ADAPTIVE BAYESIAN ESTIMATION OF CONDITIONAL DISCRETE-CONTINUOUS DISTRIBUTIONS WITH AN APPLICATION TO STOCK MARKET TRADING ACTIVITY*

By Andriy Norets† and Justinas Pelenis‡

Brown University and Vienna Institute for Advanced Studies

We consider Bayesian nonparametric estimation of conditional discrete-continuous distributions. Our model is based on a mixture of normal distributions with covariate dependent mixing probabilities. We use continuous latent variables for modeling the discrete part of the distribution. The marginal distribution of covariates is not modeled. Under anisotropic smoothness conditions on the data generating conditional distribution and a possibly increasing number of the support points for the discrete part of the distribution, we show that the posterior in our model contracts at frequentist adaptive optimal rates up to a log factor. Our results also imply an upper bound on the posterior contraction rate for predictive distributions when the data follow an ergodic Markov process and our model is used for modeling the Markov transition distribution. The proposed model performs well in an application to stock market trading activity.

*The first author gratefully acknowledges the support from the NSF Award SES-1851796.

†Corresponding Author, Professor, Department of Economics, Brown University

‡ Fellow, Vienna Institute for Advanced Studies

Keywords and phrases: Bayesian nonparametrics, adaptive rates, posterior contraction, conditional density, mixtures of normal distributions, smoothly mixing regressions, mixtures of experts.

1. Introduction. In this paper, we propose a Bayesian nonparametric model for estimation of conditional discrete-continuous distributions. We show that the model has outstanding asymptotic properties and compares favorably to standard parametric and nonparametric alternatives in an application to forecasting of stock trade counts. More generally, we provide a practical and optimal adaptive nonparametric alternative to workhorse econometric parametric models such as probit, ordered probit and Poisson regression.

Nonparametric modeling of conditional distributions is especially important in the Bayesian framework. Conditional distributions can fully describe dependence of one set of variables on another. However, even if the main object of interest is not the whole conditional distribution but a conditional mean or quantiles, a Bayesian econometrician has to specify at least a conditional distribution in order to define a likelihood. The use of nonparametric or very flexible models ameliorates the risk of invalid inference due to misspecification.

The theory and practical implementation of Bayesian nonparametric methods for continuous data are very well developed at this point, see Ghosal and van der Vaart (2017) for aets thorough exposition of theoretical developments and Dey, Muller, and Sinha (1998), Chamberlain and Hirano (1999), Burda, Harding, and Hausman (2008), Chib and Greenberg (2010), and Jensen and Maheu (2014) among many others for applications. In most applications in economics, the data contain both continuous and discrete variables. Nonparametric methods for conditional discrete-continuous distributions and their theoretical properties are less understood and developed.

Starting from Aitchison and Aitken (1976), researchers observed that smoothing discrete data in nonparametric estimation improves estimation results. Hall and Titterington (1987) provided a theoretical justification for improvements resulting from smoothing in estimation of a univariate discrete distribution with a support that can increase with the sample size. Norets and Pelenis (2021) extended these results to estimation of joint multivariate discrete-continuous distributions. In their framework, discrete variables have the support that can become finer with the sample size; the data generating joint distribution can be smooth to a different degree (and not smooth at all) with respect to different discrete and continuous variables. They derived optimal estimation rates for these settings and show that smoothing is beneficial only for a subset of discrete variables with a quickly growing number of support points and/or a high level of smoothness. They also show that a Bayesian nonparametric model based on latent variables and mixtures of multivariate normal distributions has posterior contraction rates that are no larger than the derived optimal estimation rates with an additional log factor. In the present paper, we adopt a similar asymptotic framework and apply it to estimation of conditional discrete-continuous distributions. Simply extracting conditional distributions from optimally estimated joint distributions does not in general result in the optimal estimation of conditional distributions since the joint and conditional distributions can have different smoothness and other properties. Therefore, in the present paper, we model the conditional distributions directly.

There are additional important reasons for constructing nonparametric priors for conditional distributions directly. First, in regression settings, a ubiquitous problem of covariate selection can be conveniently addressed by standard means (special priors and Bayesian model selection and comparison). Second, nonparametric priors for conditional distributions can also be used for modeling of Markov transition probabilities, and, thus for nonparametric modeling of Markovian time series. Such nonparametric time series models have a wide range of applications in empirical macroeconomics and, especially, in empirical finance with its abundance of large datasets.

Our nonparametric model for conditional discrete-continuous distributions is based on a mixture of normal distributions with covariate dependent mixing weights and a variable number of mixture components. It is closely related to mixture-of-experts or smoothly mixing regressions (Jacobs, Jordan, Nowlan, and Hinton (1991), Jordan and Xu (1995), Peng, Jacobs, and Tanner (1996), Wood, Jiang, and Tanner (2002), Geweke and Keane (2007), Villani et al. (2009), Norets (2010), Norets and Pelenis (2014), Norets and Pati (2017)). Discrete dependent variables in our model are represented by continuous latent variables, which jointly with continuous dependent variables are modeled by the mixture of multivariate normals. The covariate dependent mixing weights are proportional to a normal density and an integral of a normal density for continuous and discrete covariates correspondingly. The model can be thought of as a generalization of a covariate dependent mixture model for continuous data from Norets and Pati (2017) to mixed discrete-continuous data. Posterior simulation for our covariate dependent mixture with a variable number of components is performed by a reversible jump algorithm from Norets (2021).

