some results on frequentist asymptotic properties of the posterior distribution in such models have also been established. Norets and Pelenis (2012) obtained approximation results in Kullback-Leibler distance and weak posterior consistency for mixture models with a prior on the number of mixture components. DeYoreo and Kottas (2017) establish weak posterior consistency for Dirichlet process mixtures. In similar settings, Canale and Dunson (2015) derived posterior contraction rates that are not optimal. In the present paper, we show that a mixture of normals model with a prior on the number of mixture components that uses latent variables for modeling the discrete part of the distribution can deliver optimal posterior contraction rates for nonparametric estimation of discrete-continuous distributions. The obtained optimal posterior contraction rates are adaptive since the priors we consider do not depend on the size of the support and the smoothness of the data generating process.

We illustrate our theoretical results in an application to the first stage estimation of discrete choice models. Specifically, we use data from Monte Carlo experiments in Pakes, Ostrovsky, and Berry (2007) who compare various two stage estimation procedures on a model of firm's entry decisions. Our procedure delivers 2.5 times reduction in the estimation error relative to the frequency estimator. Overall, our theoretical and simulation results suggest that models for discrete data based on mixtures and latent variables should be an important part of the econometric toolkit.

The rest of the paper is organized as follows. In Section 2, we describe our framework and the Bayesian model. Section 3 presents simulation results and favorable comparisons with frequency and kernel estimators. The asymptotic theoretical results are presented in Section 4. MCMC algorithm for model estimation and proof outlines are given in Appendices. Auxiliary results and proof details are delegated to the online supplement.

2. DATA GENERATING PROCESS AND BAYESIAN MODEL

Let us denote the continuous part of observations by $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$ and the discrete part by $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$, where

$$\mathcal{Y} = \prod_{j=1}^{d_y} \mathcal{Y}_j, \text{ with } \mathcal{Y}_j = \left\{ \frac{1 - 1/2}{N_j}, \frac{2 - 1/2}{N_j}, \ldots, \frac{N_j - 1/2}{N_j} \right\},$$

is a grid on $[0,1]^{d_y}$ (a product symbol Π applied to sets hereafter denotes a Cartesian product). The number of values that the discrete coordinates y_j can take, N_j, can potentially grow with the sample size or stay constant. For each discrete coordinate value $y_j \in \mathcal{Y}_j$, let

$$A_{y_j} = \begin{cases} (-\infty, y_j + 0.5/N_j] & \text{if } y_j = 0.5/N_j \\ (y_j - 0.5/N_j, \infty) & \text{if } y_j = 1 - 0.5/N_j \\ (y_j - 0.5/N_j, y_j + 0.5/N_j] & \text{otherwise} \end{cases}$$

be an interval that includes y_j and has a length of $1/N_j$, except for the first and the last intervals that are expanded to include the rest of the negative and positive parts of the real line correspondingly. Then, every value of the discrete part of observations $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$ can be associated with a hyper-rectangle $A_y = \prod_{j=1}^{d_y} A_{y_j}$. Let us represent the data generating density-probability mass function $p_0(y,x)$ as an integral of a latent density f_0 over A_y,

$$p_0(y,x) = \int_{A_y} f_0(\tilde{y}, x) d\tilde{y}, \tag{1}$$

where f_0 belongs to the set of probability density functions (pdf) on $\mathbb{R}^d$ with respect to the Lebesgue measure, and $d = d_x + d_y$. The representation of a mixed discrete-continuous distribution in (1) is so far without a loss of generality since for any given p_0 one could always define f_0 using a mixture of densities with non-overlapping supports included in A_y, $y \in \mathcal{Y}$.

We assume that the data available for estimation of p_0 are comprised of n independently identically distributed observations from p_0: $(Y^n, X^n) = (Y_1, X_1, \ldots, Y_n, X_n)$. Let P_0, E_0, P_0^n, and E_0^n denote the probability measures and expectations corresponding to p_0 and its product p_0^n.

When N_j's grow with the sample size n the generality of the representation in (1) can be lost when assumptions such as smoothness are imposed on f_0. Nevertheless, in what follows we do allow for f_0 to be smooth. The interpretation of the smoothness is that the values of discrete variables can be ordered and that borrowing of information from nearby discrete points can be useful in estimation.

2.1. *Bayesian Model*

Our nonparametric Bayesian model for the data generating process in (1) is based on a mixture of normal distributions with a variable number of components for modelling the joint distribution of $(\tilde{y}, x)$,

$$f(\tilde{y}, x|\theta, m) = \sum_{k=1}^{m} \alpha_k \phi(\tilde{y}, x; \mu_k, \sigma \cdot \nu_k^{-1/2})$$

$$p(y, x|\theta, m) = \int_{A_y} f(\tilde{y}, x|\theta, m) d\tilde{y}, \tag{2}$$

where $\theta = (\mu_k, \nu_k, \alpha_k, k = 1, 2, \ldots; \sigma)$ and $\phi(\cdot; \mu_k, \sigma \cdot \nu_k^{-1/2})$ denotes a multivariate normal density with mean $\mu_k \in \mathbb{R}^d$ and a diagonal covariance matrix with the squared elements of vector $\sigma \cdot \nu_k^{-1/2} = (\sigma_1 \nu_{k1}^{-1/2}, \ldots, \sigma_d \nu_{kd}^{-1/2})$ on the diagonal.

We use the following prior for (θ, m). The prior for $(\alpha_1, \ldots, \alpha_m)$ conditional on m is Dirichlet$(a/m, \ldots, a/m)$, $a > 0$,

$$\Pi(\alpha_1, \ldots, \alpha_m|m) = \frac{\Gamma(a)}{\Gamma(a/m)^m} \prod_{j=1}^{m} \alpha_j^{a/m-1}.$$

It is a standard conjugate prior for discrete probability distributions and it is commonly used in finite mixture models. The prior means and variances of the mixing weights are equal to $1/m$ and $(m-1)/([a+1]m^2)$ correspondingly. The hyperparameter a is called the concentration parameter: when a is large, the prior concentrates on equal mixing weights; when a is small, a considerable fraction of mixing weights tend to be close to 0 a priori. For applications of Dirichlet priors in econometrics, see, for example, Chamberlain and Imbens (2003). The prior probability mass function for the number of mixture components m is

$$\Pi(m) \propto e^{-\gamma m (\log m)^{\tau_1}}, \; m = 1, 2, \ldots, \quad \gamma > 0, \; \tau_1 \geq 0, \tag{3}$$

where $\propto$ means “proportional to”. The exponential tails of $\Pi(m)$ attain a tradeoff between putting just enough prior probability on the relevant finite mixture approximations of f_0 and putting appropriately small prior probabilities on rough mixtures that would overfit the data.

A popular alternative to specifying a prior on m and $(\alpha_1, \ldots, \alpha_m)$ is a Dirichlet process mixture (m is set to infinity and a "stick-breaking" prior (Sethuraman (1994)) is used for the infinite sequence of mixing weights $(\alpha_1, \ldots)$). This prior would deliver the same posterior contraction rates for continuous variables or settings where smoothing is important; however, when smoothing is not beneficial, the Dirichlet process mixture prior does not seem to put sufficient weight on the relevant finite mixture approximations, and, hence, we focus on the mixtures of finite mixtures here.

The component specific scale parameters ν_k are not necessary for asymptotic results; it is a common practice in the literature to include them (see, for example, Geweke (2005)) and they seem to improve the finite sample performance. We use independent conditionally conjugate gamma-normal priors for (μ_{kj}, ν_{kj}). The common scale parameters σ are required to ensure that the prior puts sufficient probability on small values of the variances of all mixture components at once (the variances play a role of the bandwidth in asymptotic results). We use independent inverse Gamma priors for the components of σ. A detailed description of the model, priors, and the MCMC algorithm for model estimation is given in Appendix A. Section 4.3.1 provides more general conditions on the prior that deliver adaptive posterior contraction rates for the model in (2).

The Bayesian model, the MCMC estimation algorithm, and the theoretical results presented below can be easily modified to accommodate settings with variables that take both discrete and continuous values. A standard example of such variables is the consumer expenditure on a good that can be zero with positive probability and otherwise is continuous on $\mathbb{R}_+$. To accommodate this example, we can associate the discrete value of 0 with interval $A_0 = \mathbb{R}_-$ for a latent variable $\tilde{y}$ and treat the continuous positive expenditure values as x in model (2). We do not pursue such modifications here for brevity.

An important issue in kernel smoothing estimation of densities with bounded support is the estimator bias near the boundary. It is not known if a similar problem arises in Bayesian normal mixture models as the locations of the normal distributions in the mixture models are chosen effectively by the penalized likelihood maximization rather than set equal to the observations as in kernel smoothing. Nevertheless, the normal densities have unbounded support and normal mixture models appear to perform better when continuous variables with known bounds are appropriately transformed into unbounded variables, which is also a common remedy in the literature on kernel smoothing.

