Journal of Econometrics 168 (2012) 332–346

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian modeling of joint and conditional distributions☆

Andriy Norets*, Justinas Pelenis

Princeton University, United States
Institute for Advanced Studies, Vienna, Austria

ARTICLE INFO

Article history:
Received 12 October 2009
Received in revised form
14 December 2011
Accepted 5 February 2012
Available online 16 February 2012

Keywords:
Mixture of normal distributions
Consistency
Bayesian conditional density estimation
Heteroscedasticity and non-linearity robust inference

ABSTRACT

In this paper, we study a Bayesian approach to flexible modeling of conditional distributions. The approach uses a flexible model for the joint distribution of the dependent and independent variables and then extracts the conditional distributions of interest from the estimated joint distribution. We use a finite mixture of multivariate normals (FMMN) to estimate the joint distribution. The conditional distributions can then be assessed analytically or through simulations. The discrete variables are handled through the use of latent variables. The estimation procedure employs an MCMC algorithm. We provide a characterization of the Kullback–Leibler closure of FMMN and show that the joint and conditional predictive densities implied by the FMMN model are consistent estimators for a large class of data generating processes with continuous and discrete observables. The method can be used as a robust regression model with discrete and continuous dependent and independent variables and as a Bayesian alternative to semi- and non-parametric models such as quantile and kernel regression. In experiments, the method compares favorably with classical nonparametric and alternative Bayesian methods.

1. Introduction

In this paper, we study a Bayesian approach to flexible modeling of conditional distributions. The approach uses a flexible model for the joint distribution of the dependent and independent variables and then extracts the conditional distributions of interest from the estimated joint distribution. We use finite mixtures of multivariate normals (FMMN) to estimate the joint distribution. The conditional distributions can then be assessed analytically or through simulations. The discrete variables are handled through the use of latent variables. The estimation procedure employs an MCMC algorithm. We show that the joint and conditional predictive densities implied by the FMMN model can consistently estimate data generating processes with continuous and discrete observables. The method can also be used as a robust regression model with discrete and continuous dependent and independent variables and as a Bayesian alternative to semi- and non-parametric models such as quantile and kernel regression.

Estimation of conditional distributions has become increasingly important in applied economics as evidenced by a large body of research that uses quantile regression methodology, see, for example, Koenker and Hallock (2001). This area seems to be somewhat overlooked in the Bayesian framework. Moreover, there seems to be no universally accepted regression methodology in the Bayesian literature that would be robust to various violations of assumptions of the normal linear model such as OLS with robust standard errors in the classical framework. The shape of the distribution of the regression error term can be flexibly approximated by mixtures of normals, see, for example, Geweke (2005). Heteroscedasticity in this model can be accommodated by multiplying the error term by a factor that flexibly depends on the covariates, see, for example, Leslie et al. (2007). However, further elaborations on this approach might become too cumbersome if other assumption violations are addressed such as asymmetry of the error distribution that depends on covariates. A flexible and convenient model for conditional distributions seems to be an attractive approach for handling these issues in the Bayesian framework.

If researchers are interested only in conditional distributions, modeling the distribution of covariates might seem to be an unnecessary complication. A promising Bayesian alternative to our approach, a smoothly mixing regression (SMR) also known as a mixture of experts in computer science literature (see, Jacobs et al. (1991), Jordan and Xu (1995), Peng et al. (1996), Wood et al. (2002), Geweke and Keane (2007), Villani et al. (2009)), directly models the conditional distribution of interest by a mixture of regressions where the mixing probabilities are modeled by a multinomial choice model and thus depend on covariates. Norets (2010) and Norets and Pelenis (2011) established that large nonparametric classes of conditional densities can be approximated

☆ First version: November 25, 2008, current version: February 5, 2012.

* Corresponding author. Tel.: +1 609 258 4012.
E-mail addresses: anorets@princeton.edu (A. Norets), pelenis@ihs.ac.at (J. Pelenis).

0304-4076/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2012.02.001

and consistently estimated by several different specifications of SMR and related dependent Dirichlet processes.[1] In contrast to available results for SMR, our results for FMMN do not require compact support for the distribution of covariates. This is an advantage for the approach based on FMMN. Another advantage of FMMN over SMR and other direct conditional approaches is that it is much easier to estimate by MCMC methods. An advantage of the direct approach to conditional density estimation is that it can be combined with procedures for selection of relevant covariates at the estimation stage. This can be accomplished by methods similar to those employed by Villani et al. (2009). We do not consider the issue of covariate selection in FMMN based models.

Ideally, a theoretical comparison of a direct conditional approach and a joint density approach to estimation of conditional densities should be based on the magnitude of the estimation errors or the convergence rates. To the best of our knowledge, posterior convergence rates have not been obtained for either FMMN or SMR (posterior convergence rates for univariate mixture models were obtained in Ghosal and van der Vaart (2007)). Even classical literature on optimal rates of convergence for conditional distributions is very limited. Efromovich (2007) derived the minimax rates for conditional densities $f(y|x)$ with univariate x and y. His results suggest that if the joint and conditional densities are equally smooth then the minimax convergence rates for them are the same and if the conditional density is smoother then it can be estimated at a faster rate. However, it is not clear if a slower rate for the joint density estimator implies a slower rate for the conditional density estimator derived from it. Thus, a definitive theoretical resolution of the issue of the conditional approach versus the unconditional approach is yet to be obtained and it is an important direction for future research.

Our method is global and it does not have logical inconsistencies that some frequentist methods have, for example, crossing quantiles in the quantile regression. Moreover, experiments on real data show that out of sample prediction quality of FMMN compares favorably with that of the state of the art kernel based methods, DPM, and SMR.

An approach similar to ours can be implemented with Dirichlet process mixtures (DPM). Muller et al. (1996) and Taddy and Kottas (2010) suggest using DPM models for regression and quantile regression correspondingly. However, these papers do not study theoretical properties of these procedures. An advantage of a DPM based model is that every number of mixture components has a positive probability and there is no need to select it. At the same time, in finite samples the number of mixture components generating the data is necessarily finite and the number of components that appears in estimation is determined by the prior. Also, the estimation algorithm is easier to implement and the prior is more flexible for the FMMN model. Therefore, we chose to work with FMMN.

Section 2 sets up the model for the joint distribution and shows how to extract the conditional distributions of interest. The Gibbs sampler for exploring the posterior distribution of the model parameters and a log scoring rule for evaluating model prediction quality are presented in Section 3. The consistency of the predictive density is shown in Section 4. Section 5 applies the method to several datasets that were previously analyzed by quantile regression and kernel methods. Appendix contains proofs of theoretical results. Experiments with artificial data, joint distribution tests for checking correctness of posterior simulator implementation (Geweke, 2004), an algorithm for computing the marginal likelihood, and some extra estimation experiments are delegated to a web appendix, Norets and Pelenis (2009).

[1] A growing literature on dependent Dirichlet processes includes the following papers, among others: MacEachern (1999), De Iorio et al. (2004), Griffin and Steel (2006), Dunson and Park (2008), Chung and Dunson (2009), and Pati et al. (2011).

2. Finite mixture of normals model

A model in the Bayesian framework specifies the joint distribution of the observables, unobservables, and objects of interest. First, we describe the model for continuous observables. Then, we show how to extend the model to the case of discrete components in the observables.

2.1. Continuous observables

The observables in the model are denoted by $Y_T = \{y_t, t = 1, \ldots, T\}$, where $y_t = (y_{t,1}, \ldots, y_{t,d}) \in R^d$. In the context of a regression model, $y_{t,1}$ is a dependent variable and $y_{t,-1} = (y_{t,2}, \ldots, y_{t,d})$ are covariates. The observables density is given by

$$p(y_t|\theta, \mathcal{M}_m) = \sum_{j=1}^{m} \alpha_j \cdot \phi(y_t; \mu_j, H_j^{-1}), \tag{1}$$

where $\mathcal{M}_m$ stands for the model with m mixture components, $\phi(y_t; \mu_j, H_j^{-1})$ is a density of a multivariate normal distribution with mean μ_j and variance H_j^{-1} (H_j is called precision), $\alpha = (\alpha_1, \ldots, \alpha_m)$ are mixing probabilities, vector $\theta = (\alpha, \mu_1, H_1, \ldots, \mu_m, H_m) \in \Theta_m$ collects all the parameters in the model, and Θ_m is the parameter space. We use the (conditionally) conjugate prior distribution $p(\theta|\mathcal{M}_m)$, which is described in Section 3.1.

Predictive joint and conditional distributions of y are of interest in our analysis. The predictive joint distribution is

$$p(y|Y_T, \mathcal{M}_m) = \int p(y|\theta, \mathcal{M}_m) p(\theta|Y_T, \mathcal{M}_m) d\theta, \tag{2}$$

where $p(y|\theta, \mathcal{M}_m)$ is given by the observables distribution in (1) and $p(\theta|Y_T, \mathcal{M}_m)$ is the posterior distribution of the parameters. The predictive conditional distribution is

$$p(y_1|y_{-1}, Y_T, \mathcal{M}_m) = \int p(y_1|y_{-1}, \theta, \mathcal{M}_m) p(\theta|Y_T, \mathcal{M}_m) d\theta.$$

The conditional distribution $p(y_1|y_{-1}, \theta, \mathcal{M}_m)$ is a mixture of (conditional) normals:

$$p(y_1|y_{-1}, \theta, \mathcal{M}_m) \propto \sum_{j=1}^{m} \alpha_j \phi(y_{-1}; \mu_{j,-1}, H_{j,-1}^{-1}) \times \phi(y_1|y_{-1}; \mu_j, H_j^{-1}), \tag{3}$$

where $\phi(y_{-1}; \mu_{j,-1}, H_{j,-1}^{-1})$ is the marginal normal distribution of y_{-1} implied by the joint normal $\phi(y; \mu_j, H_j^{-1})$, $\phi(y_1|y_{-1}; \mu_j, H_j^{-1})$ is the conditional normal distribution of y_1 implied by the joint normal $\phi(y; \mu_j, H_j^{-1})$, and the mixing probabilities are given by

$$\frac{\alpha_j \phi(y_{-1}; \mu_{j,-1}, H_{j,-1}^{-1})}{\sum_k \alpha_k \phi(y_{-1}; \mu_{k,-1}, H_{k,-1}^{-1})}.$$

The joint and conditional densities of interest and expectations with respect to them can be evaluated by simulation: $\theta^k \sim p(\theta|Y_T, \mathcal{M}_m)$ (draws from the posterior), $y^k \sim p(y|\theta^k, \mathcal{M}_m)$, and $y_1^k \sim p(y_1|y_{-1}, \theta^k, \mathcal{M}_m)$.

2.2. Discrete components in observables

It is common in the Bayesian framework to model discrete variables by continuous latent variables for computational reasons, see, for example, Albert and Chib (1993) and Chapter 6 in Geweke (2005). We also use latent variables to handle discrete observables.

Let us denote continuous components in observables vector $y \in R^{d+K}$ by $y_c \in R^d$ and discrete components by y_{-c}, where the subscript c stands for continuous and K is the number of discrete variables. Suppose the kth discrete variable can take N_k different values, where $k \in \{1, \ldots, K\}$. We map possible values of each discrete variable into a partition of R by intervals. Thus, $y_{-c} = [a^1_{l_1}, b^1_{l_1}] \times \cdots \times [a^K_{l_K}, b^K_{l_K}]$, $l_k \in \{1, \ldots, N_k\}$ and $R = \cup_{l_k=1}^{N_k} [a^k_{l_k}, b^k_{l_k}]$ for every $k \in \{1, \ldots, K\}$. For each discrete variable we introduce a corresponding latent variable in the model, $y^*_{-c} \in y_{-c}$. The density of the latent variables and continuous observables is modeled as a mixture of normals,

$$p(y_c, y^*_{-c}|\theta, \mathcal{M}_m) = \sum_{j=1}^{m} \alpha_j \cdot \phi(y_c, y^*_{-c}; \mu_j, H_j^{-1}). \tag{4}$$

The conditional density of the discrete observables with respect to the counting measure is an indicator function

$$p(y_{-c}|y_c, y^*_{-c}, \theta, \mathcal{M}_m) = 1_{y_{-c}}(y^*_{-c}). \tag{5}$$

The observables density with respect to the product of the Lebesgue and counting measures conditional on parameters is given by the integral of the product of (4) and (5) with respect to y^*_{-c},

$$p(y_c, y_{-c}|\theta, \mathcal{M}_m) = \sum_{j=1}^{m} \alpha_j \phi(y_c; \mu_{j,c}, H_{j,c}^{-1}) \times \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu_j, H_j^{-1}) d(y^*_{-c}), \tag{6}$$

where $\phi(y_c; \mu_{j,c}, H_{j,c}^{-1})$ is the marginal normal distribution of y_c implied by the joint normal $\phi(y_c, y^*_{-c}; \mu_j, H_j^{-1})$ and $\phi(y^*_{-c}|y_c; \mu_j, H_j^{-1})$ is the conditional normal distribution of y^*_{-c} given y_c implied by the joint normal $\phi(y_c, y^*_{-c}; \mu_j, H_j^{-1})$. As described in the previous subsection, the draws from $p(y_1|y_{-1}, \theta, \mathcal{M}_m)$ can be used for evaluating the predictive conditional densities of interest.