There are potentially many different ways of handling discrete variables, especially covariates, in a covariate dependent mixture model. The main practical contribution of our paper is to develop a model specification that has optimal asymptotic properties. Specifically, we show that the posterior contraction rates in our model are equal (up to a log factor) to the optimal estimation rates. In our framework, it means that the model optimally takes advantage of smoothness in the data generating conditional distribution in both continuous and discrete variables. If the data generating conditional distribution is not sufficiently smooth or does not have a sufficiently fine support for some discrete variables, then the resulting posterior contraction rate corresponds to the standard estimation rate for (the smoothness and dimension of) the continuous and the rest of the discrete variables. The derived posterior contraction rates are adaptive as the prior distribution does not depend on the smoothness and the support of the data generating process. Our results for conditional distributions also imply the same convergence rates for predictive distributions when our prior is used for nonparametric modeling of Markov transition distributions for ergodic Markovian time series. To the best of our knowledge, such asymptotic guarantees for estimation of conditional discrete-continuous distributions are not currently available for any other Bayesian model or a frequentist nonparametric estimator.

We evaluate the practical performance of our model in an out-of-sample forecasting exercise for stock trades count data and two additional applications to cross-sectional data. The model compares favorably with a parametric Poisson regression and a nonparametric discrete-continuous conditional density estimator based on discrete and continuous kernels with a cross-validation procedure for bandwidth selection (Li and Racine (2008)).

Let us briefly review additional related references in the literature. Our posterior contraction results are derived from general sufficient conditions for posterior contraction introduced by Ghosal et al. (2000). Optimal adaptive posterior contraction rates for joint densities were obtained in Scricciolo (2006), Rousseau (2010), Kruijer et al. (2010), Shen, Tokdar, and Ghosal (2013) among others. Shen and Ghosal (2016) and Norets and Pati (2017) obtained optimal adaptive posterior contraction rates for nonparametric conditional density models for continuous observations. Norets and Pelenis (2012), DeYoreo and Kottas (2017), and Canale and Dunson (2015) derived posterior consistency and non-optimal bounds on posterior contraction rates for nonparametric models of joint discrete-continuous distributions in asymptotic settings without smoothness for discrete variables. Albert and Chib (1993) and McCulloch and Rossi (1994) pioneered the use of continuous latent variables for handling discrete observations in Markov chain Monte Carlo algorithms for parametric limited dependent variable models. In frequentist framework, nonparametric estimation of discrete distributions with and without smoothness assumptions was considered in Aitchison and Aitken (1976), Hall and Titterington (1987), Burman (1987), Dong and Simonoff (1995), Aerts et al. (1997), and Efromovich (2011) among many others.

The rest of the paper is organized as follows. Section 2 describes the data generating process and the asymptotic framework. The model and main posterior concentration results are presented in Section 3. Sections 4 and 5 evaluate the model performance in out-of-sample forecasting exercises. Technical assumptions, intermediate results, and proofs are given in Sections 6 and 7. Additional proof details are delegated to the Appendix.

2. Data Generating Process. Let us denote the response space by $\mathcal{Y} \times \mathcal{X}$ and the covariate space by $\mathcal{Z} \times \mathcal{W}$. The continuous part of observations is denoted by $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$ and $w \in \mathcal{W} \subset \mathbb{R}^{d_w}$ and the discrete part by $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$ and $z = (z_{d_y+1}, \ldots, z_{d_y+d_z}) \in \mathcal{Z}$, where

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

are grids on $[0,1]^{d_y}$ and $[0,1]^{d_z}$ (a product symbol Π applied to sets hereafter denotes a Cartesian product). The number of values that the discrete coordinates y_j or z_j can take, N_j, can potentially grow with the sample size or stay constant.

For $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$ and $z = (z_{d_y+1}, \ldots, z_{d_y+d_z}) \in \mathcal{Z}$, let $A_y = \prod_{j=1}^{d_y} A_{y_j}$ and $A_z = \prod_{j=d_y+1}^{d_y+d_z} A_{z_j}$, where

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

and A_{z_j} is defined analogously.

Let us represent the data generating density-probability mass function as an integral of a density over latent variables

$$p_0(y, x, z, w) = \int_{A_y} \int_{A_z} f_0(\tilde{y}, x|\tilde{z}, w) g_0(\tilde{z}, w) d\tilde{y} d\tilde{z}, \tag{2.1}$$

where f_0 is a conditional probability density function on $\mathbb{R}^{d_x+d_y+d_z+d_w}$ and g_0 is a probability density function on $\mathbb{R}^{d_z+d_w}$ with respect to the Lebesgue measure, and the discrete part of the observation (y, z) is mapped into the latent variables $(\tilde{y}, \tilde{z}) \in A_y \times A_z$. The representation of a mixed discrete-continuous distribution in (2.1) is so far without a loss of generality since for any given p_0 one could always define f_0 and g_0 using a mixture of densities with non-overlapping supports included in $A_y \times A_z$, $(y,z) \in \mathcal{Y} \times \mathcal{Z}$.