3. APPLICATION

In applied economics literature, nonparametric estimation of multivariate discrete or mixed discrete-continuous distributions is often used in the first stage of two stage estimation procedures for structural discrete choice models. Pakes, Ostrovsky, and Berry (2007) compare various two stage estimation procedures on a model of firm's entry decisions. Their Monte Carlo experiments provide convenient and realistic settings for demonstrating the performance of the mixture based models in practice.

The first stage in Pakes et al. (2007) requires estimation of entry and exit probabilities conditional on the number of entrants currently in the market and a discretized market size measure. These conditional probabilities are essentially obtained from the standard frequency estimator of the joint distribution for the four-dimensional vector of discrete random variables: the market size, the number of firms currently in the market, the number of new entrants, and the number of exiting firms. In what follows, we use the simulated data from Pakes et al. (2007) to compare our estimator with the standard frequency estimator and a classical kernel estimator with special discrete kernels from a publicly available R package *np* (Hayfield and Racine (2008)). The kernel bandwidth parameters are selected in the package by cross-validation as described in Li and Racine (2003); the latter authors provide simulation evidence that their methods outperform several other alternatives in the classical literature; the package *np* implements a wide variety of nonparametric methods presented in a textbook on nonparametric econometrics by Li and Racine (2007).

Pakes et al. (2007) simulate a structural entry exit model to obtain one million draws for their Monte Carlo experiments. We use this one million simulated draws as a population distribution to estimate. The support of this population distribution consists of 2617 values of the four dimensional random vectors describing the market size, the current number of firms, the number of entrants and the number of exits. The marginal population distributions of each vector component are depicted in Figure 1. All the discrete values shown in the figure have non-zero probabilities, although some of those probabilities are small. From this population, we draw 50 random samples of size $n = 500$ (Pakes et al. (2007) use $n = 250$ and $n = 1000$ in their Monte Carlo experiments). For each sample, we compute the standard frequency estimator, the kernel estimator, and the mixture model estimators for a fixed $m \in \{1, \ldots, 30\}$ and a variable m. The MCMC algorithm for the fixed m model is standard in the literature (Diebolt and Robert (1994)). For the variable m model, we implemented two MCMC algorithms: an adaptation of a split-merge algorithm for Dirichlet process mixtures from Jain and Neal (2004) and an approximately optimal reversible jump algorithm from Norets (2020); they produce the same estimation results in the Monte Carlo experiments but the latter algorithm converges much faster. The reversible jump algorithm is described in detail in Appendix A.

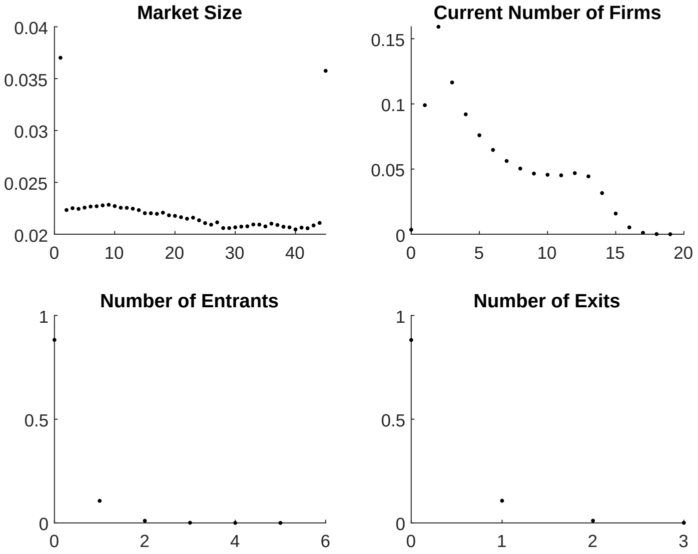

FIGURE 1.— Marginal population distributions

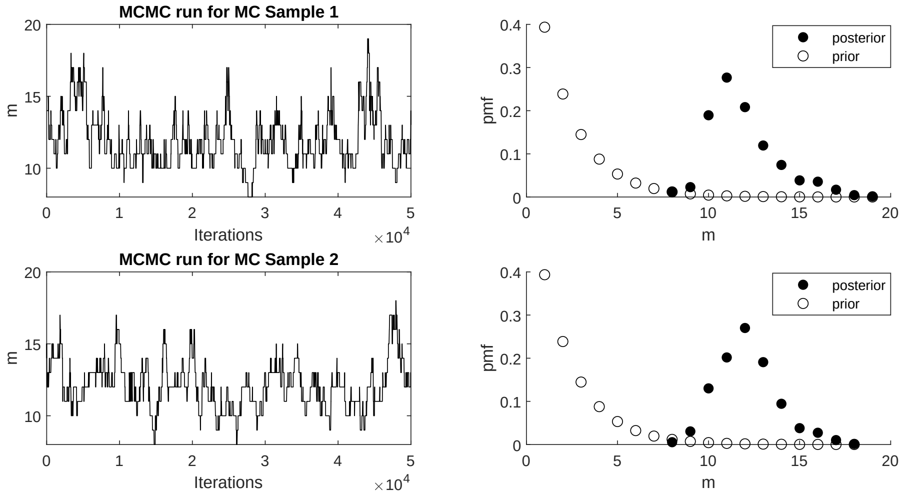

FIGURE 2.— MCMC trace plot and prior and posterior of m for two samples

Figure 2 presents the reversible jump MCMC draws and the prior and the posterior distributions of m for the first two samples used in the Monte Carlo experiment. Estimation results for the fixed and variable m models are obtained from 10,000 and 50,000 MCMC draws correspondingly, as MCMC convergence is slower for the variable m models. As can be seen from the MCMC trace plots in the figure, the posterior simulator reliably explores the posterior distribution; MCMC results for other samples are similar.

The priors used in estimation experiments are roughly based on the first two sample moments: the prior for the location parameter μ_{kj} is centered at the corresponding sample average, $\bar{Y}_j = \sum_{i=1}^n Y_{ij}/n$ and has variance equal to the sample variance, $\hat{\sigma}_j^2 = \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2/n$. The prior mode of the precision parameter σ_j^{-2} is set to the inverse of the sample variance, $\hat{\sigma}_j^{-2}$ and its variance is set to 1. The component specific scale parameters have prior mode and precision equal to 1. These empirical Bayes priors are similar to unit variance priors centered at 0 for location parameters and 1 for scale parameters used in conjunction with standardized data.

Choosing reasonable values for the Dirichlet parameter a and the prior hyper-parameters for m is less straightforward. We set $\tau_1 = 0$ as the finer adjustments that it can provide to the penalization of larger values of m in the prior do not appear to be important in simulations. We set $\gamma = 0.5$, approximately the smallest value at which every mixture component has at least several observations assigned to it by latent mixture allocation variables (defined in Appendix A) on most iterations of MCMC sampler runs. The Dirichlet parameter $a = 15$ is set to be comparable to the values of MCMC draws of m (larger values of a shrink towards equal mixing probabilities). Prior robustness and sensitivity checks are important, especially for these hyper-parameters. The estimation results are not sensitive to moderate variations in the prior ($a \in \{10, 15, 20\}$ and $\gamma \in \{0.25, 0.5, 1\}$), as we illustrate in the online supplement.

The estimation errors in L_1, L_2, and L_∞ averaged over the 50 random samples are presented in Figure 3. The L_r distance between discrete-continuous distributions p_1 and p_2 can be defined by

$$d_{L_r}(p_1, p_2) = \left(\sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} |p_1(y,x) - p_2(y,x)|^r dx \right)^{1/r}, \ r > 0.$$

The L_1 distance is also equal to two times the largest difference between the probabilities that the two distributions can assign to the same event; in the case of only discrete variables, L_∞ is the sup-norm; L_2 is most commonly used in classical nonparametrics for analytical tractability.