3. Estimation method

3.1. Gibbs sampler

The posterior distribution of the parameters is explored by the Gibbs sampler. A convenient parameterization of the Gibbs sampler for finite mixture models involves introduction of latent state variables (Diebolt and Robert, 1994): $s_t \in \{1, \ldots, m\}$, $p(y_t|s_t, \theta, \mathcal{M}_m) = \phi(\cdot; \mu_{s_t}, H_{s_t}^{-1})$ and $P(s_t = j|\theta, \mathcal{M}_m) = \alpha_j$. The posterior is proportional to the joint distribution of the observables and unobservables,

$$\begin{aligned} p(\{y_t, s_t\}_{t=1}^T; \{\alpha_j, \mu_j, H_j\}_{j=1}^m|\mathcal{M}_m) &\propto \prod_{t=1}^{T} \alpha_{s_t} |H_{s_t}|^{0.5} \exp\{-0.5(y_t - \mu_{s_t})' H_{s_t} (y_t - \mu_{s_t})\} \\ &\times \alpha_1^{\underline{a}-1} \cdots \alpha_m^{\underline{a}-1} \\ &\times \prod_j |H_j|^{0.5} \exp\{-0.5(\mu_j - \underline{\mu})' \underline{\lambda} H_j (\mu_j - \underline{\mu})\} \\ &\times \prod_j |H_j|^{(\underline{\nu}-d-1)/2} \exp\{-0.5 \mathrm{tr} \underline{S} H_j\}. \end{aligned} \tag{7}$$

We used conditionally conjugate priors: Normal–Wishart for (μ_j, H_j) and Dirichlet for α. Hyper-parameters $(\underline{\nu}, \underline{S}, \underline{\mu}, \underline{\lambda}, \underline{a})$ have to be specified by the researcher in each particular application. We provide some suggestions on this in Section 5.

The densities for the Gibbs sampler blocks are proportional to the joint distribution in (7). The Gibbs sampler block for the latent states has a multinomial distribution,

$$p(s_t = j| \ldots) \propto \alpha_j |H_j|^{0.5} \exp\{-0.5(y_t - \mu_j)' H_j (y_t - \mu_j)\}.$$

The block for mixing probabilities is Dirichlet,

$$p(\alpha| \ldots) \propto \alpha_1^{\sum_t 1\{s_t=1\}+\underline{a}-1} \cdots \alpha_m^{\sum_t 1\{s_t=m\}+\underline{a}-1}. \tag{8}$$

The block for the mean and precision of the mixture components is given by

$$\begin{aligned} p(\mu_j, H_j| \ldots) &\propto \prod_{t:s_t=j} |H_j|^{0.5} \exp\{-0.5(y_t - \mu_j)' H_j (y_t - \mu_j)\} \\ &\quad \times |H_j|^{0.5} \exp\{-0.5(\mu_j - \underline{\mu})' \underline{\lambda} H_j (\mu_j - \underline{\mu})\} \\ &\quad \times |H_j|^{(\underline{\nu}-d-1)/2} \exp\{-0.5 \mathrm{tr} \underline{S} H_j\} \\ &\propto |H_j|^{(T_j+\underline{\nu}-d)/2} \\ &\quad \times \exp\left\{-0.5 \mathrm{tr}\left(H_j \left[\sum_{t:s_t=j} (y_t - \mu_j)(y_t - \mu_j)' \right.\right.\right. \\ &\quad \left.\left.\left. + \underline{\lambda}(\mu_j - \underline{\mu})(\mu_j - \underline{\mu})' + \underline{S}\right]\right)\right\} \\ &\propto |H_j|^{(T_j+\underline{\nu}-d)/2} \\ &\quad \times \exp\left\{-0.5 \mathrm{tr}\left(H_j \left[\sum_{t:s_t=j} (y_t - \bar{y}_j)(y_t - \bar{y}_j)' \right.\right.\right. \\ &\quad + T_j(\bar{y}_j - \mu_j)(\bar{y}_j - \mu_j)' \\ &\quad \left.\left.\left. + \underline{\lambda}(\mu_j - \underline{\mu})(\mu_j - \underline{\mu})' + \underline{S}\right]\right)\right\}, \end{aligned}$$

where $T_j = \sum_t 1\{s_t = j\}$ and $\bar{y}_j = T_j^{-1} \sum_{t:s_t=j} y_t$. Thus, $p(\mu_j|H_j, \ldots)p(H_j| \ldots)$ is a Normal–Wishart distribution:

$$\begin{aligned} H_j &\sim \mathit{Wishart}\left(T_j + \underline{\nu}, \left[\sum_{t:s_t=j} (y_t - \bar{y}_j)(y_t - \bar{y}_j)' \right.\right. \\ &\quad \left.\left. + \frac{T_j \underline{\lambda}}{T_j + \underline{\lambda}} (\bar{y}_j - \underline{\mu})(\bar{y}_j - \underline{\mu})' + \underline{S}\right]^{-1}\right) \\ \mu_j &\sim N\left(\frac{T_j \bar{y}_j + \underline{\lambda}\underline{\mu}}{T_j + \underline{\lambda}}, [(T_j + \underline{\lambda}) H_j]^{-1}\right). \end{aligned} \tag{9}$$

We initially chose a Normal–Wishart prior for (μ_j, H_j) because it simplifies computation of the marginal likelihood (see web appendix, Norets and Pelenis (2009)). With independent conditionally conjugate priors for μ_j and H_j, the Gibbs sampler would have a normal block for μ_j and a Wishart block for H_j (the derivation of the block densities is similar in this case).

If the observables have discrete components then in the Gibbs sampler described above one can replace y_t with $(y_{t,c}, y^*_{t,-c})$ and add blocks for components of the latent variables $y^*_{t,-c}$. A block for the kth component of $y^*_{t,-c}$ has a truncated normal distribution,

$$p(y^*_{t,-c,k}| \ldots) \propto \exp\{-0.5((y_{t,c}, y^*_{t,-c}) - \mu_{s_t})' H_{s_t} ((y_{t,c}, y^*_{t,-c}) - \mu_{s_t})\} \cdot 1_{y_{t,-c}}(y^*_{t,-c}).$$

In the model we described, the posterior for parameters is symmetric with respect to label switching for the parameters. For example, marginal posteriors of (μ_j, H_j, α_j) are the same for every j. For larger values of m the described MCMC algorithm might not produce enough label switching to obtain identical marginal posteriors for (μ_j, H_j, α_j). However, as demonstrated by Geweke (2007), the lack of label switching in MCMC is usually not a problem in mixture models as long as objects of interest are label invariant, which is the case in this paper.

3.2. Log scoring rules

We initially used the marginal likelihood (ML) to evaluate model performance. An algorithm for computing the ML based on Chib (1995) and Marin and Robert (2008) is presented in a web appendix Norets and Pelenis (2009). When the number of variables, especially discrete ones, is large, computation of the ML is numerically unstable. Therefore, we use log scoring rules instead. Another important reason for using log scores is that they can be computed for non-Bayesian models and thus can be useful in comparison of FMMN with classical alternatives. A full cross-validated log score (Gelfand et al., 1992) is given by

$$\sum_{t=1}^{T} \log p(y_t|Y_{T/t}, \mathcal{M}_m) \approx \sum_{t=1}^{T} \log\left(\frac{1}{N}\sum_{n=1}^{N} p(y_t|Y_{T/t}, \theta^n, \mathcal{M}_m)\right) \tag{10}$$

$$\sum_{t=1}^{T} \log p(y_{t,i}|y_{t,-i}, Y_{T/t}, \mathcal{M}_m) \approx \sum_{t=1}^{T} \log\left(\frac{1}{N}\sum_{n=1}^{N} p(y_{t,i}|y_{t,-i}, Y_{T/t}, \theta^n, \mathcal{M}_m)\right), \tag{11}$$

where $Y_{T/t}$ denotes the sample with observation t removed and θ^n's are draws from posterior $p(\theta|Y_{T/t}, \mathcal{M}_m)$. Eq. (10) should be used if the joint probability distribution is of interest, while Eq. (11) should be used if the conditional distribution of the i-th element is of interest (additional advantage of log scoring rules over the ML is that the former can evaluate models when the conditional distribution is of interest). A full cross-validated log scoring rule requires T posterior simulators for each model specification. A modified cross-validated log scoring rule Geweke and Keane (2007) is more computationally efficient. Under this rule, the sample is ordered randomly and the first T_1 observations are used for estimation and the rest for computing the log score. This procedure is repeated K times and the means or medians of the obtained log scores are used for model comparison. The following formula explicitly shows how the mean log score is computed,

$$\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{t=T_1+1}^{T} \log p(y_{t,i}^k|y_{t,-i}^k, Y_{T_1}^k, \mathcal{M}_m)\right), \tag{12}$$

where Y^k denotes a random reordering of Y and $p(y_{t,i}^k|y_{t,-i}^k, Y_{T_1}^k, \mathcal{M}_m)$ is computed as in (11).

4. Consistency of FMMN

In this section, we study frequentist properties of the predictive densities and the posterior distribution when the sample size converges to infinity. Consistency has been accepted as a minimal requirement on priors in the Bayesian nonparametrics literature, see Ghosh and Ramamoorthi (2003) for a textbook treatment. Below, we briefly review the most closely related previous work and then discuss our consistency results.

4.1. Existing results

In the classical framework, Genovese and Wasserman (2000) showed that if the true density f on R is a general normal mixture then a maximum likelihood sieve is consistent in the Hellinger distance. In the Bayesian framework, the theoretical results of Roeder and Wasserman (1997) are most closely related to what we do in this section of the paper. Roeder and Wasserman (1997) show that the posterior probability of any total variation neighborhood of the true density f converges to 1 if f on R is in the Kullback–Leibler (KL) closure of finite mixtures of normals and $m = o(T/\log(T))$. The Roeder and Wasserman (1997) prior was chosen so that their finite mixture of normals model approached a model based on the Dirichlet process prior. The result the authors get is related to analogous results in the literature on the Dirichlet process priors, see Ghosh and Ramamoorthi (2003) and Ghosal and van der Vaart (2007). Our results hold for the true density on R^d not R.

Additionally, we provide a characterization of the Kullback–Leibler closure of FMMN. In some papers, the true density is often assumed to be in some special class, for example, general mixtures of normals in Genovese and Wasserman (2000) or KL closure of finite mixtures of normals in Roeder and Wasserman (1997). However, no description of these classes is provided. It is not immediately clear what densities can actually be estimated in practice. Thus, our characterization of the KL closure of FMMN can be useful for developing and applying other theoretical results for FMMN.

When this manuscript was presented at a conference we learned about recent related work by Wu and Ghosal (2010) who study posterior consistency of Dirichlet process mixtures (DPM) of multivariate kernels in multivariate density estimation. Some of their sufficient conditions for consistency look similar to our characterization of the Kullback–Leibler closure of FMMN. Our conditions are weaker. However, we note that the model and the type of consistency results in Wu and Ghosal (2010) differ from what we consider in this paper. Another distinction of our work from the previous literature is that we develop theoretical results for categorical observables in addition to continuous ones.

4.2. Theoretical results

First, we consider the case when the number of mixture components m is fixed. Under mild regularity conditions, we demonstrate that for a given $\epsilon > 0$ there exists m such that the L_1 distance between the predictive density and the data generating process (DGP) density is less than ϵ almost surely (a.s.). This result is presented in Theorems 1–3. Theorem 1 states that if the posterior concentrates on the parameter values, Θ_m^*, that minimize the KL distance between the DGP density and the model and if this distance is small then the L_1 distance between the predictive density and the DGP density is small as well. The concentration of the posterior on Θ_m^* in parametric problems is a standard result (see Theorems 3.4.1–3.4.2 in Geweke (2005)), which is related to similar results for the maximum likelihood estimator. In Theorem 2, we provide a set of mild sufficient conditions for FMMN that imply the posterior concentration result. In Theorem 3, we characterize a class of the DGP densities that can be arbitrarily well approximated by FMMN in the KL distance. The theory is first formulated for continuous observables. Analogous results for observables that can be discrete are given in Corollaries 1 and 2.

At the end of the section, we discuss the posterior consistency for FMMN based models, in which a prior on the number of mixture components is specified. Our characterization of the KL closure of FMMN in Theorem 3 combined with the Schwartz (1965) posterior consistency theorem immediately implies that the posterior is weakly consistent. More generally, the characterization of the KL closure of FMMN we obtain in Theorem 3 is also of independent interest as KL neighborhoods of the DGP densities play a major role in the general theory of weak and strong posterior consistency in Bayesian nonparametrics (Ghosh and Ramamoorthi, 2003).

We assume that the parameter space Θ_m in model with m mixture components $\mathcal{M}_m$ is compact and the observables

$Y_T = (y_1, \ldots, y_T)$ are independently identically distributed (i.i.d.) random variables with density $f(y) \equiv p(y|D)$, where D stands for the data generating process. Frequentist probabilistic statements in which Y_T is treated as random are written conditional on D as in $p(y|D)$. Proofs of the theoretical results in this section are given in Appendix.

Theorem 1. *Suppose the following two conditions hold. First, the DGP density f is in the KL closure of the finite mixtures of normals, i.e., for any $\epsilon > 0$ there exists m and $\theta \in \Theta_m$ such that*

$$d_{KL}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m)) = \int \log \frac{f(y)}{p(y|\theta, \mathcal{M}_m)} F(dy) \leq \epsilon.$$

Second, the posterior concentrates on the parameter values that minimize the KL distance between the true density and the model, $\Theta_m^* = \arg\min_{\theta \in \Theta_m} d_{KL}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m))$, i.e., for any open set $N(\Theta_m^*)$ such that $\Theta_m^* \subset N(\Theta_m^*)$,

$$P(N(\Theta_m^*)|Y_T, \mathcal{M}_m) \underset{T\to\infty}{\to} 1 \quad a.s.$$

Then, for any $\epsilon > 0$ and all sufficiently large m the probability that the L_1 distance between the predictive density defined in (2) and the DGP density is smaller then ϵ converges to 1,

$$\lim_{T\to\infty} P[d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)) < \epsilon | D] = 1,$$

where $d_{L1}(f(\cdot), p(\cdot)) = \int |f(y) - p(y)| dy$. Actually, a stronger result holds,

$$\lim_{T\to\infty} P\left(\bigcap_{t=T}^{\infty} [d_{L1}(f(\cdot), p(\cdot|Y_t, \mathcal{M}_m)) < \epsilon] | D\right) = 1,$$

which means that $[d_{L1}(f(\cdot), p(\cdot|Y_t, \mathcal{M}_m)) < \epsilon]$ holds a.s. The same results hold for the conditional predictive density if the following distance between conditional distributions is used,

$$d(f(\cdot|\cdot), p(\cdot|\cdot, Y_T, \mathcal{M}_m)) = \int d_{L1}(f(\cdot|y_{-1}), p(\cdot|y_{-1}, Y_T, \mathcal{M}_m)) \times f(y_{-1}) dy_{-1}.$$

A parameter value that minimizes the KL distance between the true density and the FMMN model is not unique; Θ_m^* includes at least $m!$ parameters that differ only by labels. Furthermore, it is not clear whether Θ_m^* can contain more than $m!$ elements. Fortunately, this issue is not important for our results. The following theorem gives conditions under which the posterior concentrates on set Θ_m^*.