Suppose that $(Y^n, X^n, Z^n, W^n) = (Y_1, X_1, Z_1, W_1, \ldots, Y_n, X_n, Z_n, W_n)$ is a random sample from the joint density $p_0(y, x|z, w)p_0(z, w)$. Let P_0 and P_0^n represent the probability measures corresponding to p_0 and its product p_0^n. When N_j's grow with the sample size then it is possible that the generality of the representation in (2.1) can be diminished if one imposed some assumption on $f_0(\cdot|\cdot)g_0(\cdot)$ such as smoothness. Nonetheless, in what follows we do allow for smoothness in f_0 to formalize the scenarios where for ordered discrete variables borrowing of information from nearby discrete points can be beneficial in estimation.

To get more refined results, we allow for anisotropic smoothness, which means that smoothness can vary across coordinate j, and we consider the possibility of N_j's growing at different rates for different j's. Let $\mathbb{Z}_+$ denote the set of non-negative integers. For smoothness coefficients $\beta_i > 0$, $i = 1, \ldots, d$, $d = d_x + d_y + d_z + d_w$, and an envelope function $L : \mathbb{R}^{2d} \to \mathbb{R}$, an anisotropic $(\beta_1, \ldots, \beta_d)$-Holder class, $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, is defined as follows.

Definition 2.1. *$f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ if for any $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, $\sum_{i=1}^d k_i/\beta_i < 1$, mixed partial derivative of order k, $D^k f$, is finite and*

$$|D^k f(z + \Delta z) - D^k f(z)| \leq L(z, \Delta z) \sum_{j=1}^{d} |\Delta z_j|^{\beta_j (1 - \sum_{i=1}^d k_i/\beta_i)}, \tag{2.2}$$

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

This definition of the Holder class has been proposed in Norets and Pelenis (2021) and its similarities and slight differences with other Holder smoothness definitions are discussed in that paper. It allows for functions that can be differentiated with respect to different coordinates different number of times. If all β_j's are the same, then the definition reduces to standard Holder smoothness.

Let $\mathcal{A}$ denote a collection of all subsets of indices for discrete coordinates $\{1, \ldots, d_y, d_y + 1, \ldots, d_y + d_z\}$. For $J \in \mathcal{A}$, define $J^c = \{1, \ldots, d\} \setminus J$,

$$N_J = \prod_{i \in J} N_i, \qquad \beta_{J^c} = \left[\sum_{i \in J^c} \beta_i^{-1}\right]^{-1},$$

$N_\emptyset = 1$, $\beta_\emptyset = \infty$, and $\beta_\emptyset/(2\beta_\emptyset + 1) = 1/2$.

Norets and Pelenis (2021) show that for joint distributions with underlying densities for continuous and latent variables variables that belong to the anisotropic Holder class $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, lower bounds on estimation rates in total variation distance are given by

$$\min_{J\in\mathcal{A}}\left[\frac{N_J}{n}\right]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}} \tag{2.3}$$

(no estimator can have a faster rate of convergence for this class of data generating processes). They also show that in a model based on a mixture of normal distributions for the underlying density, posterior contraction rates are equal (up to a log factor) to the lower bounds, and thus are optimal up to a log factor. Since the distance between joint distributions can be bounded by the sum of the distances between the corresponding conditional and marginal distributions (by the triangle inequality), (2.3) also provides a lower bound on the estimation rates for conditional distributions with underlying conditional densities in $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$. Expression $[N_J/n]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$ in (2.3) is the standard estimation rate for a $card(J^c)$-dimensional density with anisotropic smoothness coefficients $\{\beta_j,\, j\in J^c\}$ and the sample size n/N_J (Ibragimov and Hasminskii (1984)). One interpretation of this expression is that smoothing is performed only over coordinates in J^c and the coordinates in J are treated as discrete. The minimum over J in (2.3) suggests that an estimator that achieves this lower bound rate needs in a sense to optimally choose a subset of discrete coordinates over which smoothing is beneficial.

3. Model and Main Results on Posterior Concentration.

We propose the following model for conditional discrete-continuous distributions

$$p(y,x|z,w;\theta,m)=\frac{\int_{A_y\times A_z} f(\tilde{y},\tilde{z},x,w|\theta,m)d\tilde{y}d\tilde{z}}{\int_{A_z}\left[\int f(\tilde{y},\tilde{z},x,w|\theta,m)d\tilde{y}dx\right]d\tilde{z}}, \tag{3.1}$$

where

$$f(\tilde{y},\tilde{z},x,w|\theta,m)=\sum_{j=1}^{m}\alpha_j\phi(\tilde{y},\tilde{z},x,w;\mu_j,\sigma) \tag{3.2}$$

is a mixture of multivariate normal distributions with a variable number of components m and parameters collected in $\theta=(\sigma,\mu_j,\alpha_j,j=1,2,\ldots)$. The multivariate normal distributions in the mixture, $\phi(\cdot;\mu_j,\sigma)$, have a diagonal variance matrix with the square roots of diagonal elements contained in $\sigma\in\mathbb{R}^d_+$. Thus, this conditional density-probability mass function can be expressed explicitly through standard univariate normal densities and cumulative distribution functions. This model can be thought of as a generalization of a covariate dependent mixture model for continuous data from Norets and Pati (2017) to mixed discrete-continuous data.