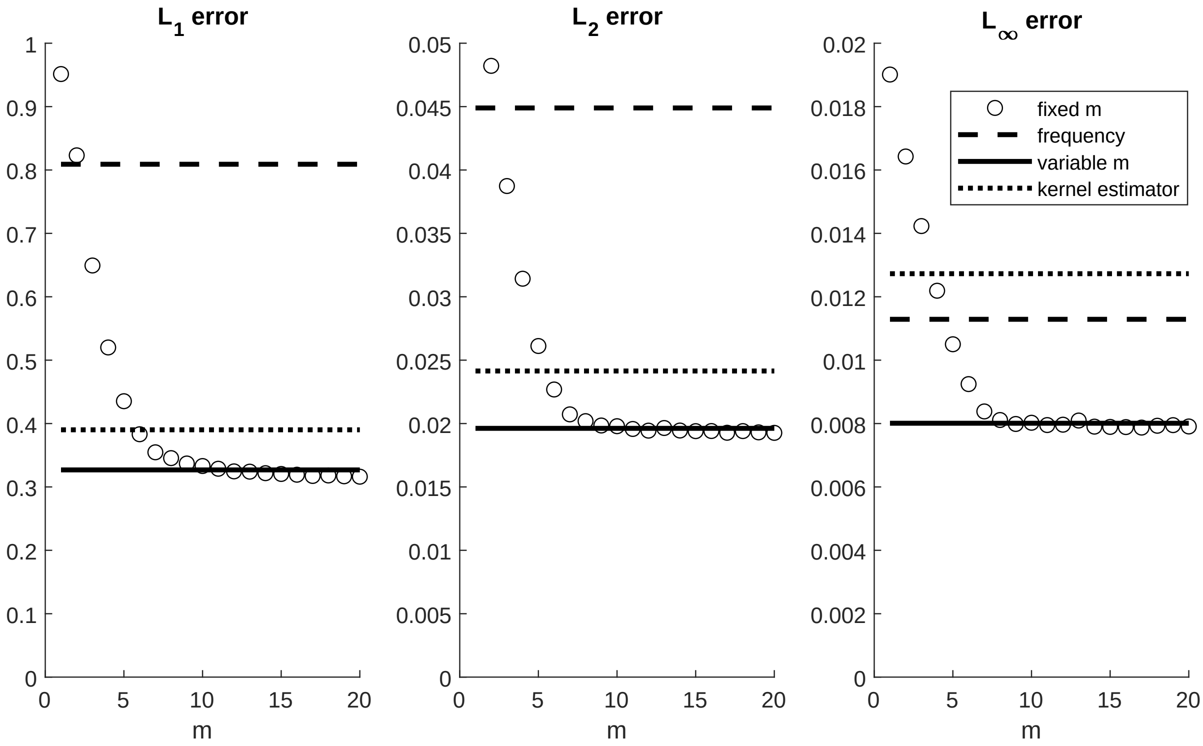

FIGURE 3.— Average Estimation Errors for Bayesian Fixed and Variable m Estimators and Frequency and Kernel Estimators

As can be seen from the figure, the mixture based estimators match the average L_1 error of the frequency estimator with just two mixture components and that of the kernel estimator with six mixture components. The results for L_2 and L_∞ are similar, except the kernel estimator performs slightly worse than the frequency estimator in the sup-norm. The use of a higher number of mixture components and a variable number of components further reduces the estimation error of the Bayesian estimators. The improvements of the mixture model over the standard frequency estimator are expected given the smooth appearance of the probability mass functions in Figure 1, the sample size ($n = 500$), and the cardinality of the population support, 2617. The mixture models outperform the kernel estimator on average as shown in the figure and in each of the 50 random samples. Theoretical properties (beyond the consistency and the asymptotic normality for a fixed discrete support) are not known for the discrete kernel estimator in our asymptotic settings with smoothness and a possibly growing support. Our conjecture is that at least without considerable modifications this kernel estimator is unlikely to deliver the adaptive optimal estimation rates that are established for mixture models in the following section; and, perhaps, that is why the kernel estimator is outperformed by the mixture model in our simulations. A few other applications and favorable comparisons of a fixed m mixture model with standard parametric and nonparametric alternatives can be found in Norets and Pelenis (2012).

The performance of the variable m model is practically the same as the performance of models with a large fixed m. Somewhat unexpectedly, the estimation results for the models with fixed m do not deteriorate when m is large ($m = 30$). The estimation errors are slightly more volatile for larger m, but on average, the errors decrease in m as can be seen in Figure 3. Of course, the performance can be easily evaluated in simulation settings, when the data generating process is known. As far as we are aware, theoretically justified Bayesian procedures for choosing a fixed m have not been developed in nonparametric settings and their development is an interesting subject for future research. Hence, presently the variable m model with the asymptotic guarantees obtained in this paper is the preferred option, and the fixed m models should be used for sensitivity and robustness checks.

Overall, the Monte Carlo simulations presented in this section suggest that models for discrete data based on mixtures and latent variables should be an important part of the toolkit in empirical industrial organization and economics more generally. The following section presents asymptotic results that further justify this claim from the theoretical perspective.

4. ASYMPTOTIC FRAMEWORK AND RESULTS

To get more refined results and to accommodate discrete variables that are not ordered or "smooth", we allow N_j's to grow at different rates for different j's or to be constant for some j's. For the same reason, we allow for anisotropic smoothness of the density f_0 that accommodates the existence of derivatives of different orders along different coordinates.

4.1. *Anisotropic Smoothness*

For each coordinate $j \in \{1, \ldots, d\}$, we introduce a smoothness coefficient, $\beta_j > 0$, such that $\lfloor \beta_j \rfloor$ (the largest integer that is strictly smaller than β_j) is the highest possible order of the partial derivative with respect to the coordinate j. In the univariate case, $\lfloor \beta_j \rfloor$'th derivative is often assumed to satisfy a Holder condition with the exponent $\beta_j - \lfloor \beta_j \rfloor$ to accommodate noninteger smoothness coefficients and to deliver Taylor expansion approximations with remainders of the appropriate order. Different generalizations of these ideas to the multivariate case are possible. We introduce a generalization below that is suitable for our purposes. Let $\mathbb{Z}_+$ denote the set of non-negative integers. For smoothness coefficients $(\beta_1, \ldots, \beta_d)$ and an envelope constant L, an anisotropic $(\beta_1, \ldots, \beta_d)$-Holder class, $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, is defined as follows.

DEFINITION 1 *$f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ if for any $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, $\sum_{l=1}^d k_l/\beta_l < 1$, mixed partial derivative of order k, $D^k f$, is finite and*

$$|D^k f(z + \Delta z) - D^k f(z)| \leq L \sum_{j=1}^{d} |\Delta z_j|^{\beta_j(1-\sum_{l=1}^d k_l/\beta_l)}, \tag{4}$$

for any Δz such that $\Delta z_j = 0$ when $\sum_{l=1}^d k_l/\beta_l + 1/\beta_j < 1$.

In this definition, a Holder condition is imposed on $D^k f$ for a coordinate j when $D^k f$ cannot be differentiated with respect to z_j anymore ($\sum_{l=1}^d k_l/\beta_l < 1$ but $\sum_{l=1}^d k_l/\beta_l + 1/\beta_j \geq 1$). This definition slightly differs from definitions available in the literature on anisotropic smoothness that we found. Section 13.2 in Schumaker (2007) presents some very general anisotropic smoothness definitions but restricts attention to integer smoothness coefficients. Ibragimov and Hasminskii (1984), and most of the literature on minimax rates under anisotropic smoothness that followed including Barron et al. (1999) and Bhattacharya et al. (2014), do not restrict mixed derivatives. Shen et al. (2013) use $|\Delta z_j|^{\min(\beta_j - k_j, 1)}$ instead of $|\Delta z_j|^{\beta_j(1-\sum_l k_l/\beta_l)}$ in (4). Their requirement is stronger than ours for functions with bounded support, and it appears too strong for our derivation of lower bounds on the estimation rate. However, our definition is sufficiently strong to obtain a Taylor expansion with remainder terms that have the same order as those in Shen et al. (2013) (while the definitions that do not restrict mixed derivatives do not deliver such an expansion).

When $\beta_j = \beta$, $\forall j$ and $\sum_{l=1}^d k_l/\beta + 1/\beta \geq 1$, $\beta_j(1 - \sum_{l=1}^d k_l/\beta_l) = \beta - \lfloor \beta \rfloor$, and we get the standard definition of β-Holder smoothness for the isotropic case.

The envelope L can be assumed to be a function of $(z, \Delta z)$ to accommodate densities with unbounded support. We derive lower bounds on estimation rates for a constant envelope function; the derived bounds are applicable to functions with non-constant envelops as a constant envelop is just a special case of a non-constant one. Upper bounds on posterior contraction rates are derived under more general assumptions on L.

4.2. *Lower Bounds on Estimation Rates*

For a class of probability distributions $\mathcal{P}$, ζ is said to be a lower bound on the estimation error in metric ρ if there exists a positive constant c independent of n such that

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}} P\left(\rho(\hat{p},p) \geq \zeta\right) \geq c > 0.$$

This definition means that there does not exist an estimator that asymptotically delivers an estimation error in ρ that is smaller than ζ for all data generating distributions in $\mathcal{P}$. If the estimation error for a given estimator for distributions in $\mathcal{P}$ matches (up to a multiplicative constant) a lower bound for $\mathcal{P}$, then this estimator is considered rate optimal. A comprehensive introduction into the theory of lower bounds can be found in Tsybakov (2008). In this section, we present lower bounds for discrete continuous distributions that are matched with upper bounds on estimation errors for the mixture based models in Section 4.3.