Theorem 2. *Suppose that*

1. *$p(y|D)$ has finite second moments;*
2. *the prior distribution of θ_m is absolutely continuous: $P(\theta_m \in C|\mathcal{M}_m) > 0$ for all $C \subseteq \Theta_m$ for which $\int_C d\theta_m > 0$;*
3. *any precision matrix in a parameter vector from Θ_m is non-negative definite with eigenvalues in $[\underline{\lambda}_m, \overline{\lambda}_m]$, where $\underline{\lambda}_m > 0$ and $\overline{\lambda}_m < \infty$.*

Then, $T^{-1} \log p(Y_T|\theta_m, \mathcal{M}_m) \overset{a.s.}{\to} l(\theta_m; \mathcal{M}_m)$ uniformly for all $\theta_m \in \Theta_m$, where $l(\theta_m; \mathcal{M}_m)$ is a continuous function of θ_m with a set of maxima

$$\Theta_m^* = \arg\max_{\theta\in\Theta_m} l(\theta; \mathcal{M}_m) = \arg\min_{\theta\in\Theta_m} d_{KL}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m))$$

and for any open set $N(\Theta_m^*)$ such that $\Theta_m^* \subset N(\Theta_m^*)$,

$$\lim_{T\to\infty} P(\theta \in N(\Theta_m^*)|Y_T, \mathcal{M}_m) = 1 \quad a.s.$$

The following theorem describes the conditions on $f(\cdot)$ that guarantee that $f(\cdot)$ can be approximated in KL distance by finite mixtures of normals. In other words, it characterizes the KL closure of FMMN.

Theorem 3. *Let A_j^m, $j = 0, 1, \ldots, m$, be a partition of Y, where $A_1^m, \ldots, A_m^m$ are adjacent cubes with side length h_m and A_0^m is the rest of set Y. Assume that*

1. *$f(y)$ is continuous on Y except on a set of F measure zero.*
2. *The second moments of y are finite.*
3. *For any y there exists a cube $C(r, y)$ with side length $r > 0$ and $y \in C(r, y)$ such that* (i)

$$\int \log \frac{f(y)}{\inf_{z\in C(r,y)} f(z)} F(dy) < \infty \tag{13}$$

and (ii) *there exists an M such that for any $m \geq M$, if $y \in A_0^m$ then $C(r, y) \cap A_0^m$ contains a cube $C_0(r, y)$ with side length $r/2$ and a vertex at y and if $y \in Y \setminus A_0^m$ then $C(r, y) \cap (Y \setminus A_0^m)$ contains a cube $C_1(r, y)$ with side $r/2$ and a vertex at y.*
4. *An upper bound on the eigenvalues of a precision matrix, $\overline{\lambda}_m$, in a parameter vector from Θ_m satisfies $\overline{\lambda}_m \to \infty$.*

Then, for any $\epsilon > 0$ there exists m and $\theta \in \Theta_m$ such that

$$d_{KL}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m)) \leq \epsilon.$$

The strongest assumption of Theorem 3 is that of the finite second moments. The proof of the theorem suggests that it can be weakened if components with tails heavier than normal, for example, Student t, are added to the mixture of normals.

Condition 4 implies that the variance of mixture components can be arbitrarily close to zero as m increases. Since the variance of mixture components plays the role of bandwidth, arbitrarily small variances of mixture components are required for approximation of large non-parametric classes of DGP densities by FMMN.

It seems hard to find a positive everywhere density that would violate condition 3 of the theorem. For example, normal, exponential, extreme value, and Student t densities do satisfy this condition. Part 3(i) of the condition requires local difference in $\log f(y)$ to be integrable. When $f(y)$ is positive everywhere, part 3(ii) of the condition is not needed. It always holds if $C(r, y)$ is a hypercube with center at y. Part 3(ii) becomes important when $f(y)$ can be equal to zero. In particular, the sets $C_0(r, y)$ and $C_1(r, y)$ in condition 3(ii) are introduced to specify that $C(r, y)$ needs to be defined differently near the boundary of the support and in the tails if one wants to use condition (13) in its present form. The following example illustrates how one could verify the theorem conditions.

Example 1. Consider an exponential distribution, $f(y) = \gamma \exp\{-\gamma y\} 1\{y \geq 0\}$, $\gamma > 0$. The density is continuous in y on $Y = [0, \infty)$ and it has finite second moments. Thus conditions 1 and 2 hold. To verify part (i) of condition 3, for some $r > 0$ let $C(r, y) = [y, y + r]$ for $y \in [0, r]$ and $C(r, y) = [y - r/2, y + r/2]$ for $y \in (r, \infty)$. Then,

$$\int \log \frac{f(y)}{\inf_{z\in C(r,y)} f(z)} F(dy) = \int_0^r rF(dy) + \int_r^\infty \frac{r}{2} F(dy) \leq r.$$

Note that if we defined $C(r, y) = [y - r/2, y + r/2]$ for all y then $\inf_{z\in C(r,y)} f(z) = 0$ for $y \in [0, r/2)$ and the condition would fail. To verify part (ii) of condition 3 define the partition A_j^m and intervals $C_0(r, y)$ and $C_1(r, y)$ as follows. For $h_m > 0$ such that $h_m m \to \infty$, let $A_j^m = [(j-1)h_m, jh_m)$ for $j > 0$ and $A_0^m = [mh_m, \infty)$. For all sufficiently large m, $r < h_m m$ and for $y \in A_0^m$, $C_0(r, y) = [y, y + r/2] \subset A_0^m \cap C(r, y)$. For $y \in Y \setminus A_0^m$, $C_1(r, y) = [y - r/2, y] \subset (Y \setminus A_0^m) \cap C(r, y)$. Thus, part (ii) of 3 is satisfied.

Corollaries 1 and 2 state that the theoretical results obtained in Theorems 1–3 for continuous observables also hold for categorical observables and the latent variable model defined in Section 2.2. Since Theorem 3 has an independent value for establishing more general consistency results we present its extension to the discrete variable case separately in Corollary 2. The corollaries' assumptions about the DGP density of the continuous observables conditional on the categorical observables are essentially the same as the corresponding assumptions about the DGP density of the observables in the continuous case.

Corollary 1. *When some of the observables are discrete Theorems 1 and 2 apply without any changes for the model with the observables density $p(y_c, y_{-c}|\theta, \mathcal{M}_m)$ with respect to the product of the Lebesgue and counting measures defined in (6).*

Corollary 2 (*Analog of Theorem 3 When Some of the Observables are Discrete*)**.** *Let A_j^m, $j = 0, 1, \ldots, m$, be a partition of the continuous part of the observables space Y_c, where $A_1^m, \ldots, A_m^m$ are adjacent cubes with side length h_m and A_0^m is the rest of set Y_c. Assume that*

1. *$f(y_c|y_{-c})$ is continuous on Y_c except on a set of F measure zero, $\forall y_{-c} \in Y_{-c}$.*
2. *The second moments of y_c are finite.*
3. *For any y_c and y_{-c} there exists a cube $C(r, y_c, y_{-c})$ with side length $r > 0$ and $y_c \in C(r, y_c, y_{-c})$ such that* (i)

$$\int \log \frac{f(y_c|y_{-c})}{\inf_{z \in C(r,y_c,y_{-c})} f(z|y_{-c})} F(dy) < \infty \tag{14}$$

and (ii) *there exists an M such that for any $m \geq M$, if $y_c \in A_0^m$ then $C(r, y_c, y_{-c}) \cap A_0^m$ contains a cube $C_0(r, y_c, y_{-c})$ with side length $r/2$ and a vertex at y_c and if $y_c \in Y_c \setminus A_0^m$ then $C(r, y_c, y_{-c}) \cap (Y_c \setminus A_0^m)$ contains a cube $C_1(r, y_c, y_{-c})$ with side $r/2$ and a vertex at y_c.*
4. *An upper bound on the eigenvalues of a precision matrix in a parameter vector from Θ_m satisfies $\bar{\lambda}_m \to \infty$.*

Then, for any $\epsilon > 0$ there exists m and $\theta \in \Theta_m$ such that

$$d_{KL}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m)) \leq \epsilon.$$

We next consider a model with a prior on the number of mixture components. Let $\mathcal{M}_\infty$ stand for a collection of FMMN models $\{\mathcal{M}_m\}_{m=1}^\infty$ with corresponding prior model probabilities $\{p_m\}_{m=1}^\infty$. Model $\mathcal{M}_\infty$ defines a prior probability measure on the space of densities. To demonstrate the posterior consistency in this model we use the following immediate implication of Schwartz (1965) posterior consistency theorem.

Theorem 4. *Suppose a prior, P, on the space of densities on Y puts a positive probability on any KL neighborhood of DGP $f(\cdot)$:*

$$P\left(p : \int \log \frac{f}{p} dF < \epsilon\right) > 0, \quad \forall \epsilon > 0.$$

Then, the corresponding posterior is weakly consistent. Specifically, for any neighborhood U of $f(\cdot)$ in the topology of weak convergence,

$$P(U|Y_T) \underset{T\to\infty}{\to} 1 \quad a.s.$$

A proof of this theorem can be found in Ghosh and Ramamoorthi (2003) (see Theorems 4.4.1 and 4.4.2).

In Theorem 3 we found a sequence of parameter values θ_m for models $\mathcal{M}_m$ such that $d_{KL}(f(\cdot), p(\cdot|\theta_m, \mathcal{M}_m)) \to 0$ as $m \to \infty$. Using the Lebesgue dominated convergence theorem as in the proof of Theorem 3, one can show that $d_{KL}(f(\cdot), p(\cdot|\theta_m, \mathcal{M}_m))$ is continuous in parameters at θ_m for all sufficiently large m, even when general variance covariance matrices are used instead of the diagonal $(\sigma_j^m)^2 I$ in $\mathcal{M}_m$. Therefore, as long as $p_m > 0$ and $p(\theta|\mathcal{M}_m)$ puts positive probability on the neighborhoods of θ_m for all m, conditions of Theorem 4 are satisfied. Thus, Theorem 4 applies to any DGP f that satisfies the assumptions of Theorem 3.

5. Applications

In this section, we present five applications of FMMN (see a web appendix Norets and Pelenis (2009) for more applications and artificial data experiments). In the first application, we apply FMMN to a large dataset and explore the sensitivity of the estimation results to the prior specification. We also provide some suggestions on choosing reasonable values for prior hyper-parameters. In Sections 5.2 and 5.3, we compare out-of-sample performance of FMMN to that of kernel smoothing methods, Dirichlet Process mixtures, and smoothly mixing regressions. We employ kernel estimation methods from Hall et al. (2004) who use cross-validation to select estimation parameters such as bandwidth. Comparison with Hall et al. (2004) methods is particularly relevant since these methods were shown to outperform many other alternatives, see Hall et al. (2004), Li and Racine (2007), and Li and Racine (2008). Hall et al. (2004) methods are implemented by a publicly available R package **np** (Hayfield and Racine, 2008), which we use in this paper. In Section 5.4, we show that FMMN is capable of handling discrete variables that take a large number of different values. In Section 5.5 we use FMMN for estimating a conditional density for a two-dimensional dependent variable.[2] Overall, the section demonstrates that in a variety of settings FMMN performs very well against a wide range of alternatives.

5.1. Infant birth weight

Abrevaya (2001) and Koenker and Hallock (2001) use linear quantile regression to study factors that affect infant birth weight. Their data include observations on infant birth weight, infant sex, pregnancy trimester of the first doctor visit, cigarettes per day smoked by the mother during pregnancy, mother's weight gain, age, education, marital status and race. We use the same data as Koenker and Hallock (2001): June 1997 Detailed Natality Data published by the National Center for Health Statistics. In our specification, we use the infant weight, demeaned mother's weight gain, and demeaned age as continuous variables. The other six variables are treated as discrete. The total number of observations available in the dataset is around 200,000. Experiments below are conducted for three random subsamples from these data that have different sizes: $T = 1000, 10{,}000, 100{,}000$. In reporting the results below, we specify which subsample was used by the sample size T. We also consider different number of mixture components: $m \in \{10, 20, 50, 100\}$.

We employ the following benchmark prior:

$$(\underline{\nu} = 20, \underline{S} = 20I, \underline{\mu} = 0, \underline{\lambda} = 1, \underline{a} = 10). \tag{15}$$

Fig. 1 shows marginal densities estimated by kernel smoothing and marginal posterior predictive densities estimated by a model with $m = 10$ from $T = 1000$ observations. The density estimates produced by the two methods are expected to be similar in large samples as the methods are consistent. Fig. 2 shows that the fit for marginal predictive densities improves considerably when larger m and T are used.

We next explore how sensitive the results are to prior specification. For each of the following model sizes: $m \in \{10, 20, 50\}$, we consider the following five changes to the benchmark prior (15):

$$\begin{aligned} &(1)\ \underline{\nu} = 20, \quad (2)\ \underline{\nu} = 50, \quad (3)\ \underline{S} = .2I, \\ &(4)\ \underline{a} = 50, \quad (5)\ \underline{a} = 3. \end{aligned} \tag{16}$$

[2] Anonymous referees suggested we check the model performance in the settings of Sections 5.4 and 5.5.