Under standard assumptions on the priors for (θ,m) and some additional technical conditions on the data generating process presented in Section 6, the posterior contraction rate for this model is equal up to a log factor to the lower bound on estimation rate given in (2.3).

THEOREM 3.1. *Suppose the assumptions from Sections 6.1 and 6.2 hold for every $J \in \mathcal{A}$. Let*

$$\epsilon_n = \min_{J \in \mathcal{A}} \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{3.3}$$

where $t_J > 0$ is defined in Section 7. Suppose also $n\epsilon_n^2 \to \infty$. Then, there exists a constant $\bar{M} > 0$ such that

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

where d_{TV} denotes the total variation distance between conditional distributions integrated over the data generating distributions of covariates.

The proof of the theorem verifies the sufficient conditions for the posterior contraction from Ghosal et al. (2000). It is conceptually similar to the proof of related results for continuous data in Norets and Pati (2017). In order to show that Kullback-Leibler neighborhoods of the data generating distribution have sufficient prior probability, which is one of the main sufficient conditions, Norets and Pati (2017) bound a distance between conditional distributions by a distance between the appropriate joint distributions and then exploit approximation results for mixtures of multivariate normal distributions from Shen et al. (2013). Similarly, here we also bound a distance between conditional distributions by a distance between the appropriate joint distributions and then exploit approximation results from Norets and Pelenis (2021). The actual proof contains new additional arguments handling the discrete variables in the conditioning set; it is rather long, and we present it in Section 7 and the Appendix.

Our results on the bounds for the prior probabilities of Kullback-Leibler neighborhoods imply that ϵ_n defined in the theorem is also a posterior contraction rate for predictive distributions when our model and prior are used for nonparametric modeling of the Markov transition distributions for Markovian time series. We provide details in Section 7.2.

4. Application to Trade Counts.

4.1. *Model Specification and Forecasting Performance.* In this section, we compare forecasting performance of our Bayesian nonparametric model for conditional discrete-continuous distribution with a parametric Poisson regression and a classical nonparametric discrete-continuous conditional density estimator from Li and Racine (2008) who use discrete and continuous kernels with a cross-validation procedure for bandwidth selection.

We use the following version of our model

$$p(y|w,\theta,m) = \int_{A_y} \sum_{j=1}^{m} \frac{\alpha_j \exp\{-0.5 \sum_{k=1}^{d_w}(w_k - \mu_{jk}^w)^2/(\sigma_k^w s_{jk}^w)^2\}}{\sum_{i=1}^{m} \alpha_i \exp\{-0.5 \sum_{k=1}^{d_w}(w_k - \mu_{ik}^w)^2/(\sigma_k^w s_{ik}^w)^2\}} \phi_{w'\beta_j, \sigma^y s_j^y}(\tilde{y}) d\tilde{y}, \quad (4.1)$$

where discrete y is one-dimensional and $w \in \mathbb{R}^{d_w}$. The location parameters for $\tilde{y}$ have a specification linear in covariates, $w'\beta_j$, and the scale parameters can differ across the mixture components but also have a common factor. Such richer specifications for mixture components lead to better finite sample performance (Villani et al. (2009)). The asymptotic results are not affected by the presence of linear coefficients β_j and component specific scales (s_{jk}^w, s_j^y) under standard priors, see Norets and Pati (2017) for a proof for the version of the model without discrete variables.

We specify the prior as follows,

$$\begin{aligned}
&\beta_j \stackrel{iid}{\sim} N(\underline{\beta}, \underline{H}_\beta^{-1}),\ \mu_j \stackrel{iid}{\sim} N(\underline{\mu}, \underline{H}_\mu^{-1}),\\
&(s_j^y)^{-2} \stackrel{iid}{\sim} G(\underline{A}_{sy}, \underline{B}_{sy}),\ (s_{jk}^w)^{-2} \stackrel{iid}{\sim} G(\underline{A}_{swk}, \underline{B}_{swk}),\ k = 1, \ldots, d_w,\\
&(\sigma^y)^{-1} \stackrel{iid}{\sim} G(\underline{A}_{\sigma y}, \underline{B}_{\sigma y}),\ (\sigma_k^w)^{-1} \stackrel{iid}{\sim} G(\underline{A}_{\sigma wk}, \underline{B}_{\sigma wk}),\ k = 1, \ldots, d_w,\\
&(\alpha_1, \ldots, \alpha_m)|m \stackrel{iid}{\sim} D(\underline{a}/m, \ldots, \underline{a}/m),\\
&\Pi\big(m = k\big) = (e^{\underline{A}_m} - 1)e^{-\underline{A}_m \cdot k},
\end{aligned}$$

where $G(A, B)$ stands for a Gamma distribution with shape A and rate B.