We consider the following class of probability distributions: for a positive constant L, let

$$\mathcal{P} = \left\{ p:\ p(y,x) = \int_{A_y} f(\tilde{y},x)d\tilde{y},\ f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L},\ f \text{ is a pdf} \right\}. \tag{5}$$

To define our lower bounds we need the following additional notation. Let $\mathcal{A}$ denote a collection of all subsets of indices for discrete coordinates $\{1,\ldots,d_y\}$. For $J \in \mathcal{A}$, let $J^c = \{1,\ldots,d\} \setminus J$ and y_J denotes the sub-vector $\{y_j,\ j\in J\}$ for a vector y. Then,

$$N_J = \prod_{j\in J} N_j$$

denotes the number of values a discrete subvector y_J can take, $d_J = \mathrm{card}(J)$, and

$$\beta_{J^c} = \left[\sum_{j\in J^c} \beta_j^{-1}\right]^{-1}$$

denotes an aggregate smoothness coefficient for the subvector containing the coordinates of the continuous part of observations x and the continuous latent variables $\tilde{y}$ with indices in J^c. For $J=\emptyset$ or $J^c=\emptyset$, we set $N_\emptyset = 1$, $\beta_\emptyset=\infty$, and $\beta_\emptyset/(2\beta_\emptyset+1)=1/2$.

THEOREM 1 *For $\mathcal{P}$ defined in (5),*

$$\Gamma_n = \min_{J\in\mathcal{A}} \left[\frac{N_J}{n}\right]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}} = \left[\frac{N_{J_*}}{n}\right]^{\frac{\beta_{J_*^c}}{2\beta_{J_*^c}+1}} \tag{6}$$

multiplied by a positive constant is a lower bound on the estimation error in the L_1 distance.

One could recognize expression $[N_J/n]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$ in (6) as the standard estimation rate for a d_{J^c}-dimensional density with anisotropic smoothness coefficients $\{\beta_j,\, j \in J^c\}$ and the sample size n/N_J (Ibragimov and Hasminskii (1984)). One way to interpret this is that the density of $\{x, \tilde{y}_j,\, j \in J^c\}$ conditional on y_J is $\{\beta_j,\, j \in J^c\}$-smooth and the number of observations available for its estimation (observations with the same value of y_J) should be of the order n/N_J; also, the estimation rate for the marginal probability mass function for y_J is $[N_J/n]^{1/2}$, which is at least as fast as $[N_J/n]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$. In this interpretation, smoothing is not performed over the discrete coordinates with indices in set J, and the lower bound is obtained when J minimizes $[N_J/n]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$. Thus, an estimator that delivers the rate in (6) should, in a sense, optimally choose the subset of discrete variables over which to perform smoothing. In the standard asymptotic settings, when the support of the discrete variables stays constant and the smoothness coefficients for all the continuous variables are the same, $\beta_{d_y+1} = \ldots = \beta_d$, the lower bound on the estimation rate in Theorem 1 simplifies to the familiar expression, $n^{-\beta_d/(2\beta_d+d_x)}$, which explicitly showcases the curse of dimensionality inherent in nonparametric estimation.

It should be possible to extend the results on the lower bounds to other distances. However, suitable sufficient conditions in the Bayesian nonparametrics literature for the corresponding upper bounds appear to be currently available only for the L_1 distance (or the Hellinger and the total variation distances, which are equivalent); hence, we focus on L_1 here. The proof of Theorem 1 is given in Appendix B.

4.2.1. *Related Literature on Lower Bounds*

Let us briefly review most relevant results on lower bounds and place our results in that context. The most closely related results on minimax rates for anisotropic continuous distributions are developed in Ibragimov and Hasminskii (1984). The minimax estimation rates for mixed discrete continuous distributions appear to be studied first by Efromovich (2011). He considers discrete variables with a fixed support and no smoothness assumptions on the discrete part of the distribution. He shows that in these settings the optimal rates for discrete continuous distributions are equal to the optimal nonparametric rates for the continuous part of the distribution. Relaxing the assumption of the fixed support for the discrete part of the distribution is very desirable in nonparametric settings. It has been commonly observed at least since Aitchison and Aitken (1976) that smoothing discrete data in nonparametric estimation improves results in practice. Hall and Titterington (1987) introduced an asymptotic framework that provided a precise theoretical justification for improvements resulting from smoothing in the context of estimating a univariate discrete distribution with a support that can grow with the sample size. In their setup, the support is an ordered set and the probability mass function is β-smooth (in a sense that analogs of β-order Taylor expansions hold). They show that in their setup the minimax rate is the smaller one of the following two: (i) the optimal estimation rate for a continuous density with the smoothness level β, $n^{-\beta/(2\beta+1)}$, and (ii) the rate of convergence of the standard frequency estimator, $(N/n)^{1/2}$, where N is the cardinality of the support and n is the sample size. Hall and Titterington (1987) refer to their setup as "Sparse Multinomial Data" since N can be larger than n and this is the reason we refer to sparsity in the title of the paper. Burman (1987) established similar results for $\beta = 2$. Subsequent literature in multivariate settings (e.g., Dong and Simonoff (1995), Aerts et al. (1997)) did not consider lower bounds but demonstrated that when the support of the discrete distribution grows sufficiently fast then estimators that employ smoothing can achieve the standard nonparametric rates for β-smooth densities on $\mathbb{R}^d$, $n^{-\beta/(2\beta+d)}$.

We generalize the results of Hall and Titterington (1987) on lower bounds for univariate discrete distributions to multivariate mixed discrete-continuous case and anisotropic smoothness. Alternatively, our results can be viewed as a generalization of results in Efromovich (2011) to settings with anisotropic smoothness and potentially growing supports for discrete variables.

4.3. *Posterior Contraction Rates for a Mixture of Normals Model*

4.3.1. *Assumptions on Prior*

The assumptions on the prior for model (2) in Section 2.1 can be slightly generalized as follows. For positive constants $a_1, a_2, \ldots, a_9$, for each $j \in \{1, \ldots, d\}$, σ_j is assumed independent of other parameters a priori and the prior satisfies

$$\Pi(\sigma_j^{-2} \geq s) \leq a_1 e^{-a_2 s^{a_3}} \text{ for all sufficiently large } s > 0 \tag{7}$$

$$\Pi(\sigma_j^{-2} < s) \leq a_4 s^{a_5} \text{ for all sufficiently small } s > 0 \tag{8}$$

$$\Pi\{s < \sigma_j^{-2} < s(1+t)\} \geq a_6 s^{a_7} t^{a_8} e^{-a_9 s^{1/2}}, \; s > 0, \; t \in (0,1). \tag{9}$$

The inverse Gamma prior for σ_i satisfies (7)-(9).

A priori, the components of μ_k, μ_{kj}, $k = 1, \ldots, m$, $j = 1, \ldots, d$ are assumed independent from each other, other parameters, and across k. Prior density for μ_{kj} is bounded below for some $a_{11}, a_{12}, \tau_2 > 0$ by

$$(10) \qquad a_{11} \exp(-a_{12}|\mu_{kj}|^{\tau_2}),$$

and for some $a_{13}, \tau_3 > 0$ and all sufficiently large $\mu_{kj} > 0$,

$$(11) \qquad \Pi(\mu_{kj} \notin [-\mu, \mu]) \leq e^{-a_{13}\mu^{\tau_3}}.$$

Normal priors for μ_{kj} satisfy these conditions.

A prior on m that can be bounded above and below by functions in the form of the right hand side of (3), possibly with different constants, would work; to simplify the notation we assume (3). We also set the component specific scale parameters ν_{ji} to 1. An extension of the posterior contraction results to variable ν_{kj}'s is straightforward, see, for example, Theorem A.5 in Norets and Pati (2017) for continuous variables, and it is not presented here for brevity.

4.3.2. *Posterior Contraction Rates*

This section presents upper bounds on the posterior contraction rates for the Bayesian mixture model that match the lower bounds in Section 4.2 up to a log factor. That means that the Bayesian mixture model deliver a rate optimal (up to a log) estimator for the data generating process in (1) under our smoothness assumptions. The estimator is adaptive since the prior and model specification do not depend on the smoothness of the data generating density and the fineness of the support relative to the sample size. To simplify the exposition we present the results below in Theorem 2 for the case when the data generating latent density f_0 has a bounded support.

THEOREM 2 *Assume the conditions on the prior in Sections 4.3.1. Suppose $f_0 \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ and $\overline{f} \geq f_0 \geq \underline{f} > 0$ holds on the support of f_0, where L, $\overline{f}$, and $\underline{f}$ are finite positive constants. Let*

$$(12) \qquad \epsilon_n = \min_{J \in \mathcal{A}} \left(\left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J} \right)$$

where

$$t_J > \left(d_{J^c} + \beta_{J^c}^{-1} + \max\{\tau_1, 1\}\right) / \left(2 + \beta_{J^c}^{-1}\right) + \max\{0, (1-\tau_1)/2\}$$

and τ_1 is a parameter in the prior on m. Suppose also $n\epsilon_n^2 \to \infty$ and for J_* that attains the minimum in (12), $N_{J_*} = o(n^{1-\nu})$ for some small $\nu > 0$. Then, the posterior contracts *at the rate ϵ_n: there exists $\bar{M} > 0$ such that*

$$\Pi\left(p : d_{L_1}(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

As in Section 4.2, when $J^c = \emptyset$, β_{J^c} can be defined to be infinity and $\beta_{J^c}/(2\beta_{J^c}+1) = 1/2$ in (12). The assumption $N_{J_*} = o(n^{1-\nu})$ excludes the cases with very slow (non-polynomial) rates as some parts of the proof require $\log(1/\epsilon_n)$ to be of order $\log n$.