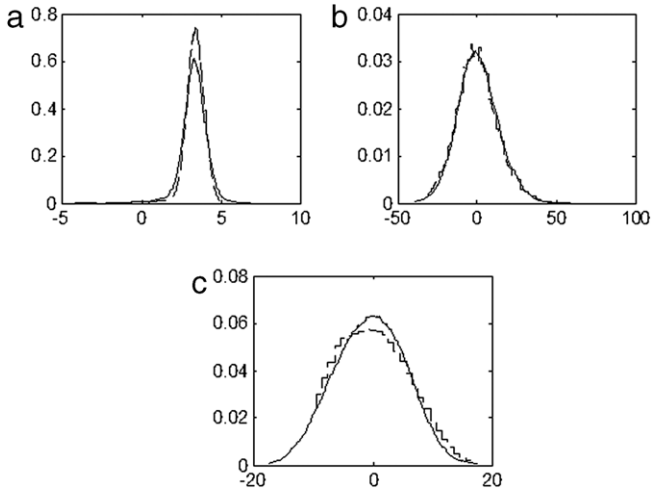

Fig. 1. Marginal densities estimated by kernel smoothing (dotted) and posterior predictive densities (solid), $T = 1000$, $m = 10$. (a) Birth weight, (b) mother weight gain, (c) age.

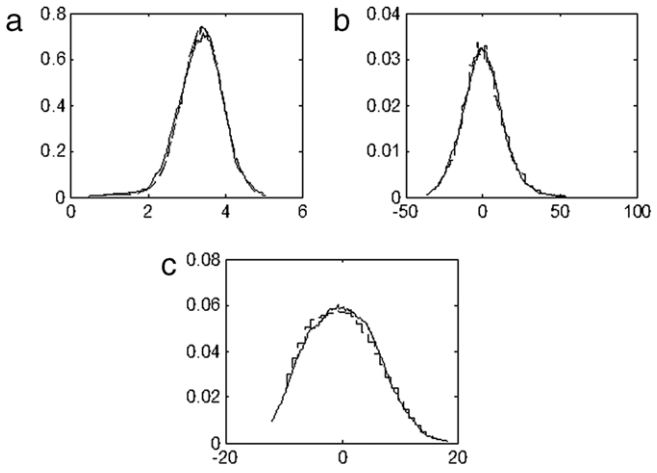

Fig. 2. Marginal densities estimated by kernel smoothing (dotted) and posterior predictive densities (solid), $T = 100{,}000$, $m = 100$. (a) Birth weight, (b) mother weight gain, (c) age.

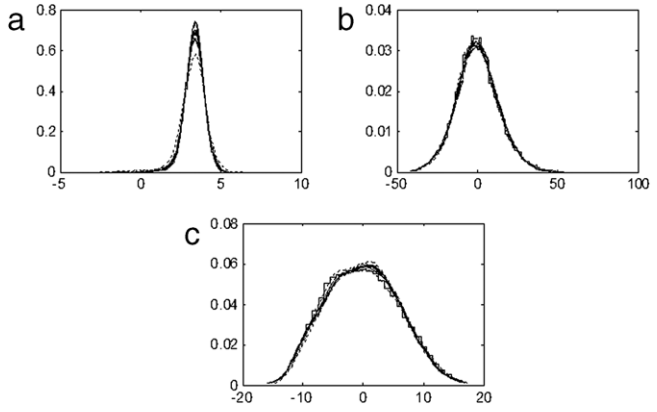

Fig. 3. Marginal densities for combinations of 5 different priors and $m \in \{10, 20, 50\}$ (total 15 models) and $T = 10{,}000$: (a) birth weight, (b) mother weight gain, (c) age.

Fig. 3 shows that the fit for marginal densities is not very sensitive to priors. In the prior sensitivity experiments, we use 20,000 draws from the MCMC algorithm. There seems to be no need for a burn-in period for models with $m \in \{10, 20\}$. For $m = 50$, 2000 first draws are discarded. On each MCMC iteration we also produce

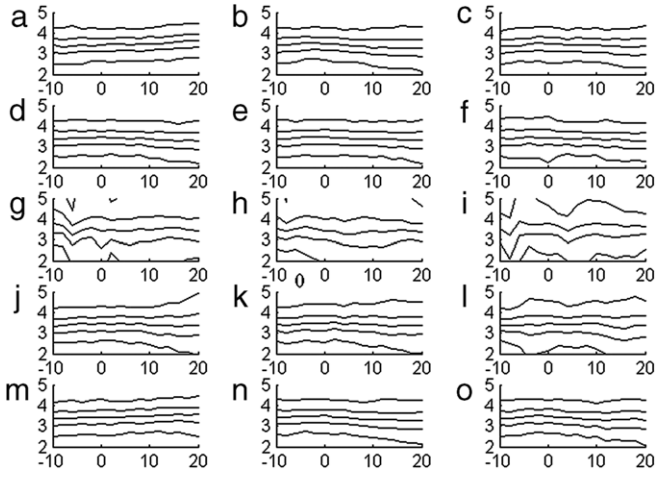

Fig. 4. 5%, 25%, 50%, 75%, 95% quantiles of birth weight conditional on demeaned mother age. Combinations of the 5 priors in (16) (rows) and $m \in \{10, 20, 50\}$ (columns).

a predictive distribution draw from $p(y|\theta, \mathcal{M}_m)$ conditional on the parameter values at that iteration. The relative numerical efficiencies (RNEs) for draws from the predictive distribution, which are label invariant, are in 0.4–1 range for $m \in \{10, 20\}$ and in 0.15–1 range for $m = 50$. To produce 100 draws from the posterior for $T = 10{,}000$ and $m = 10$, $m = 20$, and $m = 50$ it takes correspondingly 46, 83, and 323 s on a laptop with Intel 1.6 GHz processor and 4 GB of RAM memory (the MCMC algorithm is implemented in the C programming language).

Fig. 4 demonstrates that in contrast to the marginal densities, the conditional quantiles can be very sensitive to prior specification.

The conditional quantiles of birth weight shown in the figure, are computed for $[-10, 20]$ range of demeaned age and the following values of the rest of the variables: infant sex — girl, demeaned weight gain — zero, cigarettes smoked — zero, education — at least high school, natal visit — first trimester, marital status — married, and race — non-black.

Most of the observations have the demeaned age in the $[-10, 10]$ range. In this range, the results in Fig. 4 are similar across different priors and model sizes, except for row (g)–(i), that corresponds to prior (3) in (16). Apparently, this prior shrinks variances toward zero too much: the prior mean for H is $100I$ while sample variances are in 1–20 range (the other priors use the prior mean for H equal to I). Thus, one needs to be careful in setting prior hyperparameters for variances in FMMN to avoid excessive shrinkage. Another important point about prior specification is that the Dirichlet hyper-parameter should exceed one. Otherwise, all the observations get assigned to one or two mixture components and the estimation results for conditional distributions could be nonsensical. Changing other prior hyperparameters in reasonable ranges does not seem to have a considerable effect on estimation results. Our estimation experiments on data from this and following subsections also suggest that centering prior for means around sample means and prior for variances around a half or a smaller fraction of the sample variance works well. Prior sensitivity analysis with smaller sample sizes ($T = 500$) and a smaller number of mixture components ($m = 3$) deliver similar results.

5.2. Boston housing data

In this section, we consider 1970s-era Boston housing data that has been analyzed by a number of authors, see for example, Li and Racine (2008). This dataset contains $T = 506$ observations

with the response variable being the median price of the house in an area. We focus on three important covariates: average number of rooms in the area, percentage of the population having lower economic status in the area and the weighted distance to five Boston employment centers. This particular set of variables is chosen to replicate the analysis in Li and Racine (2008). We want to determine whether FMMN can perform equally well or even better than the nonparametric conditional density kernel estimator used by Li and Racine (2008) for this specific empirical application. Furthermore, we evaluate the performance of FMMN against DPM and SMR models as well.

We estimate the distribution of the median price conditional on the other three covariates. All the variables are treated as continuous. Modified cross-validated log scoring rule (see Section 3.2) is used as a measure of performance. For $K = 100$ random reorderings of the sample, the sample is split into an estimation part with $T_1 = 400$ observations and an evaluation part of $T_2 = 106$ observations. For FMMN models with $m \in \{3, 4, 5, 6, 7\}$, we employ the following prior: $(\underline{\nu} = 5, \underline{S} = \underline{\nu} \cdot \mathrm{diag}(100, 50, 5, 0.5) \cdot 0.25, \underline{\mu} = (23, 13, 4, 6)', \underline{\lambda} = 1, \underline{a} = 3)$. The prior of the mean is chosen to be close to the sample mean, and the prior mean of the precision is similar to the sample precision multiplied by a factor of 4 (the theory suggests that the precision of each mixture component should be larger than the precision of the DGP density).

In the SMR specification, the means of the mixed normals are linear in x and the mixing probabilities are modeled by a multinomial logit with linear indices in x. The prior for the coefficients in the means of the mixed normals is centered at zero and the prior for the intercepts is centered at the sample average. The prior variance for the coefficients is 100. The prior mean for the precision of the mixed normals is centered at the sample precision multiplied by 2. The prior standard deviation for the precision of the mixed normals is equal to the prior mean multiplied by 2. The prior on the logit parameters is normal with zero mean and variance 100. The prior for DPM specification is chosen as suggested in Section 2.3 of Taddy and Kottas (2010) (those priors are sample mean and sample range dependent).

The results summarizing the predictive ability of different models are presented in Table 1 below. For FMMN models the number of MCMC draws was 10,000 with first 2500 draws discarded. The convergence of MCMC chain is assessed using separated partial means test for MCMC for first and second moments of the predictive distribution draws and out-of-sample log scores. We focus on these variables rather than posterior draws of the parameters since they are label invariant. Posterior simulation takes approximately 80 s for a single FMMN model with $m = 7$ and less for smaller values of m on a desktop with Intel 2.80 GHz processor and 4 GB of RAM memory. The numerical standard errors for individual out-of-sample log scores for each simulation of FMMN models range from 0.04 to 0.2 and the RNEs range from 0.05 to 0.5.

Table 1 reveals that FMMN models with $m > 3$ outperform nonparametric kernel conditional density estimator and DPM, and perform comparably with SMR. The superior performance of FMMN seems to be a result of the in-sample overfitting by the DPM and kernel smoothing methods.

5.3. Labor market participation

In this section, we use Gerfin (1996) cross-section dataset containing $T = 872$ observations of seven variables that are used to model labor market participation of married Swiss women.[3]

Table 1
Modified cross-validated log scores.

Model	Log score	
	Mean	Median
Kernel	−293.17	−289.44
FMMN ($m = 3$)	−291.57	−289.16
FMMN ($m = 4$)	−282.10	−281.66
FMMN ($m = 5$)	−278.66	−278.93
FMMN ($m = 6$)	−278.40	−278.53
FMMN ($m = 7$)	−278.26	−278.57
SMR ($m = 4$)	−280.29	−280.31
SMR ($m = 7$)	−280.23	−279.19
DPM	−286.97	−284.41

Table 2
Modified cross-validated log scores and classification rates.

Model	Log score		% Correct rate	
	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

A binary variable LFP is equal to 1 if the woman is an active labor force participant and is equal to 0 otherwise. We wish to evaluate the predictive performance of alternative estimators for this binary variable. Moreover, this dataset contains both discrete and continuous variables and we would like to check whether FMMN model performs well when some variables are categorical. We consider a FMMN model, a linear probit model as in Gerfin (1996), and a nonparametric conditional density kernel estimator of Hall et al. (2004).

We estimate the distribution of the variable LFP conditional on log of non-labor income, age, education, number of young children, number of old children, and foreign dummy. We treat the age and log of non-labor income as continuous and the rest of the variables as categorical. Modified cross-validated log scoring rule (see Section 3.2) and correct classification rates are used as two alternative measures of the predictive performance of an estimator. For $K = 100$ random reorderings of the data, the sample is split into a prediction part with $T_1 = 650$ observations and an evaluation part of $T_2 = 222$ observations. For FMMN models with $m \in \{1, 2, 3, 4\}$ we employ the following prior: $(\underline{\nu} = 8, \underline{S} = \underline{\nu} \cdot \mathrm{diag}(0.2, 1, 0.25, 10, 0.5, 1, 0.2) \cdot 0.25, \underline{\mu} = (0, 11, 4, 9, 0, 1, 0)', \underline{\lambda} = 1, \underline{a} = 3)$. The prior mean for coefficients is chosen to be similar to the sample mean, and the prior mean for the precision matrix is chosen to be close to the sample precision multiplied by 4. For FMMN models the number of MCMC draws was 10,000 with first 2500 discarded. The convergence of MCMC chain was assessed using separated partial means test for MCMC for first and second moments of the predictive distribution draws. Posterior simulation takes approximately 200 s for FMMN model with $m = 4$ on a desktop with an Intel 2.80 GHz processor and 4 GB of RAM. We use every 75th of remaining 7500 iterations for log-score computation as evaluating multivariate truncated integrals of normal distributions is computationally intensive. The results summarizing the predictive ability of different models are presented in Table 2. The numerical standard errors for individual log scores range from 0.2 to 0.6 and the RNEs are higher than 0.9. For the correct classification rates the numerical standard errors range between 0.1% and 0.2% and the RNEs are higher than 0.9 (the serial correlation is very low for the thinned draws). As can be seen from Table 2, the FMMN model with $m = 2$ already outperforms both the kernel and probit methods in out-of-sample prediction judging by both the modified log score and

[3] The data for this study can be obtained online at http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/.