Similarly to Norets and Pati (2017), we use the following (data-dependent) values for prior hyper-parameters,

$$\begin{aligned}
&\underline{\beta} = \left(\sum_i w_i w_i'\right)^{-1} \sum_i w_i y_i,\ \underline{H}_\beta^{-1} = \underline{c}_\beta \left(\sum_i w_i w_i'\right)^{-1} \sum_i (y_i - w_i'\underline{\beta})^2/n,\\
&\underline{\mu} = \sum_i w_i/n,\ \underline{H}_\mu^{-1} = \sum_i (w_i - \underline{\mu})(w_i - \underline{\mu})'/n,\\
&\underline{A}_{\sigma y} = \underline{B}_{\sigma y} = \underline{A}_{\sigma wl} = \underline{B}_{\sigma wl} = \underline{A}_{swk} = \underline{B}_{swk} = \underline{A}_{sy} = \underline{B}_{sy} = 1,\\
&\underline{a} = 15,\ \underline{A}_m = 1,
\end{aligned}$$

where $\underline{c}_\beta = 100$. Thus, a modal prior draw would have one mixture component with linear coefficients and scale parameters estimated by the ordinary least squares. Scricciolo (2015) shows that in a related conditional distribution model for continuous data from Norets and Pati (2017), such dependence of prior hyperparameters on data does not affect the posterior contraction rates; we conjecture that such a result holds for our model as well.

For evaluating forecast performance, we use the time series count data from Jung et al. (2011). The dataset contain the number of trades on the New York Stock Exchange in 5 minute intervals for Geltfelter Company (GLT) over 39 trading days Jan 3 - Feb 18 2005. Cameron and Trivedi (2013) estimated an autoregressive Poisson model for these data using lagged trade counts for GLT and trigonometric terms, like $\cos(2\pi t/75)$, where t is time period, to account for intraday seasonality in the data. The total number of observations is 2925. We use a rolling window of $T = 1125$ observations (15 days) for model estimation and the subsequent $T^* = 75$ observations (1 day) for one period ahead forecasts. We move the window by 75 observations at a time, for a total of 23 estimation/forecast exercises. Following common practice in the literature, see, for example, Geweke and Keane (2007), we measure forecasting performance by the pseudo out-of-sample log score (the log of the predictive distribution evaluated at the forecast portion of the data):

$$LS(c) = \sum_{t=75c+T+1}^{75c+T+T^*} \log \hat{p}_c(y_t|w_t, y_{75c+1}, w_{75c+1}, \dots, y_{75c+T}, w_{75c+T}),$$

where $c \in \{0, 1, \dots, 22\}$ is the rolling window index. For our nonparametric Bayesian procedure, the predictive distribution is approximated by

$$\hat{p}_c(y_t|w_t, y_{75c+1}, w_{75c+1}, \dots, y_{75c+T}, w_{75c+T}) = \frac{1}{S}\sum_{s=1}^{S} p(y_t|w_t, \theta^{(s,c)}, m^{(s,c)}),$$

where $\{\theta^{(s,c)}, m^{(s,c)}, s = 1, \dots, S\}$ are MCMC draws obtained for the rolling window c, $\{y_{75c+1}, w_{75c+1}, \dots, y_{75c+T}, w_{75c+T}\}$.

We found that including more than 5 lags and more than one trigonometric term did not improve out-of-sample predictive performance. Hence, we present results below for $w_t = (\cos(2\pi t/75), y_{t-1}, \dots, y_{t-5})'$. To obtain estimation results for our model we use a reversible jump MCMC algorithm developed in Norets (2021) with an additional Gibbs block that simulates the latent variables $\tilde{y}_i$'s. For each MCMC run $c \in \{0, 1, \dots, 22\}$, we perform 10^4 iterations, of which the first 10^3 are discarded for burn-in. We present some evidence of MCMC convergence in Section 4.2.

The obtained predictive log scores are presented in Table 1. The kernel estimation results are obtained by the publicly available R package `np` (Hayfield and Racine (2008)).

It can be seen from Table 1 that the Bayesian nonparametric approach delivers the largest average predictive log score. It outperforms the kernel and Poisson estimators in 70% and 96% of cases correspondingly. The results are qualitatively the same for moderate changes in prior hyperparameters.

In addition to the out of sample performance of the whole predictive density (evaluated through the predictive log scores), the performance of point estimators could also be of interest.

Table 1. *Predictive Log Scores*

c	$LS^{Bayes}(c)$	$LS^{Kernel}(c)$	$LS^{Poisson}(c)$
0	-198.6982	-195.738	-216.583
1	-185.4372	-186.642	-181.545
2	-195.8008	-196.904	-206.36
3	-172.3776	-173.985	-172.69
4	-193.9293	-193.067	-201.649
5	-221.9528	-229.703	-242.763
6	-178.8345	-179.163	-183.172
7	-198.9912	-198.492	-210.458
8	-162.9572	-166.119	-168.421
9	-166.4577	-167.01	-170.67
10	-173.5684	-172.923	-174.255
11	-178.9721	-182.575	-189.41
12	-186.3612	-190.602	-191.57
13	-202.8732	-216.146	-221.549
14	-181.3866	-181.912	-181.497
15	-190.1113	-193.053	-204.388
16	-211.9176	-213.484	-227.713
17	-209.3348	-210.05	-232.783
18	-198.495	-198.441	-215.227
19	-198.0299	-200.466	-210.714
20	-202.8891	-202.121	-233.747
21	-200.7652	-199.856	-220.487
22	-188.0144	-197.366	-220.697
Average	-191.22	-193.30	-203.41

Table 2 shows the root mean squared error (RMSE) for the three methods computed for the prediction part of the sample and averaged over the 23 estimation/forecast exercises. The point estimator is the mean of the predictive distribution. As can be seen from the table, the Bayesian point estimator performs slightly better than the classical parametric and nonparametric alternatives.