The theorem is a special case of the results presented in Appendix C that can accommodate unbounded support for f_0. The proof of Theorem 2 follows from the discussion of the more general assumptions in the appendix as the bounded support case is used there to illustrate the assumptions. Similarly to other papers on posterior contraction for mixtures of normal densities though, the more general sufficient conditions in the appendix require subexponential tails for f_0. The results for f_0 with an unbounded support also require the envelope function L in the smoothness definition to be comparable to f_0.

The proof of the posterior contraction results is based on the general sufficient conditions from Ghosal et al. (2000). It exploits approximations of smooth densities by mixtures of normal distributions developed in the Bayesian nonparametrics literature (Rousseau (2010), Kruijer et al. (2010), de Jonge and van Zanten (2010), and Shen, Tokdar, and Ghosal (2013)) and also develops appropriate approximations for nonsmooth discrete distributions. Posterior contraction rates for nonparametric density estimation by mixture models derived in the aforementioned papers also include a log factor similar to $(\log n)^{t_J}$ in (12). It is not known in the literature whether the log factor can be avoided; however, it is not a very important issue as the log factor is negligible compared to the polynomial part of the rate.

The results on the upper bounds in this section and lower bounds in Section 4.2 also hold for the data generating processes where f_0 is not smooth at all in some discrete coordinates. The resulting rates can be obtained from those we derive by setting the corresponding coordinates in β to (values arbitrarily close to) zero in (6), so that for the optimal rate, smoothing is effectively not performed for these coordinates. Thus, the proposed Bayesian model achieves the objective outlined in the introduction: it optimally takes advantage of smoothness in the data generating process if it is present and at the same time performs no worse than the standard frequency estimators if the data generating process is not (sufficiently) smooth. Simulations in Section 3 suggest that the model performs better in practice than available parametric and nonparametric alternatives and appears to live up to its excellent theoretical properties.

5. FUTURE WORK

In many applications, conditional rather than joint distributions are actually of interest. Of course, one could always estimate the joint distribution and then extract the conditional distributions of interest. When the smoothness of the joint and conditional distributions is the same then rate optimality of joint distribution estimator implies rate optimality for the corresponding conditional distribution estimator. However, when the conditional distribution is smoother then it could be beneficial to estimate the conditional distribution directly. In an ongoing work, Norets and Pelenis (2021), we pursue an extension of our posterior contraction results to conditional distribution models based on covariate dependent mixtures; the extension is similar to work by Norets and Pati (2017) on continuous distributions.

It would also be of interest to explore whether other Bayesian nonparametric models (for example, those based on Gaussian process priors) or classical nonparametric methods based on higher order kernels or orthogonal series expansions can deliver estimators with adaptive optimal convergence rates in our asymptotic framework.

REFERENCES

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.

Albert, J. H. and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Barron, A., L. Birgé, and P. Massart (1999): "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, 113, 301–413.

Bhattacharya, A., D. Pati, and D. Dunson (2014): "Anisotropic function estimation using multi-bandwidth Gaussian processes," *The Annals of Statistics*, 42, 352–381.

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

Canale, A. and D. B. Dunson (2011): "Bayesian Kernel Mixtures for Counts," *Journal of the American Statistical Association*, 106, 1528–1539.

——— (2015): "Bayesian multivariate mixed-scale density estimation," *Statistics and its Interface*, 8, 195–201.

Chamberlain, G. and G. W. Imbens (2003): "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*, 21, 12–18.

de Jonge, R. and J. H. van Zanten (2010): "Adaptive nonparametric Bayesian inference using location-scale mixture priors," *The Annals of Statistics*, 38, 3300–3320.

DEY, D., P. MULLER, AND D. SINHA, eds. (1998): *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics , Vol. 133, Springer.

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

DIEBOLT, J. AND C. P. ROBERT (1994): "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 363–375.

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

EFROMOVICH, S. (2011): "Nonparametric estimation of the anisotropic probability density of mixed variables," *Journal of Multivariate Analysis*, 102, 468 – 481.

GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

GHOSAL, S., J. K. GHOSH, AND A. W. V. D. VAART (2000): "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531.

GHOSAL, S. AND A. VAN DER VAART (2007): "Posterior convergence rates of Dirichlet mixtures at smooth densities," *The Annals of Statistics*, 35, 697–723.

GHOSAL, S. AND A. W. VAN DER VAART (2001): "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *The Annals of Statistics*, 29, 1233–1263.

GREEN, P. J. (1995): "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

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

HAYFIELD, T. AND J. S. RACINE (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

HOTZ, J. AND R. MILLER (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–530.

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

JAIN, S. AND R. M. NEAL (2004): "A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158–182.

KRUIJER, W., J. ROUSSEAU, AND A. VAN DER VAART (2010): "Adaptive Bayesian density estimation with location-scale mixtures," *Electronic Journal of Statistics*, 4, 1225–1257.

LI, Q. AND J. RACINE (2003): "Nonparametric estimation of distributions with categorical and continuous data," *Journal of Multivariate Analysis*, 86, 266–292.

LI, Q. AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

NORETS, A. (2020): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models," *Econometric Theory*, Forthcoming.

NORETS, A. AND D. PATI (2017): "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 980–1012.

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

——— (2014): "Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures," *Econometric Theory*, 30, 606–646.

——— (2021): "Adaptive Bayesian Estimation of Conditional Discrete-Continuous Distributions with an Application to Stock Market Trading Activity," *Journal of Econometrics*, forthcoming.

Pakes, A., M. Ostrovsky, and S. Berry (2007): "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)," *the RAND Journal of Economics*, 38, 373–399.

Rousseau, J. (2010): "Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density," *The Annals of Statistics*, 38, 146–180.

Schumaker, L. (2007): *Spline functions : basic theory*, Cambridge New York: Cambridge University Press.

Sethuraman, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

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.

APPENDIX A: MODEL, PRIORS, AND MCMC ALGORITHM

A.1. *Model and Priors*

For the MCMC implementation and description, it is convenient to formulate the model in (2) using mixture allocation latent variables (Diebolt and Robert (1994)), $(s_1, \ldots, s_n)$, latent variables $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$ corresponding to discrete observations, and precision parameters $h_j = \sigma_j^{-2}$ so that for each observation index $i \in \{1, \ldots, n\}$ and mixture component index $k \in \{1, \ldots, m\}$,

$$(\tilde{Y}_i, X_i)|s_i = k, \mu_k, h, \nu_j, m \sim \phi\left(\cdot; \mu_k, (h_1^{-1/2}\nu_{k1}^{-1/2}, \ldots, h_d^{-1/2}\nu_{kd}^{-1/2})\right),$$

$$p(s_i = k|\theta, m) = \alpha_k.$$

The joint distribution of observables and unobservables in the model is

$$\begin{aligned} &p\left(Y_i, \tilde{Y}_i, X_i, s_i, \; i = 1, \ldots, n; \mu_1, \nu_1, \ldots, \mu_m, \nu_m; h, m\right) = \\ &\prod_{i=1}^{n} 1\{\tilde{Y}_i \in A_{Y_i}\}\phi\left(\tilde{Y}_i, X_i; \mu_{s_i}, (h_1^{-1/2}\nu_{s_i 1}^{-1/2}, \ldots, h_d^{-1/2}\nu_{s_i d}^{-1/2})\right) \; \alpha_{s_i} \\ &\cdot \Pi(\alpha_1, \ldots, \alpha_m|m) \cdot \prod_{j=1}^{d} \Pi(h_j) \prod_{k=1}^{m} \Pi(\mu_{kj}|\nu_{kj})\Pi(\nu_{kj}) \cdot \Pi(m). \end{aligned} \tag{13}$$

The common precision parameter, h_j, is a priori distributed as a square of a gamma distributed random variable with shape $\underline{A}_{h_j}$ and rate $\underline{B}_{h_j}$, which is consistent with the conditions in Section 4.3.1:

$$\Pi(h_j) \propto h_j^{\underline{A}_{h_j}/2-1} e^{-\underline{B}_{h_j} \cdot h_j^{1/2}}.$$

The priors for (ν_{kj}, μ_{kj}) are conditionally conjugate gamma-normal:

$$\Pi(\nu_{kj}) \propto \nu_{kj}^{\underline{A}_{\nu_j}-1} e^{-\underline{B}_{\nu_j} \cdot \nu_{kj}},$$

$$\Pi(\mu_{kj}|\nu_{kj}) \propto \nu_{kj}^{1/2} e^{-0.5 \underline{h}_{\mu_j} \nu_{kj} (\mu_{kj} - \underline{\mu}_j)^2}.$$

The priors for mixing weights and m are as described in Section 2.1:

$$\Pi(\alpha_1, \ldots, \alpha_m | m) \propto \prod_{k=1}^{m} \alpha_k^{a/m-1},$$

$$\Pi(m) \propto e^{-\gamma m (\log m)^{\tau_1}}.$$

A.2. *MCMC Algorithm*

We develop a Metropolis-within-Gibbs algorithm with a reversible jump step for m (Green (1995)) for exploring the posterior distribution. See, for example, Geweke (2005) for a textbook treatment of MCMC algorithms in general and for mixture models in particular.