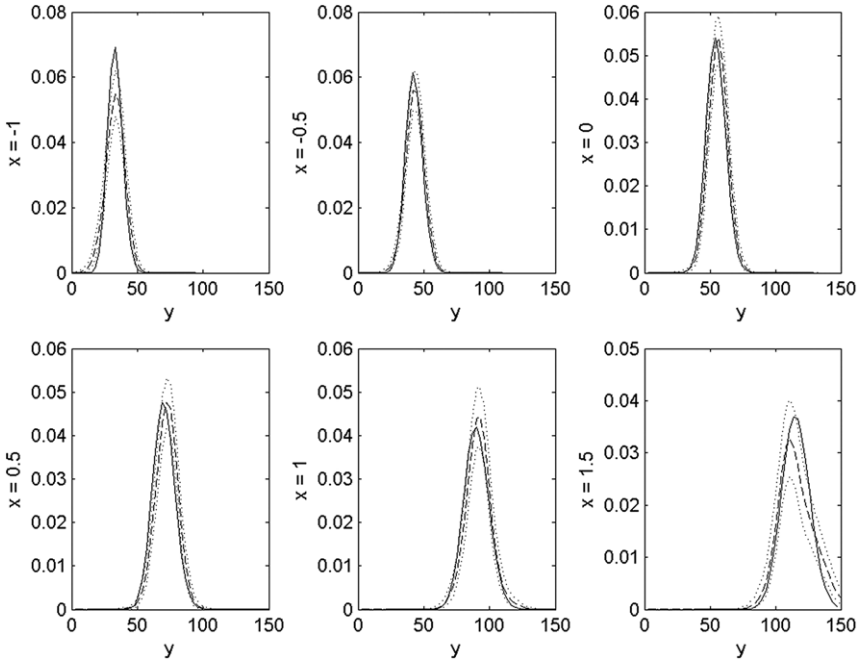

Fig. 5. The solid lines are truth, dashed lines are posterior mean and dotted lines are 95% highest posterior density region estimates of conditional probability weights $f(y|x)$, where x is the value on the y-axis. The number of observations is $T = 500$.

the correct classification rates. Results suggest that the FMMN model is an attractive alternative to classical parametric and nonparametric techniques for conditional distribution estimation for both continuous and categorical data.

5.4. Poisson regression

This section evaluates whether the FMMN method can handle discrete variables that take a large number of different values. We generate a sample of $T = 500$ observations of a continuous covariate x and a discrete response variable y. The data generating process is given by: $x_i \sim N(0, 1)$, $y_i \sim$ Poisson $(\beta_0 + \beta_1 x_i)$. The values of $\beta_0 = 4$ and $\beta_1 = 0.5$ are chosen so that the discrete variable y_i takes a large number of distinct values. We use the number of mixture components $m = 10$. The prior is set to: $(\underline{\nu} = 10, \underline{S} = \underline{\nu} \cdot \text{sample cov} \cdot 0.25, \underline{\mu} = \text{sample means}, \underline{\lambda} = 1, \underline{a} = 10)$. The results are based on 12,500 draws from posterior simulator with first 2500 draws discarded. Posterior simulation takes approximately 2 mins on a desktop with Intel 2.80 GHz processor and 4 GB of RAM memory. MCMC convergence is assessed by a separated partial means test for first and second moments of the predictive distribution draws. Numerical standard errors of the predictive distribution draws for y and x are equal to 0.8 and 0.02 and RNEs are equal to 0.78 and 0.86. In Fig. 5, we plot posterior estimates for conditional probability weights $f(y|x)$ for varying values of x. As can be seen from the figure, a FMMN model can estimate reasonably well the conditional distribution of a discrete variable with a large support.

Of course, this is a low dimensional example and more extensive theoretical work, simulations, applications to real data, and comparisons with other approaches would be necessary to better evaluate FFMN performance in settings with a large number of discrete variables. We leave this for future work.

5.5. NLSY data

This section applies FMMN to estimation of conditional distribution of a multivariate variable. We consider National Longitudinal Survey NLSY79 dataset from 2002 interview for subjects that were first surveyed in 1979.[4] We model weekly hours worked and hourly earnings as a function of years of schooling and total out-of-school work experience and we focus only on male subjects and all the observable variables are treated as continuous. The number of observations is equal to 5400.

The prior is set to: $(\underline{\nu} = 10, \underline{S} = \underline{\nu} \cdot \text{diag}([300, 100, 20, 10]) \cdot 0.25, \underline{\mu} = ([20, 04, 17, 14])', \underline{\lambda} = 1, \underline{a} = 3)$. We model the data as FMMN with $m = 10$. The results are based on 25,000 draws from the posterior simulator with first 5000 draws discarded. Posterior simulation takes approximately 40 min on a desktop with Intel 2.80 GHz processor and 4 GB of RAM memory. The numerical standard errors for the draws from the predictive distribution are on the level of 1% to 2% of the sample standard deviation. The RNEs for the predictive distribution draws are all above 0.65. The convergence of posterior is assessed through a separated partial means test. In Fig. 6, we plot predictive density for hours worked per week and hourly earnings conditional on a level of schooling and prior work experience. We condition on *schooling* $= \{12, 16\}$ and *work experience* $= \{14, 18, 22\}$.

The figure shows how earnings increase with schooling and work experience. Also, one can explore in the figure how the earnings differ for part time and full time workers. Overall, the subsection demonstrates that FMMN model can be a useful tool for estimation of conditional distributions of multivariate variables.

Acknowledgments

We thank John Geweke, Chris Sims, Bo Honore, participants of the microeconometrics seminar at Princeton, and seminar participants at the University of Montreal and University of Toronto for helpful discussions. We also thank anonymous referees for suggestions that helped to improve the paper. All remaining errors are ours.

[4] Information about this survey can be obtained online at http://www.bls.gov/nls/nlsy79.htm.

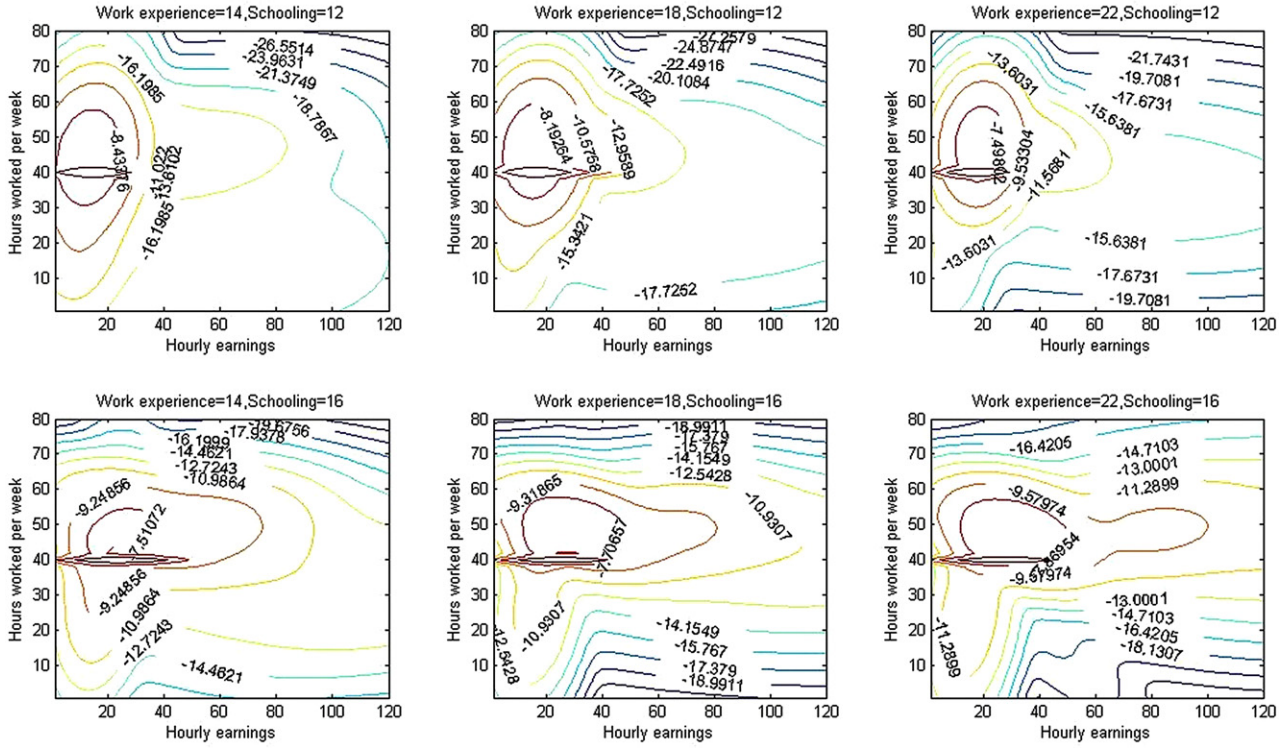

Fig. 6. Log of the predictive density of hourly earnings and weekly working hours conditional on work experience and schooling.

Appendix. Proofs

A.1. Continuous data

Proof (Theorem 1)**.** First, note that

$$
\begin{aligned}
&d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)) \\
&\quad = \int \left| \int_{\Theta_m} f(y) \cdot p(\theta_m|Y_T, \mathcal{M}_m) d\theta_m \right. \\
&\qquad \left. - \int_{\Theta_m} p(y|\theta_m, \mathcal{M}_m) \cdot p(\theta_m|Y_T, \mathcal{M}_m) d\theta_m \right| dy \\
&\quad \leq \iint_{\Theta_m} |f(y) - p(\theta_m|Y_T, \mathcal{M}_m)| \cdot p(\theta_m|Y_T, \mathcal{M}_m) d\theta_m dy \\
&\quad = \int_{\Theta_m} \int |f(y) - p(\theta_m|Y_T, \mathcal{M}_m)| dy \\
&\qquad \times p(\theta_m|Y_T, \mathcal{M}_m) d\theta_m. \qquad (17)
\end{aligned}
$$

Second, by the theorem assumptions, given $\epsilon > 0$ there exists m and $\widehat{\theta}_m \in \Theta_m$ such that

$$d_{KL}(f(y), p(y|\widehat{\theta}_m, \mathcal{M}_m)) < \frac{\epsilon}{8}.$$

If $\theta_m \in \Theta_m^*$ then

$$
\begin{aligned}
d_{L1}(f(\cdot), p(\cdot|\theta_m, \mathcal{M}_m)) &= \int |f(y) - p(y|\theta_m, \mathcal{M}_m)| dy \\
&\leq 2\, d_{KL}(f(\cdot), p(\cdot|\theta_m, \mathcal{M}_m)) \\
&= 2 \min_{\theta_m \in \Theta_m} d_{KL}(f(\cdot), p(\cdot|\theta_m, \mathcal{M}_m)) \\
&\leq 2\, d_{KL}(f(\cdot), p(\cdot|\widehat{\theta}_m, \mathcal{M}_m)) < \frac{\epsilon}{4}.
\end{aligned}
$$

Since $d_{L1}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m))$ is uniformly continuous in θ by Lemma 3 below, there exists $\delta > 0$ such that $\forall \theta \in N(\Theta_m^*) = \bigcup_{\theta \in \Theta_m^*} [\tilde{\theta} : \|\tilde{\theta} - \theta\| < \delta]$, $d_{L1}(f(\cdot), p(\cdot|\theta, \mathcal{M}_m)) < \epsilon/2$. This inequality and (17) imply

$$
\begin{aligned}
d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)) &\leq 2P(N(\Theta_m^*)^c|Y_T, \mathcal{M}_m) \\
&\quad + P(N(\Theta_m^*)|Y_T, \mathcal{M}_m) \cdot \frac{\epsilon}{2}.
\end{aligned}
$$

By the theorem assumptions, $R(Y_T) \equiv P(N(\Theta_m^*)^c|Y_T, \mathcal{M}_m) \underset{T\to\infty}{\to} 0$ a.s. So, $d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)) < 2R(Y_T) + \epsilon/2$ and

$$[d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)) < \epsilon] \supset \left[R(Y_T) < \frac{\epsilon}{4}\right].$$

As $R(Y_T) \to 0$ a.s., we have

$$P[d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)) < \epsilon|D] \geq P\left[R(Y_T) < \frac{\epsilon}{4}|D\right] \to 1$$

and

$$
\begin{aligned}
1 &= P\left(\bigcup_{T=1}^{\infty} \bigcap_{t=T}^{\infty} \left[R(Y_t) < \frac{\epsilon}{4}\right]|D\right) \\
&\leq P\left(\bigcup_{T=1}^{\infty} \bigcap_{t=T}^{\infty} [d_{L1}(f(\cdot), p(\cdot|Y_t, \mathcal{M}_m)) < \epsilon]|D\right) \\
&= \lim_{T\to\infty} P\left(\bigcap_{t=T}^{\infty} [d_{L1}(f(\cdot), p(\cdot|Y_t, \mathcal{M}_m)) < \epsilon]|D\right).
\end{aligned}
$$

The same results follow for the conditional predictive density since

$$
\begin{aligned}
d(f(\cdot|\cdot), p(\cdot|\cdot, Y_T, \mathcal{M}_m)) &= \int d_{L1}(f(\cdot|y_{-1}), p(\cdot|y_{-1}, Y_T, \mathcal{M}_m)) \\
&\quad \times f(y_{-1}) dy_{-1} \\
&\leq 2 d_{L1}(f(\cdot), p(\cdot|Y_T, \mathcal{M}_m)). \quad \square
\end{aligned}
$$

Proof (Theorem 2). First, let us show that

$$T^{-1}\log\ p(Y_T|\theta_m,\mathcal{M}_m) = \frac{1}{T}\sum_{t=1}^{T}\log\ p(y_t|\theta_m,\mathcal{M}_m) \xrightarrow{a.s.} l(\theta_m;\mathcal{M}_m)$$

uniformly for all $\theta_m \in \Theta_m$. Note that

$$p(y_t|\theta_m,\mathcal{M}_m) = \sum_{j=1}^{m}\alpha_j\cdot\phi(y_t;\mu_j,H_j^{-1}) \le \max_{j=1,\ldots,m}|H_j|^{1/2} \le \underline{\lambda}_m^{-0.5d}, \quad \forall\theta_m\in\Theta_m.$$

Also,

$$\begin{aligned}\log\ p(y_t|\theta_m,\mathcal{M}_m) &\ge \sum_{j=1}^{m}\alpha_j\cdot\Big[\log(2\pi)^{-d/2}+\frac{1}{2}\log\ |H_j| - 0.5(y-\mu_j)'H_j(y-\mu_j)\Big]\\ &\ge \log(2\pi)^{-d/2}+0.5d\log\underline{\lambda}_m - 0.5\max_j[y'H_jy-2yH_j\mu_j+\mu_j'H_j\mu_j]\\ &\ge \log(2\pi)^{-d/2}+0.5d\log\underline{\lambda}_m-0.5\bar{\lambda}y'y - \|y\|\max_j\|H_j\mu_j\|-\max_j\|\mu_j'H_j\mu_j\|.\end{aligned}$$

Since eigenvalues of H_j are bounded above and away from zero and since $\|H_j\|$ and $\|\mu_j\|$ are bounded on Θ_m,

$$|\log\ p(y_t|\theta_m,\mathcal{M}_m)| \le q(y_t),$$

where $q(y_t)$ is integrable because $p(y|D)$ has finite second moments by the theorem assumptions. Also, $\log\ p(y|\theta_m)$ is continuous in θ and measurable in y. Thus, by Theorem 2 in Jennrich (1969), we get uniform a.s. convergence. $l(\theta;\mathcal{M}_m)$ is continuous by the dominated convergence theorem.