Table 2. *Average RMSE for the prediction part of the sample*

	Bayes	Kernel	Poisson
RMSE	3.4495	3.4615	3.4595

These results suggest that our model provides an attractive and feasible alternative to standard parametric and nonparametric estimation procedures for conditional discrete-continuous distributions, including Markov transition distributions for time series.

4.2. *Evidence of MCMC Convergence.* Figures 1 and 2 below show MCMC draws of m and the in-sample log likelihood for 10^5 MCMC iterations for the first rolling window $c = 0$ in the forecast evaluation exercise in Section 4.1 above.

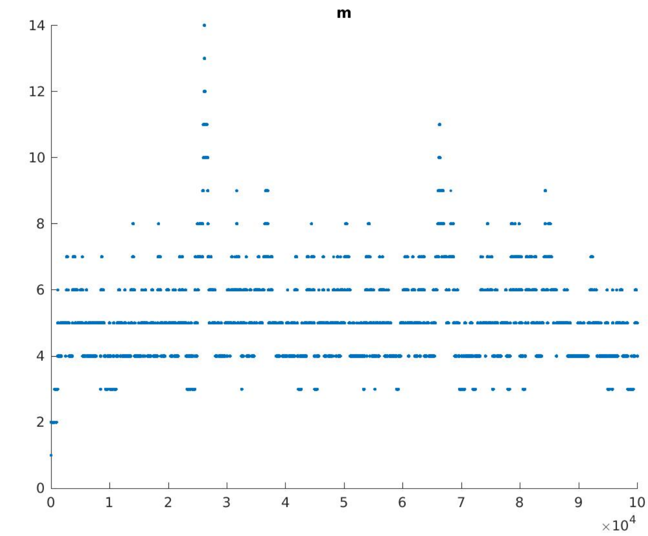

FIG 1. *MCMC draws of m.*

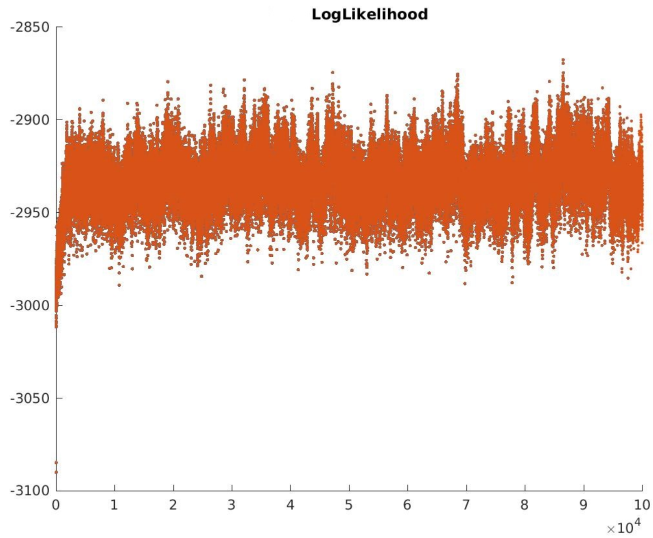

FIG 2. *In-sample log likelihood evaluated at MCMC draws.*

It is clear from the figures that while posterior probabilities for larger values of m would not be very precisely estimated even with 10^5 iterations, 10^4 MCMC iterations appear to be sufficient for exploring the posterior of the in-sample log likelihood (a label invariant function of parameters). Hence, we use 10^4 MCMC iteration for each rolling window in our forecast evaluation exercise.

5. Additional Applications. This section presents applications of the proposed method to two standard datasets in the literature on count data analysis (Cameron and Trivedi (2013)).

In the first application, we estimate the distribution of the number of patents filed by a firm in 1979 conditional on the logarithm of the total research and development expenditures over the previous six years, the logarithm of firm's capital, and an indicator of weather the firm is in the scientific sector. The data is originally from Hall et al. (1986), it "covers almost all of the firms doing appreciable amounts of R and D in the manufacturing sector" in the 1970ies. The sample size is 311.

In the second application, we estimate the distribution of the total number of children conditional on the mother's income at the time of the interview and a college degree dummy variable. The data is from Swiss Household Panel W1 (1999), it was originally analyzed in Cameron and Trivedi (2013). The sample size is 1878.

For both applications, the model and prior specifications are the same as for the application to trade counts in Section 4. The number of estimation/forecast exercises is 50 and for each of these exercises, the prediction part of the sample is a randomly selected 10 % of the available observations.