Conditional on m, the distributions for the Gibbs sampler blocks of the parameters and the latent variables are proportional to (13) and can be written as follows:

$$\tilde{Y}_{ij} | \ldots \sim \phi\left(\tilde{Y}_{ij}; \mu_{s_i j}, h_j^{-1/2} \nu_{s_i j}^{-1/2}\right) \cdot 1\{\tilde{Y}_{ij} \in A_{Y_{ij}}\} \quad \text{(truncated normal)}$$

$$p(s_i = k | \ldots) \propto \phi\left(\tilde{Y}_i, X_i; \mu_k, (h_1^{-1/2} \nu_{k1}^{-1/2}, \ldots, h_d^{-1/2} \nu_{kd}^{-1/2})\right) \alpha_k \quad \text{(multinomial)}$$

$$p(\alpha_1, \ldots, \alpha_m | \ldots) \propto \prod_{k=1}^{m} \alpha_k^{a/m + \sum_{i=1}^{n} 1\{s_i = k\} - 1} \quad \text{(Dirichlet)}$$

$$p(\mu_{kj}, \nu_{kj} | \ldots) \propto \nu_{kj}^{\bar{A}_{\nu_j} - 1/2} e^{-\bar{B}_{\nu_j} \cdot \nu_{kj} - 0.5 \bar{h}_{\mu_j} \nu_{kj} (\mu_{kj} - \bar{\mu}_j)^2} \quad \text{(gamma-normal)}$$

with parameters

$$\bar{h}_{\mu_j} = \underline{h}_{\mu_j} + h_j \cdot \sum_{i=1}^{n} 1\{s_i = k\}, \quad \bar{\mu}_j = \bar{h}_{\mu_j}^{-1} [\underline{h}_{\mu_j} \underline{\mu}_j + h_j \cdot \sum_{i:\, s_i = k} \tilde{Y}_{ij}],$$

$$\bar{A}_{\nu_j} = \underline{A}_{\nu_j} + 0.5 \sum_{i=1}^{n} 1\{s_i = k\}, \quad \bar{B}_{\nu_j} = \underline{B}_{\nu_j} + 0.5[h_j \sum_{i:\, s_i = k} \tilde{Y}_{ij}^2 + \underline{h}_{\mu_j} \underline{\mu}_j^2 - \bar{h}_{\mu_j} \bar{\mu}_j^2].$$

The block for h_j is simulated by the Metropolis-Hastings-within-Gibbs with a gamma proposal with shape parameter $\underline{A}_{h_j}/2 + n/2$, rate parameter $0.5 \sum_{i=1}^{n} \nu_{s_i j} (\tilde{Y}_{ij} - \mu_{s_i j})^2$ and the Metropolis-Hastings acceptance probability $\min\{1, e^{\underline{B}_{h_j}(h_j^{0.5} - (h_j^*)^{0.5})}\}$, where h_j^* is the proposal and h_j is the current value. In the descriptions of blocks for μ_{kj}, ν_{kj}, and h_j above, it was implicitly assumed that index j refers to discrete coordinates ($j \in \{1, \ldots, d_y\}$); for $j \geq d_y$, $\tilde{Y}_{ij}$ should be replaced by X_{ij} in the descriptions of these blocks.

For the model with variable m, a block for m is added to the MCMC algorithm. The update for m is performed by an approximately optimal reversible jump algorithm from Norets (2020). To apply the algorithm we first transform the mixing weights into unnormalized weights $\tilde{\alpha}_k$, $k = 1, \ldots$, so that conditional on m, $\alpha_k = \tilde{\alpha}_k / \sum_{l=1}^{m} \tilde{\alpha}_l$ and the Dirichlet prior on $(\alpha_1, \ldots, \alpha_m)$ corresponds to a gamma prior for the unnormalized weights: $\tilde{\alpha}_k | m \sim Gamma(a/m, 1)$, $k = 1, \ldots, m$. Let $\theta_k = (\mu_k, \nu_k, \tilde{\alpha}_k)$, $\theta_{1m} = (h, \theta_1, \ldots, \theta_m)$, $Y = \{Y_i, \tilde{Y}_i, X_i\, i = 1, \ldots, n\}$ and denote a proposal distribution for the parameter of a new mixture component $m + 1$ by $\tilde{\pi}_{m+1}(\theta_{m+1} | Y, \theta_{1m})$. The algorithm works as follows. Simulate proposal m^* from $Pr(m^* = m + 1 | m) = Pr(m^* = m - 1 | m) = 1/2$. If $m^* = m + 1$, then also simulate $\theta_{m+1} \sim \tilde{\pi}_{m+1}(\theta_{m+1} | Y, \theta_{1m})$. Accept the proposal with probability $\min\{1, \alpha(m^*, m)\}$, where

$$\begin{aligned} \alpha(m^*, m) = & \frac{p(Y | m^*, \theta_{1m^*}) \Pi(\theta_{1m^*} | m^*) \Pi(m^*)}{p(Y | m, \theta_{1m}) \Pi(\theta_{1m} | m) \Pi(m)} \\ & \cdot \left(\frac{1\{m^* = m + 1\}}{\tilde{\pi}_m(\theta_{m+1} | \theta_{1m}, Y)} + 1\{m^* = m - 1\} \tilde{\pi}_{m-1}(\theta_m | \theta_{1m-1}, Y) \right). \end{aligned} \tag{14}$$

Norets (2020) shows that an optimal choice of proposal $\tilde{\pi}_m$ is the conditional posterior $p(\theta_{m+1} | Y, m + 1, \theta_{1m})$. The conditional posterior can be evaluated up to a normalization constant; however, it seems hard to directly simulate from it and compute the required normalization constant. Hence, we use a Gaussian approximation to $p(\theta_{m+1} | Y, m + 1, \theta_{1m})$ as the proposal (with the mean equal to the conditional posterior mode, obtained by a Newton method, and the variance equal to the inverse of the negative of the Hessian evaluated at the mode).

From an initial value of parameters, $(\theta_{1m}^{(0)}, m^{(0)})$, the MCMC algorithm sequentially updates parameters by simulating from the algorithm blocks. The resulting Markov chain, $(\theta_{1m}^{(r)}, m^{(r)})$, $r = 1, \ldots, M$, is used to approximate posterior objects of interest such as the posterior predictive (or posterior mean) density-point mass

$$p(y, x|Y^n, X^n) \approx \frac{1}{M} \sum_{r=1}^{M} p(y, x|\theta_{1m}^{(r)}, m^{(r)}).$$

APPENDIX B: PROOF OUTLINE FOR LOWER BOUNDS

In this section, we set up the notation and an outline of the proof of Theorem 1. Detailed calculations are delegated to lemmas in the supplement. The proof is based on a general theorem from the literature on lower bounds, which we present next in a slightly simplified form.

LEMMA 1 *(Theorem 2.5 in Tsybakov (2008)) ζ is a lower bound on the estimation error in metric ρ for a class $\mathcal{Q}$ if there exist a positive integer $M \geq 2$ and $q_j, q_i \in \mathcal{Q}$, $0 \leq j < i \leq M$ such that $\rho(q_j, q_i) \geq 2\zeta$, $q_j << q_0$, $j = 1, \dots, M$ and*

$$\sum_{j=1}^{M} KL(Q_j^n, Q_0^n)/M < \log(M)/8, \tag{15}$$

where KL is the Kullback-Leibler divergence and Q_j^n is the distribution of a random sample from q_j.

The following standard result on bounding the number of unequal elements in binary sequences is used in our construction of q_j, $j = 1, \dots, M$.

LEMMA 2 *(Varshamov-Gilbert bound, Lemma 2.9 in Tsybakov (2008)) Consider the set of all binary sequences of length $\bar{m}$,*

$$\Omega = \{w = (w_1, \dots, w_{\bar{m}}) : \ w_r \in \{0, 1\}\} = \{0, 1\}^{\bar{m}}.$$

Suppose $\bar{m} \geq 8$. Then there exists a subset $\{w^1, \dots, w^M\}$ of Ω such that $w^0 = (0, \dots, 0)$,

$$\sum_{r=1}^{\bar{m}} 1\{w_r^j \neq w_r^i\} \geq \bar{m}/8, \ \forall 0 \leq j < i \leq M,$$

and

$$M \geq 2^{\bar{m}/8}.$$

To define q_j's for our problem, we need some additional notation. Let

$$K_0(u) = \exp\{-1/(1-u^2)\} \cdot 1\{|u| \leq 1\}.$$

This function has bounded derivatives of all orders and it smoothly decreases to zero at the boundary of its support. This type of kernel functions is usually used for constructing hypotheses for lower bounds, see Section 2.5 in Tsybakov (2008). Since we need to construct a smooth density that integrates to 1, we define (as illustrated in Figure 4)

$$g(u) = c_0[K_0(4(u+1/4)) - K_0(4(u-1/4))],$$

where $c_0 > 0$ is a sufficiently small constant that will be specified below.