Second, let $N = N(\Theta_m^*)$, $L(\theta) = l(\theta,\mathcal{M}_m)$, $L_0 = \max L(\theta)$, and $L_2 = \max_{\theta\in N^c}L(\theta) < L_0$. We claim that there exists L_1 such that $L_2 < L_1 < L_0$ and $H = \{\theta : L(\theta) > L_1\} \subset N$. Suppose that the claim is false. Then, $\forall L_1 < L_0$, $\exists\theta\in N^c$ such that $L(\theta) > L_1$. Similarly, for a sequence $L_1^n \uparrow L_0$, there exists a sequence $\theta^n \in N^c$, such that $L(\theta^n) > L_1^n$. Since N^c is compact there exists a convergent subsequence $\theta^{n_k}\to\bar{\theta}\in N^c$ and $L(\theta^{n_k})\to L(\bar{\theta}) = L_0$. Therefore, $\bar{\theta}\in N^c$ and $L(\bar{\theta}) = L_0$ and we get a contradiction to the statement that $\max_{\theta\in N^c}L(\theta) < L_0$. Hence, there exists L_1 and H such that $L_2 < L_1 < L_0$ and $H = \{\theta : L(\theta) > L_1\} \subset N$. Then,

$$\begin{aligned}&\frac{P(\theta\in N^c|Y_T,\mathcal{M}_m)}{P(\theta\in N|Y_T,\mathcal{M}_m)}\\ &\le \frac{\int_{\theta\in N^c}P(Y_T|\theta,\mathcal{M}_m)p(\theta|\mathcal{M}_m)/p(Y_T|\mathcal{M}_m)d\theta}{\int_{\theta\in H}P(Y_T|\theta,\mathcal{M}_m)p(\theta|\mathcal{M}_m)/p(Y_T|\mathcal{M}_m)d\theta}\\ &\le \frac{P(\theta\in N^c)}{P(\theta\in H)}\cdot\frac{\sup_{\theta\in N^c}p(Y_T|\theta,\mathcal{M}_m)}{\inf_{\theta\in H}p(Y_T|\theta,\mathcal{M}_m)}.\end{aligned} \tag{18}$$

Note that $P(\theta\in H) > 0$ because H is open and non-empty and the prior is absolutely continuous by assumption. Since log is a strictly monotone function, the density $p(\cdot|\cdot)$ is always positive, and the convergence to $L(\theta)$ uniform a.s.,

$$\begin{aligned}&T^{-1}\log\left(\frac{\sup_{\theta\in N^c}p(Y_T|\theta,\mathcal{M}_m)}{\inf_{\theta\in H}p(Y_T|\theta,\mathcal{M}_m)}\right)\\ &= \sup_{\theta\in N^c}T^{-1}\log(p(Y_T|\theta,\mathcal{M}_m)) - \inf_{\theta\in H}T^{-1}\log(p(Y_T|\theta,\mathcal{M}_m))\\ &\to \sup_{\theta\in N^c}L(\theta) - \inf_{\theta\in H}L(\theta) \le L_2 - L_1 < 0.\end{aligned}$$

Consequently, the a.s. limit of (18) is 0 and $P(\theta\in N|Y_T,\mathcal{M}_m)\to 1$. □

Proof (Theorem 3). Parameter values θ_m for approximating $f(y)$ by FMMN model $\mathcal{M}_m$ are defined by

$$p(y|\theta_m,\mathcal{M}_m) = \sum_{j=1}^{m}F(A_j^m)\phi(y;\mu_j^m,\sigma_m) + F(A_0^m)\phi(y;0,\sigma_0), \tag{19}$$

where σ_0 is fixed, σ_m converges to zero as m increases, and $\mu_j^m\in A_j^m$. Since d_{KL} is always non-negative,

$$\begin{aligned}0 &\le \int\log\frac{f(y)}{p(y|\theta_m,\mathcal{M}_m)}F(dy)\\ &\le \int\log\max\left\{1,\frac{f(y)}{p(y|\theta_m,\mathcal{M}_m)}\right\}F(dy).\end{aligned}$$

We will use dominated convergence theorem (DCT) to show that the last integral in the inequality above converges to zero as m increases.

First, we will show the point-wise convergence of the integrand to zero a.s. F. For fixed y and a cube $C_{\delta_m}(y)$ with center y and side length $\delta_m > 0$

$$\begin{aligned}p(y|\theta_m,\mathcal{M}_m) &= \sum_{j=1}^{m}F(A_j^m)\phi(y;\mu_j^m,\sigma_m)+F(A_0^m)\phi(y;0,\sigma_0)\\ &\ge \inf_{z\in C_{\delta_m}(y)}f(z)\sum_{j:A_j^m\subset C_{\delta_m}(y)}\lambda(A_j^m)\phi(y;\mu_j^m,\sigma_m),\end{aligned} \tag{20}$$

where λ is the Lebesgue measure.

It is always possible to construct a partition, in which elements $A_1^m,\ldots,A_m^m$ are adjacent cubes with side length h_m ($\lambda(A_j^m) = h_m^d$ for $j > 0$) and for any y there exists M_0 such that

$$\forall m\ge M_0,\quad C_{\delta_m}(y)\cap A_0^m = \emptyset. \tag{21}$$

In Lemmas 1 and 2 below, the following bounds for the Riemann sum in (20) are derived (the Riemann sum is not far from the corresponding normal integral and the integral is not far from 1)

$$\sum_{j:A_j^m\subset C_{\delta_m}(y)}\lambda(A_j^m)\phi(y;\mu_j^m,\sigma_m) \ge 1-\frac{3d\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d}-\frac{8d\sigma_m}{(2\pi)^{1/2}\delta_m}. \tag{22}$$

Let δ_m,σ_m,h_m satisfy the following

$$\delta_m\to 0,\qquad \sigma_m/\delta_m\to 0,\qquad h_m/\sigma_m^d\to 0. \tag{23}$$

Given $\epsilon > 0$ there exists M_1 such that for $m\ge M_1$ expressions in (22) are bounded below by $(1-\epsilon)$.

For any y at which f is continuous and positive there exists M_2 such that for $m\ge M_2$, $[f(y)/\inf_{z\in C_{\delta_m}(y)}f(z)]\le(1+\epsilon)$ since $\delta_m\to 0$.

For any $m\ge\max\{M_0,M_1,M_2\}$,

$$\begin{aligned}1 &\le \max\left\{1,\frac{f(y)}{p(y|\theta_m,\mathcal{M}_m)}\right\}\le\max\left\{1,\frac{f(y)}{\inf_{z\in C_{\delta_m}(y)}f(z)(1-\epsilon)}\right\}\\ &\le\frac{1+\epsilon}{1-\epsilon}.\end{aligned}$$

Thus, $\log\max\{1,f(y)/p(y|\theta_m,\mathcal{M}_m)\}\to 0$ for any y satisfying $f(y) > 0$, which implies convergence a.s. F since $f(y) > 0$ on Y except on a set of F measure zero by assumptions of Theorem 3.

Second, we will establish an integrable upper bound on $\log\max\{1, f(y|x)/p(y|x,\mathcal{M}_m)\}$.

$$\begin{aligned} p(y|\theta_m,\mathcal{M}_m) &= \sum_{j=1}^{m} F(A_j^m)\phi(y;\mu_j^m,\sigma_m) \\ &\quad + F(A_0^m)\phi(y;0,\sigma_0) \\ &\geq [1-1_{A_0^m}(y)]\cdot \inf_{z\in C_1(r,y)} f(z) \\ &\quad \times \sum_{j:A_j^m\subset C_1(r,y)} \lambda(A_j^m)\phi(y;\mu_j^m,\sigma_m) \\ &\quad + 1_{A_0^m}(y)\cdot \inf_{z\in C_0(r,y)} f(z) \\ &\quad \times \lambda(C_0(r,y))\phi(y;0,\sigma_0). \end{aligned} \tag{24}$$

Lemmas 1 and 2 imply that the Riemann sum in (24) is bounded below by $2^{-d}-2^{-(d+1)}=2^{-(d+1)}$ for any m larger than some M_4. Parameter σ_0 can be chosen so that

$$1 > 2^{-(d+1)} > \phi(y;0,\sigma_0)\lambda(C_0(r,y)). \tag{25}$$

This implies

$$\begin{aligned} &\log\max\left\{1, \frac{f(y)}{p(y|\mathcal{M}_m)}\right\} \\ &\leq \log\max\left\{1, \frac{f(y)}{\inf_{z\in C(r,y)} f(z)\cdot\phi(y;0,\sigma_0)\cdot(r/2)^d}\right\} \\ &\leq \log\frac{1}{\phi(y;0,\sigma_0)(r/2)^d} \\ &\quad \times \max\left\{\phi(y;0,\sigma_0)(r/2)^d, \frac{f(y)}{\inf_{z\in C(r,y)} f(z)}\right\} \\ &\leq -\log(\phi(y;0,\sigma_0)(r/2)^d) + \log\frac{f(y)}{\inf_{z\in C(r,y)} f(z)}. \end{aligned} \tag{26}$$

Inequality (26) follows by (25). The first expression in (26) is integrable by condition 2 in Theorem 3. The second expression in (26) is integrable by condition 3 of Theorem 3.

Since we have established pointwise convergence and integrable upper bound, we can apply the DCT. Henceforth, given $\epsilon > 0 \exists M$ such that for any $m > M$ and θ_m defined in (19), $d_{KL}(f(\cdot), p(\cdot|\theta,\mathcal{M}_m)) \leq \epsilon$. □

Lemma 1. *Define a cube $C_\delta(y)=\{\mu\in R^d : y_i\leq\mu_i\leq y_i+\delta, i=1,\ldots,d\}$. Let $A_1,\ldots,A_m$ be adjacent cubes with centers μ_j and side length h such that $C_\delta(y)\subset\cup_{j=1}^m A_j$ and $\delta>3h$. Define $J=\{j : A_j\subset C_\delta(y)\}$. Then,*

$$\sum_{j\in J}\lambda(A_j)\phi(y;\mu_j,\sigma)\geq\int_{C_\delta(y)}\phi(\mu;y,\sigma)d\mu-\frac{3d\delta^{d-1}h}{(2\pi)^{d/2}\sigma^d}.$$

By symmetry, this result holds for any cube with vertex at y and side length δ. This implies that for cube $D_\delta(y)=\{x : y_i-\delta/2\leq x_i\leq y_i+\delta/2, i=1,\ldots,d\}$,

$$\begin{aligned}\sum_{j:A_j\subset D_\delta(y)}\lambda(A_j)\phi(y;\mu_j,\sigma) &\geq \int_{D_\delta(y)}\phi(\mu;y,\sigma)d\mu \\ &\quad - 2^d\frac{3d(\delta/2)^{d-1}h}{(2\pi)^{d/2}\sigma^d}\end{aligned}$$

as long as $D_\delta(y)\subset\cup_{j=1}^m A_j$ and $\delta>6h$.