Tables 3 and 4 provide predictive log scores and RMSEs for the three methods averaged over the 50 estimation/forecast exercises.

TABLE 3. *Average Predictive Performance for Fertility Data*

	Bayes	Kernel	Poisson
Average Predictive Log Score	-124.6969	-128.2933	-133.5551
Average RMSE	1.3890	1.3987	1.3893

TABLE 4. *Average Predictive Performance for Patents Data*

	Bayes	Kernel	Poisson
Average Predictive Log Score	-114.1826	-133.3086	-361.1469
Average RMSE	34.4523	37.5913	41.9626

The results from both applications in this section confirm the results from the main application to the trade count data: our Bayesian model is an attractive approach to estimation of conditional discrete-continuous distributions.

6. Technical Assumptions.

6.1. *Assumptions on Prior.* The prior Π for (θ, m) is assumed to satisfy the conditions outlined below and matches the assumptions on the prior considered in Norets and Pelenis

(2021). The prior for σ_i satisfies

$$\begin{aligned} \Pi(\sigma_i^{-2} \geq s) &\leq a_1 \exp\{-a_2 s^{a_3}\} \quad \text{for all sufficiently large } s > 0 && (6.1)\\ \Pi(\sigma_i^{-2} < s) &\leq a_4 s^{a_5} \quad \text{for all sufficiently small } s > 0 && (6.2)\\ \Pi\{s < \sigma_i^{-2} < s(1+t)\} &\geq a_6 s^{a_7} t^{a_8} \exp\{-a_9 s^{1/2}\}, \quad s > 0, \quad t \in (0,1) && (6.3)\end{aligned}$$

for some positive constants $a_1, a_2, \ldots, a_9$ and for each $i \in \{1, \ldots, d\}$. The inverse Gamma prior for σ_i is an example of a prior that satisfies the proposed requirements.

Conditional on m, the prior for $(\alpha_1, \ldots, \alpha_m)$ is Dirichlet$(a/m, \ldots, a/m)$, $a > 0$. Prior for the number of mixture components m is

$$\Pi(m = i) \propto \exp(-a_{10} i (\log i)^{\tau_1}), i = 2, 3, \ldots, \quad a_{10} > 0, \tau_1 \geq 0. \tag{6.4}$$

The components of each μ_j, $\mu_{j,i}$, $i = 1, \ldots, d$, are independent from each other, other parameters, and across j. A sufficient condition on the prior is that the prior density for $\mu_{j,i}$ is bounded below for some $a_{12}, \tau_2 > 0$ by

$$a_{11} \exp(-a_{12} |\mu_{j,i}|^{\tau_2}), \tag{6.5}$$

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

$$\Pi(\mu_{j,i} \notin [-\mu, \mu]) \leq \exp(-a_{13} \mu^{\tau_3}). \tag{6.6}$$

6.2. *Technical Assumptions on the Data Generating Process.* In this subsection, we formulate technical assumptions on the data generating process for a fixed subset of indices for discrete variables $J \in \mathcal{A}$. In the main posterior contraction result in Theorem 3.1, these assumptions are assumed to hold for every $J \in \mathcal{A}$.

Let $d_J = card(J)$, $I = \{1, \ldots, d_y + d_z\} \setminus J$, $J^c = \{1, \ldots, d\} \setminus J$, and $d_{J^c} = card(J^c)$. Similarly to $\mathcal{Y}$, $\mathcal{Z}$, A_y and A_z defined in Section 2, we define $\mathcal{Y}_J = \prod_{j \in J} \mathcal{Y}_j$, $\mathcal{Z}_J = \prod_{j \in J} \mathcal{Z}_j$, $A_{y_J} = \prod_{i \in J} A_{y_i}$ and $A_{z_J} = \prod_{i \in J} A_{z_i}$. Also, let $y_J = \{y_i\}_{i \in J}$, $\tilde{y}_I = \{\tilde{y}_i\}_{i \in I}$, $z_J = \{z_i\}_{i \in J}$, $\tilde{z}_I = \{\tilde{z}_i\}_{i \in I}$, $\tilde{x} = (\tilde{y}_I, \tilde{z}_I, x, w) \in \tilde{\mathcal{X}} = \mathbb{R}^{d_{J^c}}$. In the proofs, we use the following notation for subsets of J^c and I: $J^c(x)$ denotes a set of indices of components of x that belong to J^c; $I(z)$, $J^c(w)$, $J^c(z)$, $J^c(x, w)$ are defined similarly.