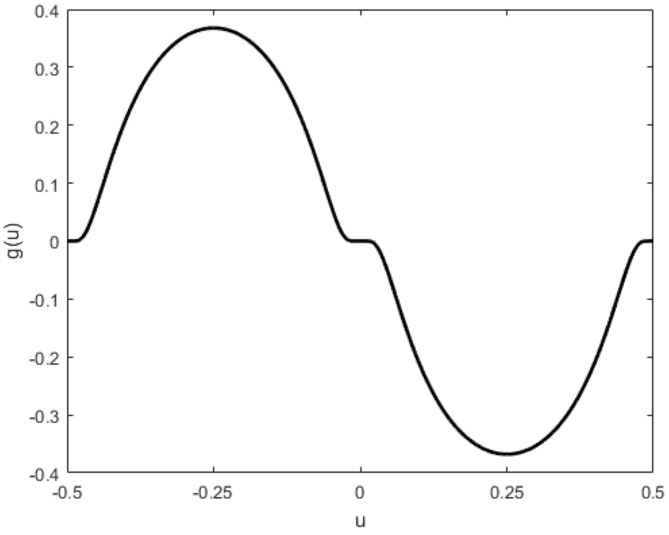

FIGURE 4.— Function g for $c_0 = 1$.

Function g will be used as a kernel in construction of q_k's. Let us define the bandwidth for these kernels first.

For the continuous coordinates, we define the bandwidth as in Ibragimov and Hasminskii (1984),

$$h_i = \Gamma_n^{1/\beta_i},\ i \in \{d_y+1, \dots, d\}.$$

For the discrete ones, over which smoothing is beneficial, we define the bandwidth as

$$h_i = \varrho_i \cdot \Gamma_n^{1/\beta_i} = \frac{2}{N_i} \cdot R_i,\ i \in J_*^c \cap \{1, \dots, d_y\},$$

where $R_i = \lfloor \Gamma_n^{1/\beta_i} N_i/2 \rfloor + 1$ is a positive integer and $\varrho_i \in (1, 2]$ as shown in Lemma 7.

For the rest of the discrete coordinates, our innovation is to first define artificial anisotropic smoothness coefficients $\beta_i^* = -\log(\Gamma_n)/\log N_i$, $i \in J_*$, at which the rate in (6) would have the same value whether we smooth over y_i ($i \in J_*^c$) or not ($i \in J_*$). Then, we define the bandwidth as

$$h_i = 2 \cdot \Gamma_n^{1/\beta_i^*} = 2/N_i, \;\; i \in J_*.$$

To streamline the notation, we also define $\beta_i^* = \beta_i$ for $i \in J_*^c$.

Let m_i be the integer part of h_i^{-1}, $i = 1, \dots, d$. Let us consider $\bar{m} = \prod_{i=1}^d m_i$ adjacent rectangles in $[0,1]^d$, B_r, $r = 1, \dots, \bar{m}$, with the side lengths $(h_1, \dots, h_d)$ and centers $c^r = (c_1^r, \dots, c_d^r)$, $c_i^r = h_i(k_{ir} - 1/2)$, $k_{ir} \in \{1, \dots, m_i\}$. For $z \in \mathbb{R}^d$ and $r = 1, \dots, \bar{m}$, define

$$g_r(z) = \Gamma_n \prod_{i=1}^d g((z_i - c_i^r)/h_i),$$

which can be non-zero only on B_r. A set of hypotheses is defined by sequences of binary weights on g_r's as follows

$$(16) \qquad q_j(y,x) = \int_{A_y} \left[g_0(\tilde{y}, x) + \sum_{r=1}^{\bar{m}} w_r^j g_r(\tilde{y}, x) \right] d\tilde{y},$$

where $w_r^j \in \{0,1\}$, $j = 0, \dots, M$, and M are defined in Lemma 2, and g_0 satisfies the following conditions: (i) it is a density on $\mathbb{R}^d$, (ii) it is bounded away from zero on $[0,1]^d$, (iii) it belongs to $\mathcal{C}^{\beta_1, \dots, \beta_d, L/2}$ for some $L \geq 2$. Examples of g_0 include uniform ($g_0 = 1_{[0,1]^d}$), a normal density, and a smoothed to zero uniform that is proportional to

$$\prod_{i=1}^d \left[1_{[0,1]}(z_i) + IK_0(z_i + 1) \cdot 1(z_i < 0) + IK_0(2 - z_i) \cdot 1(z_i > 1) \right],$$

where $IK_0(z_i) = \int_{-1}^{z_i} K_0(u)du \Big/ \int_{-1}^{1} K_0(u)du$.

The rest of the proof is delegated to lemmas in the supplement, which show that q_k in (16) satisfy the sufficient conditions from Lemma 1. Specifically, Lemma 3 derives the lower bound on the L_1 distance. Lemma 4 verifies condition (15) when $\bar{m} \geq 8$. Lemma 5, part (i) of Lemma 7, and the assumptions on g_0 imply that the latent densities in the definition of q_j belong to $\mathcal{C}^{\beta_1, \dots, \beta_d, L}$, $j = 0, \dots, M$.

This argument (Lemma 4 specifically) requires $\bar{m} \geq 8$ as it relies on Lemma 2. Observe that as $n \to \infty$, $\bar{m} \geq 8$ if there are continuous variables or there are discrete variables over which smoothing is beneficial ($J_*^c \neq \emptyset$). Thus, $\bar{m} < 8$ can happen only if there are no continuous variables and $N_{J_*} = N_1 \cdots N_d$ is bounded. This is just a problem of estimating a multinomial distribution with finite support and the standard results for parametric problems deliver the usual $n^{-1/2}$ rate.

APPENDIX C: POSTERIOR CONTRACTION RATES FOR UNBOUNDED SUPPORT

C.1. *Assumptions on the Data Generating Process for Unbounded Support*

In what follows, we consider a fixed subset of discrete indices $J \in \mathcal{A}$ and show that under regularity conditions, the posterior contraction rate is bounded above by $\left[\frac{N_J}{n}\right]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$ times a log factor. If the regularity conditions we describe below for a fixed J hold for every subset of $\mathcal{A}$, then the posterior contraction rate matches the lower bound in (6) up to a log factor.

Without a loss of generality, let $J = \{1, \ldots, d_J\}$, $I = \{d_J+1, \ldots, d_y\}$, $J^c = \{1, \ldots, d\} \backslash J$, and $d_{J^c} = card(J^c)$. Similarly to $\mathcal{Y}$ and A_y defined in Section 2, we define $\mathcal{Y}_J = \prod_{j\in J} \mathcal{Y}_j$ and $A_{y_J} = \prod_{i\in J} A_{y_i}$. Also, let $y_J = \{y_i\}_{i\in J}$, $\tilde{y}_I = \{\tilde{y}_i\}_{i\in I}$, $\tilde{x} = (\tilde{y}_I, x) \in \tilde{\mathcal{X}} = \mathbb{R}^{d_{J^c}}$.

To formulate the assumptions on the data generating process, we need additional notation,

$$\begin{aligned}
f_{0J}(y_J, \tilde{x}) &= \int_{A_{y_J}} f_0(\tilde{y}_J, \tilde{x}) d\tilde{y}_J, \\
\pi_{0J}(y_J) &= \int_{\tilde{\mathcal{X}}} f_{0J}(y_J, \tilde{x}) d\tilde{x}, \\
f_{0|J}(\tilde{x}|y_J) &= \frac{f_{0J}(y_J, \tilde{x})}{\pi_{0J}(y_J)}, \\
p_{0|J}(y_I, x|y_J) &= \int_{A_{y_I}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I.
\end{aligned}$$

Also, let $F_{0|J}$ and $E_{0|J}$ denote the conditional probability and expectation corresponding to $f_{0|J}$. If $\pi_{0J}(y_J) = 0$ for a particular y_J, then we can define the conditional density $f_{0|J}(\tilde{x}|y_J)$ arbitrarily. We make the following assumptions on the data generating process.

Assumption 1 *There are positive finite constants $b, \bar{f}_0, \tau$ such that for any $y_J \in \mathcal{Y}_J$ and $\tilde{x} \in \tilde{\mathcal{X}}$*

$$f_{0|J}(\tilde{x}|y_J) \leq \bar{f}_0 \exp\left(-b||\tilde{x}||^\tau\right). \tag{17}$$

It appears that all the papers on (near) optimal posterior contraction rates for mixtures of normal densities impose similar tail conditions on the data generating densities.