Proof. For $j\in J$ let $B_j=\{x : \mu_{ji}\leq x_i\leq\mu_{ji}+h, i=1,\ldots,d\}$ be a rotated and shifted version of A_j so that the sides of B_j are parallel to the sides of $C_\delta(y)$. Note that $\mu_j=\arg\max_{\mu\in B_j}\phi(\mu;y;\sigma)$ and therefore

$$\begin{aligned}\sum_{j\in J}\lambda(A_j)\phi(y;\mu_j,\sigma) &= \sum_{j\in J}\lambda(B_j)\phi(y;\mu_j,\sigma) \\ &\geq \int_{\cup_J B_j}\phi(\mu;y,\sigma)d\mu \\ &\geq \int_{C_\delta(y)}\phi(\mu;y,\sigma)d\mu \\ &\quad - \int_{C_\delta(y)\setminus\cup_J B_j}\phi(\mu;y,\sigma)d\mu.\end{aligned}$$

Since $\{x : \min_J\mu_{ji}\leq x_i\leq\max_J\mu_{ji}, i=1,\ldots,d\}\subset C_\delta(y)\cap[\cup_J B_j]$ and $\max_J\mu_{ji}-\min_J\mu_{ji}\geq\delta-3hd^{1/2}$, $\lambda(C_\delta(y)\cap[\cup_J B_j])\geq(\delta-3hd^{1/2})^d$ and

$$\begin{aligned}\lambda(C_\delta(y)\setminus[\cup_J B_j]) &= \lambda(C_\delta(y))-\lambda(C_\delta(y)\cap[\cup_J B_j]) \\ &\leq \delta^d-(\delta-3hd^{1/2})^d\leq 3dhd^{1/2}\delta^{d-1},\end{aligned}$$

where the last inequality follows by induction. Therefore,

$$\begin{aligned}\int_{C_\delta(y)\setminus\cup_J B_j}\phi(\mu;y,\sigma)d\mu &\leq \lambda(C_\delta(y)\setminus[\cup_J B_j])\frac{1}{(2\pi)^{d/2}\sigma^d} \\ &\leq \frac{3d^{3/2}h\delta^{d-1}}{(2\pi)^{d/2}\sigma^d}.\end{aligned}$$ □

Lemma 2. *Let $C_\delta(y)$ be a d-dimensional cube with center y and side length $\delta>0$. Then,*

$$\int_{C_\delta(y)}\phi(\mu;y,\sigma)d\mu>1-\frac{8d\sigma}{(2\pi)^{1/2}\delta}.$$

Note that this inequality immediately implies that for any sub-cube of $C_\delta(y)$, $\tilde{C}$, with vertex at y and side length $\delta/2$, for example, $\tilde{C}=C_\delta(y)\cap[\mu\geq y]$,

$$\int_{\tilde{C}}\phi(\mu;y,\sigma)d\mu=\frac{1}{2^d}\int_{C_\delta(y)}\phi(\mu;y,\sigma)d\mu>\frac{1}{2^d}-\frac{8d\sigma}{2^d(2\pi)^{1/2}\delta}.$$

Proof.

$$\begin{aligned}\int_{C_\delta(y)}\phi(\mu;y,\sigma)d\mu &= \int_{\cap_{i=1}^d[|\mu_i|\leq\delta/2]}\phi(\mu;0,\sigma)d\mu \\ &= 1-\int_{\cup_{i=1}^d[|\mu_i|\geq\delta/2]}\phi(\mu;0,\sigma)d\mu \\ &\geq 1-\sum_{i=1}^d\int_{|\mu_i|\geq\delta/2}\phi(\mu_i;0,\sigma)d\mu_i \\ &= 1-2d\int_{\delta/2}^\infty\phi(\mu_1;0,\sigma)d\mu_1 \\ &> 1-\frac{2d}{(2\pi)^{1/2}\sigma} \\ &\quad \times\int_{\delta/2}^\infty\exp\left\{-\frac{0.5(\delta/2)\mu_1}{\sigma^2}\right\}d\mu_1 \\ &= 1-\frac{2d}{(2\pi)^{1/2}\sigma}\frac{-\sigma^2}{0.5(\delta/2)} \\ &\quad \times\exp\left\{-\frac{0.5(\delta/2)\mu_1}{\sigma^2}\right\}\Bigg|_{\delta/2}^\infty\end{aligned}$$

$$= 1 - \frac{8d\sigma}{(2\pi)^{1/2}\delta}\exp\left\{-\frac{0.5(\delta/2)^2}{\sigma^2}\right\}$$
$$> 1 - \frac{8d\sigma}{(2\pi)^{1/2}\delta}. \quad \square$$

Lemma 3. *$d_{L1}(f(y), p(y|\theta, \mathcal{M}_m))$ is uniformly continuous in θ.*

Proof.

$$|d_{L1}(f(y), p(y|\theta_1, \mathcal{M}_m)) - d_{L1}(f(y), p(y|\theta_2, \mathcal{M}_m))|$$
$$= \left|\int |f - p_1| - |f - p_2|\right| \le \int |p_1 - p_2|$$
$$= \int_{B_n[0]} |p_1 - p_2| + \int_{B_n[0]^c} p_1 + \int_{B_n[0]^c} p_2,$$

where $B_n[0]$ is a closed ball with center 0 and radius n. $\int_{B_n[0]} p(y|\theta, \mathcal{M}_m)\,dy$ is continuous by the DCT ($B_n[0]$ is bounded). Thus, $\int_{B_n[0]^c} p(y|\theta, \mathcal{M}_m)\,dy = 1 - \int_{B_n[0]} p(y|\theta, \mathcal{M}_m)\,dy$ is continuous in θ and $\forall\,\theta$ it is monotone in n ($\searrow 0$). By Dini's theorem it converges to 0 uniformly. Therefore, $\exists\, n$ such that $\forall\,\theta_1, \theta_2 \in \Theta_m$

$$\int_{B_n[0]^c} p_1 + \int_{B_n[0]^c} p_2 < \frac{\epsilon}{2}.$$

$p(y|\theta, \mathcal{M}_m)$ is continuous in y and θ and uniformly continuous on $B_n[0] \times \Theta_m$. Hence, given $\epsilon > 0\ \exists\delta > 0$ such that $\forall\, |\theta_1 - \theta_2| < \delta$, $\int_{B_n[0]} |p_1 - p_2| < \frac{\epsilon}{2}$. $\square$

A.2. Continuous and discrete data

The arguments used here are almost identical to the arguments above. Therefore, only the differences in assumptions and additional steps that are necessary to deal with discrete data will be provided.

Proof (Corollary 1, *Extension of* Theorem 1 *to Discrete Observables Case*)**.** The proof is identical to the one of Theorem 1. $\square$

Proof (Corollary 1, *Extension of* Theorem 2 *to Discrete Observables Case*)**.** First, let us show that

$$T^{-1}\log\ p(Y_T|\theta_m, \mathcal{M}_m) = \frac{1}{T}\sum_{t=1}^{T}\log\ p(y_t|\theta_m, \mathcal{M}_m)$$
$$\xrightarrow{a.s.} l(\theta_m; \mathcal{M}_m)$$

uniformly for all $\theta_m \in \Theta_m$. Note that

$$p(y_t|\theta_m, \mathcal{M}_m) = \sum_{j=1}^{m}\alpha_j \cdot \phi(y_{t,c}; \mu_{j,c}, H_{j,c}^{-1})$$
$$\times \int_{y_{t,-c}} \phi(y^*_{t,-c}|y_{t,c}; \mu_j, H_j^{-1})d(y^*_{t,-c})$$
$$\le \max_{j=1,\ldots,m} |H_{j,c}|^{1/2} \le \underline{\lambda}_m^{-0.5d}, \quad \forall\theta_m \in \Theta_m.$$

By construction, $y_{-c} = [a^1_{l_1}, b^1_{l_1}] \times \cdots \times [a^K_{l_K}, b^K_{l_K}]$, where $y_{-c} \subset R^K$. Define

$$\delta = \min\left\{\min_{k,l_k}\{|b^k_{l_k} - a^k_{l_k}|\}, 1\right\} \quad \text{and}$$
$$\gamma = \max\left\{\left|\max_{k,l_k}\{a^k_{l_k}\}\right|, \left|\min_{k,l_k}\{b^k_{l_k}\}\right|\right\}.$$

Note that δ is a finite number which is either the length of the shortest possible interval or 1, and γ is the closest point to 0 in the farthest away from 0 interval. Define $D_{y_{-c}} \subset y_{-c}$ as $D_{y_{-c}} \equiv [a^1_{l_1}, a^1_{l_1} + \delta] \times \cdots \times [a^K_{l_K}, a^K_{l_K} + \delta]$ if $a^k_{l_k} \neq -\infty$ for all $k \in \{1, \ldots, K\}$. If for some k, $a^k_{l_k} = -\infty$ use $[b^k_{l_k} - \delta, b^k_{l_k}]$ instead of $[a^k_{l_k}, a^k_{l_k} + \delta]$ in the definition of $D_{y_{-c}}$. Note that if $y^* \in D_{y_{-c}}$, then $\|y^*\| \le \sqrt{K}(\delta + \gamma)$. Then,

$$p(y_t|\theta_m, \mathcal{M}_m) = \sum_{j=1}^{m}\alpha_j \cdot \phi(y_{t,c}; \mu_{j,c}, H_{j,c}^{-1})$$
$$\times \int_{y_{t,-c}} \phi(y^*_{t,-c}|y_{t,c}; \mu_j, H_j^{-1})d(y^*_{t,-c})$$
$$\ge \sum_{j=1}^{m}\alpha_j \cdot \phi(y_{t,c}; \mu_{j,c}, H_{j,c}^{-1})$$
$$\times \int_{D_{y_{t,-c}}} \phi(y^*_{t,-c}|y_{t,c}; \mu_j, H_j^{-1})d(y^*_{t,-c}).$$

Define $\overline{y}^j_{t,-c}$ as

$$\overline{y}^j_{t,-c}(y_{t,c}) = \arg\min_{y^*_{t,-c} \in D_{y_{t,-c}}} \phi(y^*_{t,-c}|y_{t,c}; \mu_j, H_j^{-1}),$$

where by construction $\|\overline{y}^j_{t,-c}\| \le \sqrt{K}(\delta + \gamma)$. Define $\overline{y}^j_t = (y_{t,c}, \overline{y}^j_{t,-c})$. Note that

$$p(y_t|\theta_m, \mathcal{M}_m) \ge \sum_{j=1}^{m}\alpha_j\phi(y_{t,c}; \mu_{j,c}, H_{j,c}^{-1})$$
$$\times \int_{D_{y_{t,-c}}} \phi(\overline{y}_{t,-c}|y_{t,c}; \mu_j, H_j^{-1})d(y^*_{t,-c})$$
$$\ge \sum_{j=1}^{m}\alpha_j\phi(y_{t,c}; \mu_{j,c}, H_{j,c}^{-1})\phi(\overline{y}^j_{t,-c}|y_{t,c}; \mu_j, H_j^{-1})\delta^K$$
$$\ge \sum_{j=1}^{m}\alpha_j\phi(\overline{y}^j_t; \mu_j, H_j^{-1})\delta^K.$$

Then,

$$\log p(y_t|\theta_m, \mathcal{M}_m) \ge \sum_{j=1}^{m}\alpha_j\left[K\log\delta + \log(2\pi)^{-(d+K)/2}\right.$$
$$\left. + \frac{1}{2}\log\ |H_j| - 0.5(\overline{y}^j - \mu_j)'H_j(\overline{y}^j - \mu_j)\right]$$
$$\ge K\log\delta + \log(2\pi)^{-(d+K)/2} + 0.5(d+K)\log\underline{\lambda}$$
$$- 0.5\max_j[\overline{y}^{j'}H_j\overline{y}^j - 2\overline{y}^{j'}H_j\mu_j + \mu_j'H_j\mu_j]$$
$$\ge K\log\delta + \log(2\pi)^{-(d+K)/2} + 0.5(d+K)\log\underline{\lambda}$$
$$- 0.5\overline{\lambda}\max_j\overline{y}^{j\prime}\overline{y}^j - \max_j\|\overline{y}^j\| \cdot \|H_j\mu_j\|$$
$$- 0.5\max_j\|\mu_j'H_j\mu_j\|$$
$$\ge K\log\delta + \log(2\pi)^{-(d+K)/2} + 0.5(d+K)\log\underline{\lambda}$$
$$- 0.5\overline{\lambda}(y_c'y_c + K(\gamma+\delta)^2)$$
$$- (\|y_c\| + \sqrt{K}(\gamma+\delta))\max_j\|H_j\mu_j\|$$
$$- 0.5\max_j\|\mu_j'H_j\mu_j\|.$$

Since eigenvalues of H_j are bounded above and away from zero and since $\|H_j\|$ and $\|\mu_j\|$ are bounded on Θ_m,

$$|\log\ p(y_t|\theta_m, \mathcal{M}_m)| \le q(y_t),$$

where $q(y_t)$ is integrable because $p(y_c|D)$ has finite second moments by the theorem assumptions. Also, $\log p(y|\theta_m)$ is

continuous in θ and measurable in y. Thus, by Theorem 2 in Jennrich (1969), we get uniform a.s. convergence. $l(\theta; \mathcal{M}_m)$ is continuous by the dominated convergence theorem. The rest of the argument is the same as in Theorem 2. □

Proof (Corollary 2). Let $A^m_{j,i} = A^m_j \times A_i$, where A^m_j is an element of partition of Y_c, $A_i = [a^1_{i_1}, b^1_{i_1}] \times \cdots \times [a^K_{i_K}, b^K_{i_K}]$ is an element of the partition of the space for the latent variables, R^K, defined by possible values of the discrete observables, and

$$i = (i_1, \ldots, i_K) \in I \equiv \{(l_1, \ldots, l_K) : l_k \in \{1, \ldots, N_k\}, k \in \{1, \ldots, K\}\}.$$

As m increases the Lebesgue measure of $A^m_j, j > 0$, decreases and the fine part of the partition $A^m_1, \ldots, A^m_m$ covers larger and larger part of Y_c, where $Y_c \subset R^d$. Parameter values for approximating $F(y) = F(y_c, y_{-c})$ by $\mathcal{M}_m$ are defined by

$$\begin{aligned} p(y|\theta_m, \mathcal{M}_m) = &\sum_{i\in I}\sum_{j=1}^m F(A^m_{j,i}) \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i}, \sigma_m) \\ &\times d(y^*_{-c}) \cdot \phi(y_c; \mu^m_j, \sigma_m) \\ &+ F(A^m_{0,i}) \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i}, \sigma_m) \\ &\times d(y^*_{-c}) \cdot \phi(y_c; 0, \sigma_0), \end{aligned} \tag{27}$$

where σ_0 is fixed, σ_m converges to zero as m increases, $\mu^m_{j,i} \in A^m_{j,i}$, $\mu^m_{j,i} = [\mu^{m\prime}_j, \mu_i']'$, μ^m_j is the center of A^m_j and μ_i is the center of A_i. Since d_{KL} is always non-negative,

$$\begin{aligned} 0 &\le \int \log \frac{f(y)}{p(y|\theta_m, \mathcal{M}_m)} F(dy) \\ &\le \int \log \max\left\{1, \frac{f(y)}{p(y|\theta_m, \mathcal{M}_m)}\right\} F(dy). \end{aligned}$$

We will use dominated convergence theorem (DCT) to show that the last integral in the inequality above converges to zero as m increases.