Assumption 2 *There exists a positive and finite $\bar{y}$ such that for any $(y_I, y_J) \in \mathcal{Y}$ and $x \in \mathcal{X}$*

$$\int_{A_{y_I}\cap\{||\tilde{y}_I||\leq\bar{y}\}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I \geq \int_{A_{y_I}\cap\{||\tilde{y}_I||>\bar{y}\}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I. \tag{18}$$

This assumption always holds for $A_{y_I} \subset [0,1]^{d_{J^c}-d_x}$. When A_{y_I} is a rectangle with at least one infinite side, an interpretation of this assumption is that the tail probabilities for $\tilde{y}_I$ conditional on (x, y_J) decline uniformly in (x, y_J). Bounded support for $\tilde{y}_I$ is a sufficient condition for this assumption.

Assumption 3 *We assume that*

$$f_{0|J} \in \mathcal{C}^{\beta_{d_J+1},\ldots,\beta_d,L}, \tag{19}$$

where for some $\tau_0 \geq 0$ *and any* $(\tilde{x}, \Delta\tilde{x}) \in \mathbb{R}^{2d_{J^c}}$

$$L(\tilde{x}, \Delta\tilde{x}) = \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}, \tag{20}$$

$$\tilde{L}(\tilde{x} + \Delta\tilde{x}) \leq \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}. \tag{21}$$

The smoothness assumption (19) on the conditional density $f_{0|J}$ is implied by the smoothness of the joint density f_0 at least under boundedness away from zero assumption, see Lemma 10 in Appendix D.3.3. A constant envelop function L used in the lower bound construction would satisfy the assumption.

Assumption 4 *There are positive finite constants* ε *and* $\bar{F}$*, such that for any* $y_J \in \mathcal{Y}_J$ *and* $k = \{k_i\}_{i \in J^c} \in \mathbb{N}_0^{d_{J^c}}$, $\sum_{i \in J^c} k_i/\beta_i < 1$,

$$\int \left[\frac{|D^k f_{0|J}(\tilde{x}|y_J)|}{f_{0|J}(\tilde{x}|y_J)}\right]^{\frac{(2+\varepsilon\beta_{J^c}^{-1} d_{J^c}^{-1})}{\sum_{i \in J^c} k_i/\beta_i}} f_{0|J}(\tilde{x}|y_J) d\tilde{x} < \bar{F}, \tag{22}$$

$$\int \left[\frac{\tilde{L}(\tilde{x})}{f_{0|J}(\tilde{x}|y_J)}\right]^{2+\varepsilon\beta_{J^c}^{-1} d_{J^c}^{-1}} f_{0|J}(\tilde{x}|y_J) d\tilde{x} < \bar{F}. \tag{23}$$

The envelope function and restrictions on its behaviour are mostly relevant for the case of unbounded support. Condition (23) suggests that the envelope function $\tilde{L}$ should be comparable to $f_{0|J}$.

Assumption 5 *For some small* $\nu > 0$,

$$N_J = o(n^{1-\nu}). \tag{24}$$

We impose this assumption to exclude from consideration the cases with very slow (non-polynomial) rates as some parts of the proof require $\log(1/\epsilon_n)$ to be of order $\log n$.

C.2. *Posterior Contraction Rates for Unbounded Support*

Let us define a constant that determines the power of the $\log n$ term in the upper bound on the posterior contraction rate derived below in Theorem 3,

$$t_{J0} = \begin{cases} \frac{d_{J^c}[1+1/(\beta_{J^c}d_{J^c})+1/\tau]+\max\{\tau_1,1,\tau_2/\tau\}}{2+1/\beta_{J^c}} & \text{if } J^c \neq \emptyset \\ \max\{\tau_1, 1\}/2 & \text{if } J^c = \emptyset \end{cases} \tag{25}$$

where (τ, τ_1, τ_2) are defined in Sections 2.1, 4.3.1, and C.1.

Theorem 3 *Suppose the assumptions from Sections 4.3.1 and C.1 hold for a given $J \in \mathcal{A}$. Let*

$$\epsilon_n = \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{26}$$

where $t_J > t_{J0} + \max\{0, (1-\tau_1)/2\}$. Suppose also $n\epsilon_n^2 \to \infty$. Then, there exists $\bar{M} > 0$ such that

$$\Pi\left(p : d_{L_1}(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

As in Section 4.2, when $J^c = \emptyset$, β_{J^c} can be defined to be infinity and $\beta_{J^c}/(2\beta_{J^c}+1) = 1/2$ in (26). Note that in the bounded support case, τ can be chosen arbitrarily large and a simplified expression in Theorem 2 can be used instead of t_{J0} in the lower bound on t_J.

Corollary 1 *Suppose the assumptions from Sections 4.3.1 and C.1 hold for every $J \in \mathcal{A}$. Let*

$$\epsilon_n = \min_{J\in\mathcal{A}} \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{27}$$

where $t_J > t_{J0} + \max\{0, (1-\tau_1)/2\}$. Suppose also $n\epsilon_n^2 \to \infty$. Then, there exists $\bar{M} > 0$ such that

$$\Pi\left(p : d_{L_1}(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

Under the assumptions of the corollary, Theorem 3 delivers a valid upper bound on the posterior contraction rate for every $J \in \mathcal{A}$ including the one for which the minimum in (27) is attained. Hence, the corollary is an immediate implication of Theorem 3. The proof of Theorem 3 is presented below.

C.3. *Proof Outline for Posterior Contraction Results*

To prove Theorem 3, we use the following sufficient conditions for posterior contraction from Theorem 2.1 in Ghosal and van der Vaart (2001). Let ϵ_n and $\tilde{\epsilon}_n$ be positive sequences with $\tilde{\epsilon}_n \leq \epsilon_n$, $\epsilon_n \to 0$, and $n\tilde{\epsilon}_n^2 \to \infty$, and c_1, c_2, c_3, and c_4 be some positive constants. Let ρ be the Hellinger or L_1 distance. Suppose $\mathcal{F}_n \subset \mathcal{F}$ is a sieve with the following bound on the metric entropy $M_e(\epsilon_n, \mathcal{F}_n, \rho)$

$$\log M_e(\epsilon_n, \mathcal{F}_n, \rho) \leq c_1 n\epsilon_n^2, \tag{28}$$

$$\Pi(\mathcal{F}_n^c) \leq c_3 \exp\{-(c_2+4)n\tilde{\epsilon}_n^2\}. \tag{29}$$

Suppose also that the prior thickness condition holds

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq c_4 \exp\{-c_2 n\tilde{\epsilon}_n^2\}, \tag{30}$$

where the generalized Kullback-Leibler neighborhood $\mathcal{K}(p_0, \tilde{\epsilon}_n)$ is defined by

$$\mathcal{K}(p_0, \epsilon) = \Bigg\{ p : \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p_0(y,x) \log \frac{p_0(y,x)}{p(y,x)} dx < \epsilon^2, \\ \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p_0(y,x) \left[\log \frac{p_0(y,x)}{p(y,x)}\right]^2 dx < \epsilon^2 \Bigg\}.$$

Then, there exists $\bar{M} > 0$ such that

$$\Pi\left(p : \rho(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

The definition of the sieve and a verification of conditions (28) and (29) closely follow analogous results in the literature on contraction rates for mixture models in the context of density estimation. The details are given in Lemma 20 in the supplement. Verification of the prior thickness condition is more involved and we formulate it as a separate result in the following theorem.

THEOREM 4 *Suppose the assumptions from Sections 4.3.1 and C.1 hold for a given $J \in \mathcal{A}$. Let $t_J > t_{J0}$, where t_{J0} is defined in (25), and*

$$\tilde{\epsilon}_n = \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}. \tag{31}$$

For any $C > 0$ and all sufficiently large n,

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq \exp\{-Cn\tilde{\epsilon}_n^2\}. \tag{32}$$

Approximation results are key for showing the prior thickness condition (32). Appropriate approximation results for $f_{0J}(y_J, \tilde{x}) = f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)$ are obtained as follows. Based on approximation results for continuous densities by normal mixtures from Shen et al. (2013), we obtain approximations for $f_{0|J}(\cdot|y_J)$ for every y_J in the form

$$f^{\star}_{|J}(\tilde{x}|y_J) = \sum_{j=1}^{K} \alpha^{\star}_{j|y_J}\phi(\tilde{x}; \mu^{\star}_{j|y_J}, \sigma^{\star}_{J^c}), \tag{33}$$

where the parameters of the mixture will be defined precisely below. For the discrete variables over which smoothing is not performed, y_J, we show that $\pi_{0J}(y_J)$ can be appropriately approximated by

$$\int_{A_{y_J}} \sum_{y'_J} \pi_{0J}(y'_J)\phi(\tilde{y}_J; y'_J, \sigma^{\star}_J)d\tilde{y}_J,$$

where $\int_{A_{y_J}} \phi(\tilde{y}_J, y'_J, \sigma^{\star}_J)d\tilde{y}_J$ behaves like an indicator $1\{y_J = y'_J\}$ for sufficiently small $\sigma^{\star}_J$. Section D.3 in the supplement presents proof details.