First, we will show the point-wise convergence of the integrand to zero a.s. F. For a fixed $y = (y_c, y_{-c})$ define a cube $C_{\delta_m}(y_c) \subset R^d$ with a center y_c and side length $\delta_m > 0$. Then,

$$\begin{aligned} p(y|\theta_m, \mathcal{M}_m) = &\sum_{i\in I}\sum_{j=1}^m F(A^m_j|A_i)F(A_i) \\ &\times \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i}, \sigma_m) d(y^*_{-c}) \\ &\times \phi(y_c; \mu^m_j, \sigma_m) + F(A^m_0|A_i)F(A_i) \\ &\times \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i}, \sigma_m) d(y^*_{-c}) \\ &\times \phi(y_c; 0, \sigma_0) \\ \ge &\sum_{j=1}^m F(A^m_j|A_{i^*})F(A_{i^*}) \\ &\times \int_{y_{-c}} \phi(y^*_{-c}; \mu_{i^*}, \sigma_m) d(y^*_{-c}) \\ &\times \phi(y_c; \mu^m_j, \sigma_m), \end{aligned}$$

where i^* is such that $y_{-c} = A_{i^*}$ and therefore $F(A_{i^*}) = f(y_{-c})$. Note that $F(\cdot)$ and $f(\cdot)$ are used interchangeably for discrete components. Furthermore, μ_{i^*} is an interior point of y_{-c} by construction. Given $\epsilon > 0$ since $\sigma_m \to 0$, $\exists M_0$ such that $\forall m \ge M_0$,

$$\begin{aligned} \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i^*}, \sigma_m) d(y^*_{-c}) &= \int_{y_{-c}} \phi(y^*_{-c}; \mu_{i^*}, \sigma_m) d(y^*_{-c}) \\ &> (1-\epsilon). \end{aligned}$$

Therefore,

$$\begin{aligned} p(y|\theta_m, \mathcal{M}_m) &\ge (1-\epsilon)F(y_{-c}) \left(\sum_{j=1}^m F(A^m_j|y_{-c})\phi(y_c; \mu^m_j, \sigma_m)\right) \\ &\ge (1-\epsilon)F(y_{-c}) \inf_{z\in C_{\delta_m}(y_c)} f(z|y_{-c}) \\ &\quad \times \sum_{j: A^m_j \subset C_{\delta_m}(y_c)} \lambda(A^m_j)\phi(y_c; \mu^m_j, \sigma_m), \end{aligned} \tag{28}$$

where λ is the Lebesgue measure.

As long as δ_m is bounded above it is always possible to construct a partition, in which elements $A^m_1, \ldots, A^m_m$ are adjacent cubes with side length h_m ($\lambda(A^m_j) = h^d_m$ for $j > 0$) and for any y_c there exists M_1 such that

$$\forall m \ge M_1, \quad C_{\delta_m}(y_c) \cap A^m_0 = \emptyset. \tag{29}$$

In Lemmas 1 and 2 below, the following bounds for the Riemann sum in (20) are derived (the Riemann sum is not far from the corresponding normal integral and the integral is not far from 1)

$$\begin{aligned} &\sum_{j: A^m_j \subset C_{\delta_m}(y_c)} \lambda(A^m_j)\phi(y_c; \mu^m_j, \sigma_m) \\ &\ge 1 - \frac{3d\delta^{d-1}_m h_m}{(2\pi)^{d/2}\sigma^d_m} - \frac{8d\sigma_m}{(2\pi)^{1/2}\delta_m}. \end{aligned} \tag{30}$$

Let δ_m, σ_m, h_m satisfy the following

$$\delta_m \to 0, \qquad \sigma_m/\delta_m \to 0, \qquad h_m/\sigma^d_m \to 0. \tag{31}$$

Given $\epsilon > 0$ there exists M_2 such that for $m \ge M_2$ expressions in (30) are bounded below by $(1-\epsilon)$.

By assumption of the corollary $f(y_c|y_{-c})$ is continuous in y_c on Y_c a.s. F. Then, for any y_c and y_{-c} satisfying $f(y_c|y_{-c}) \cdot f(y_{-c}) > 0$ there exists M_3 such that for $m \ge M_3$, $[f(y_c|y_{-c})/\inf_{z\in C_{\delta_m}(y_c)} f(z|y_{-c})] \le (1+\epsilon)$ since $\delta_m \to 0$.

For any $m \ge \max\{M_0, M_1, M_2, M_3\}$,

$$\begin{aligned} 1 &\le \max\left\{1, \frac{f(y)}{p(y|\theta_m, \mathcal{M}_m)}\right\} \\ &\le \max\left\{1, \frac{f(y_c|y_{-c})f(y_{-c})}{f(y_{-c}) \inf_{z\in C_{\delta_m}(y_c)} f(z|y_{-c})(1-\epsilon)^2}\right\} \\ &\le \max\left\{1, \frac{f(y_c|y_{-c})}{\inf_{z\in C_{\delta_m}(y_c)} f(z|y_{-c})(1-\epsilon)^2}\right\} \le \frac{1+\epsilon}{(1-\epsilon)^2}. \end{aligned}$$

Thus, $\log\max\{1, f(y)/p(y|\theta^m)\} \to 0$ for any y satisfying $f(y) > 0$, which implies convergence a.s. F.

Second, we will establish an integrable upper bound on $\log\max\{1, f(y)/p(y|\theta^m)\}$.

$$\begin{aligned} p(y|\theta_m, \mathcal{M}_m) = &\sum_{i\in I}\sum_{j=1}^m F(A^m_j|A_i)F(A_i) \\ &\times \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i}, \sigma_m) d(y^*_{-c}) \\ &\times \phi(y_c; \mu^m_j, \sigma_m) + F(A^m_0|A_i)F(A_i) \\ &\times \int_{y_{-c}} \phi(y^*_{-c}|y_c; \mu^m_{j,i}, \sigma_m) d(y^*_{-c}) \\ &\times \phi(y_c; 0, \sigma_0) \\ \ge &\sum_{j=1}^m F(A^m_j|A_{i^*})F(A_{i^*}) \end{aligned}$$

$$\times \int_{y_{-c}} \phi(y^*_{-c};\mu_{i^*},\sigma_m)d(y^*_{-c})$$
$$\times \phi(y_c;\mu_j^m,\sigma_m) + F(A_0^m|A_{i^*})F(A_{i^*})$$
$$\times \int_{y_{-c}} \phi(y^*_{-c};\mu_{i^*},\sigma_m)d(y^*_{-c})$$
$$\times \phi(y_c;0,\sigma_0).$$

Since $\sigma_m \to \infty$ there exists M_4 such that $\forall m > M_4$

$$\int_{y_{-c}} \phi(y^*_{-c};\mu_{i^*},\sigma_m)d(y^*_{-c}) > 0.5.$$

Then,

$$\begin{aligned} p(y|\theta_m,\mathcal{M}_m) &\geq f(y_{-c})\cdot 0.5\left(\sum_{j=1}^{m} F(A_j^m|y_{-c})\phi(y_c;\mu_j^m,\sigma_m)\right. \\ &\quad \left. +\ F(A_0^m|y_{-c})\phi(y_c,0,\sigma_0)\right) \\ &\geq [1-1_{A_0^m}(y_c)]\cdot f(y_{-c})\cdot 0.5 \\ &\quad \times \inf_{z\in C_1(r,y_c,y_{-c})} f(z|y_{-c}) \\ &\quad \times \sum_{j:A_j^m\subset C_1(r,y_c,y_{-c})} \lambda(A_j^m)\phi(y_c;\mu_j^m,\sigma_m) \\ &\quad + 1_{A_0^m}(y_c)\cdot f(y_{-c})\cdot 0.5\cdot \inf_{z\in C_0(r,y_c,y_{-c})} f(z|y_{-c}) \\ &\quad \times \lambda(C_0(r,y_c,y_{-c}))\phi(y_c;0,\sigma_0). \end{aligned} \tag{32}$$

Lemmas 1 and 2 imply that the Riemann sum in (32) is bounded below by $2^{-d}-2^{-(d+1)}=2^{-(d+1)}$ for any m larger than some M_5. Parameter σ_0 can be chosen so that

$$1 > 2^{-(d+1)} > \phi(y_c;0,\sigma_0)\lambda(C_0(r,y)). \tag{33}$$

This implies

$$\begin{aligned} &\log\max\left\{1,\frac{f(y)}{p(y|\theta_m,\mathcal{M}_m)}\right\} \\ &\leq \log\max\left\{1,\frac{f(y_c|y_{-c})f(y_{-c})}{f(y_{-c})0.5\inf_{z\in C(r,y_c,y_{-c})} f(z|y_{-c})\cdot\phi(y_c;0,\sigma_0)\cdot(r/2)^d}\right\} \\ &\leq \log\frac{1}{0.5\phi(y_c;0,\sigma_0)(r/2)^d} \\ &\quad \times\max\left\{0.5\phi(y_c;0,\sigma_0)(r/2)^d,\frac{f(y_c|y_{-c})}{\inf_{z\in C(r,y_c,y_{-c})} f(z|y_{-c})}\right\} \\ &\leq -\log(0.5\phi(y;0,\sigma_0)(r/2)^d)+\log\frac{f(y_c|y_{-c})}{\inf_{z\in C(r,y_c,y_{-c})} f(z|y_{-c})}. \end{aligned} \tag{34}$$

Inequality (34) follows by (33). The first expression in (34) is integrable by corollary assumption 2. The second expression in (34) is integrable by corollary assumption 3.

Since we have established pointwise convergence and integrable upper bound, we can apply the DCT. Henceforth, given $\epsilon > 0 \exists M$ such that for any $m > M$ and θ_m defined in (27), $d_{KL}(f(\cdot),p(\cdot|\theta,\mathcal{M}_m)) \leq \epsilon$. □

References

Abrevaya, J., 2001. The effects of demographics and maternal behavior on the distribution of birth outcomes. Empirical Economics 26, 247–257.

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

Chung, Y., Dunson, D.B., 2009. Nonparametric Bayes conditional distribution modeling with variable selection. Journal of the American Statistical Association 104, 1646–1660.

De Iorio, M., Muller, P., Rosner, G.L., MacEachern, S.N., 2004. An ANOVA model for dependent random measures. Journal of the American Statistical Association 99, 205–215.

Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society. Series B (Methodological) 56, 363–375.

Dunson, D.B., Park, J.-H., 2008. Kernel stick-breaking processes. Biometrika 95, 307–323.

Efromovich, S., 2007. Conditional density estimation in a regression setting. The Annals of Statistics 35, 2504–2535.

Gelfand, A., Dey, D., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based methods. In: Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics, vol. 4. Oxford University Press, pp. 147–168.

Genovese, C.R., Wasserman, L., 2000. Rates of convergence for the Gaussian mixture sieve. The Annals of Statistics 28, 1105–1127.

Gerfin, M., 1996. Parametric and semi-parametric estimation of the binary response model of labour market participation. Journal of Applied Econometrics 11, 321–339.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. Wiley-Interscience.

Geweke, J., 2004. Getting it right: joint distribution tests of posterior simulators. Journal of the American Statistical Association 99, 799–804.

Geweke, J., 2007. Interpretation and inference in mixture models: simple MCMC works. Computational Statistics and Data Analysis 51, 3529–3550.

Geweke, J., Keane, M., 2007. Smoothly mixing regressions. Journal of Econometrics 138, 252–290.

Ghosal, S., van der Vaart, A., 2007. Posterior convergence rates of Dirichlet mixtures at smooth densities. The Annals of Statistics 35, 697–723.

Ghosh, J., Ramamoorthi, R., 2003. Bayesian Nonparametrics, first ed.. Springer.

Griffin, J.E., Steel, M.F.J., 2006. Order-based dependent Dirichlet processes. Journal of the American Statistical Association 101, 179–194.

Hall, P., Racine, J., Li, Q., 2004. Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association 99, 1015–1026.

Hayfield, T., Racine, J.S., 2008. Nonparametric econometrics: the np package. Journal of Statistical Software 27, 1–32.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixtures of local experts. Neural Computation 3, 79–87.

Jennrich, R.I., 1969. Asymptotic properties of non-linear least squares estimators. The Annals of Mathematical Statistics 40, 633–643.

Jordan, M., Xu, L., 1995. Convergence results for the EM approach to mixtures of experts architectures. Neural Networks 8, 1409–1431.

Koenker, R., Hallock, K.F., 2001. Quantile regression. The Journal of Economic Perspectives 15, 143–156.

Leslie, D.S., Kohn, R., Nott, D.J., 2007. A general approach to heteroscedastic linear regression. Statistics and Computing 17, 131–146.

Li, Q., Racine, J.S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press.

Li, Q., Racine, J.S., 2008. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. Journal of Business and Economic Statistics 26, 423–434.

MacEachern, S.N., 1999. Dependent nonparametric processes. ASA Proceedings of the Section on Bayesian Statistical Science.

Marin, J.M., Robert, C., 2008. Approximating the marginal likelihood in mixture models. Bulletin of the Indian Chapter of ISBA 1, 2–7.

Muller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83, 67–79.

Norets, A., 2010. Approximation of conditional densities by smooth mixtures of regressions. The Annals of Statistics 38, 1733–1766.

Norets, A., Pelenis, J., 2011. Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures, Unpublished manuscript, Princeton University.

Norets, A., Pelenis, J., 2009. Web appendix for Bayesian modeling of joint and conditional distributions, Unpublished manuscript, Princeton University.

Pati, D., Dunson, D., Tokdar, S., 2011. Posterior consistency in conditional distribution estimation, Duke University, Dept. of Statistics.

Peng, F., Jacobs, R.A., Tanner, M.A., 1996. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. Journal of the American Statistical Association 91, 953–960.

Roeder, K., Wasserman, L., 1997. Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association 92, 894–902.

Schwartz, L., 1965. On Bayes procedures. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 4, 10–26.

Taddy, M.A., Kottas, A., 2010. A Bayesian nonparametric approach to inference for quantile regression. Journal of Business and Economic Statistics 28, 357–369.

Villani, M., Kohn, R., Giordani, P., 2009. Regression density estimation using smooth adaptive Gaussian mixtures. Journal of Econometrics 153, 155–173.

Wood, S., Jiang, W., Tanner, M., 2002. Bayesian mixture of splines for spatially adaptive nonparametric regression. Biometrika 89, 513–528.

Wu, Y., Ghosal, S., 2010. The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimations. Journal of Multivariate Analysis 101, 2411–2419.