The assumptions we formulate below are key to deriving optimal approximation results for the conditional data generating distribution that deliver (up to a log factor) the optimal posterior contraction rates. The approximation results for the conditional distribution are obtained by constructing a mixture of normals approximation to the following artificial joint distribution first

$$\bar{f}_0(\tilde{y}, x, \tilde{z}, w) = f_0(\tilde{y}, x|\tilde{z}, w)\bar{g}_0(\tilde{z}, w).$$

This artificial joint distribution has to have the conditional distribution equal to the data generating conditional distribution, but its marginal distribution, which we denote $\bar{g}_0(\tilde{z}, w)$ does not have to be equal to the data generating marginal distribution $g_0(\tilde{z}, w)$. Importantly, we can choose $\bar{g}_0(\tilde{z}, w)$ to be sufficiently smooth so that $\bar{f}_0(\tilde{y}, x, \tilde{z}, w)$ and $f_0(\tilde{y}, x|\tilde{z}, w)$ belong to the same smoothness class, and, hence, optimal approximations for $\bar{f}_0(\tilde{y}, x, \tilde{z}, w)$ can deliver optimal approximations for $f_0(\tilde{y}, x|\tilde{z}, w)$. The simplest way to interpret the following assumptions is to consider $\tilde{\mathcal{Z}} \times \mathcal{W} = [0, 1]^{d_z+d_w}$. In this case, one could take uniform $\bar{g}_0(\tilde{z}, w) = 1_{[0,1]^{d_z+d_w}}(\tilde{z}, w)$ and $\bar{f}_0(\tilde{y}, x, \tilde{z}, w) = f_0(\tilde{y}, x|\tilde{z}, w)$, and the following assumptions on $\bar{f}_0$ are straightforward to interpret in terms of f_0. Alternatively, if $g_0(\tilde{z}, w)$ is more smooth than $f_0(\tilde{y}, x|\tilde{z}, w)$, then one can consider $\bar{g}_0 = g_0$ and the Holder smoothness assumptions below would essentially restrict only $f_0(\tilde{y}, x|\tilde{z}, w)$. If $g_0(\tilde{z}, w)$ has a lower smoothness level than $f_0(\tilde{y}, x|\tilde{z}, w)$, then the assumptions below effectively require a well behaved and smooth $\bar{g}_0$ that bounds g_0 from above up to a multiplicative constant. We use the general form of assumptions below in order to accommodate unbounded $\tilde{\mathcal{Z}} \times \mathcal{W}$ and arbitrary smoothness in g_0.

Let us introduce the notation for marginal and conditional distributions implied by $\bar{f}_0$,

$$\begin{aligned}
\bar{f}_{0|J}(\tilde{x}|y_J, z_J) &= \frac{\bar{f}_{0J}(y_J, z_J, \tilde{x})}{\bar{\pi}_{0J}(y_J, z_J)}, \\
\bar{f}_{0J}(y_J, z_J, \tilde{x}) &= \int_{A_{y_J} \times A_{z_J}} \bar{f}_0(\tilde{y}_J, \tilde{z}_J, \tilde{x}) d\tilde{y}_J d\tilde{z}_J, \\
\bar{\pi}_{0J}(y_J, z_J) &= \int_{\tilde{\mathcal{X}}} \bar{f}_{0J}(y_J, z_J, \tilde{x}) d\tilde{x},
\end{aligned}$$

where the conditional density $\bar{f}_{0|J}(\tilde{x}|y_J, z_J)$ can be defined arbitrarily when $\bar{\pi}_{0J}(y_J, z_J) = 0$. Also, let $\bar{F}_{0|J}$ denote the conditional probability corresponding to the conditional density $\bar{f}_{0|J}$.

Assumption 6.1. *Assume that there exists a constant $\eta > 0$ and a probability density function $\bar{g}_0(\tilde{z}, w)$ with respect to the Lebesque measure such that $\eta\bar{g}_0(\tilde{z}, w) \geq g_0(\tilde{z}, w)$ for all $(\tilde{z}, w) \in \tilde{\mathcal{Z}} \times \mathcal{W}$.*

Assumption 6.2. *There are positive finite constants $b, \bar{f}_0, \tau$ such that for any $(y_J, z_J) \in \mathcal{Y}_J \times \mathcal{Z}_J$ and $\tilde{x} \in \tilde{\mathcal{X}}$*

$$\bar{f}_{0|J}(\tilde{x}|y_J, z_J) \leq \bar{f}_0 \exp\left(-b||\tilde{x}||^\tau\right). \tag{6.7}$$

Similar tail conditions on data generating densities are imposed in most of the papers on (near) optimal posterior contraction rates for mixtures of normal densities.

Assumption 6.3. *We assume that*

$$\bar{f}_{0|J} \in \mathcal{C}^{\beta_{d_J+1},\ldots,\beta_d,L}, \tag{6.8}$$

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

$$L(\tilde{x}, \Delta\tilde{x}) = \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}, \tag{6.9}$$

$$\tilde{L}(\tilde{x} + \Delta\tilde{x}) \leq \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}. \tag{6.10}$$

Simple sufficient conditions for $\bar{f}_{0|J} \in \mathcal{C}^{\beta_{d_J+1},\ldots,\beta_d,L}$ for all $J \in \mathcal{A}$ are $\bar{f}_0$ is bounded away from zero, has bounded support and belongs to $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ (Lemma 5.8. in Norets and Pelenis (2021)).

Assumption 6.4. *There are positive finite constants* ε *and* $\bar{F}$*, such that for any* $(y_J, z_J) \in \mathcal{Y}_J \times \mathcal{Z}_J$ *and* $k = \{k_i\}_{i\in J^c} \in \mathbb{N}_0^{d_{J^c}}$, $\sum_{i\in J^c} k_i/\beta_i < 1$,

$$\int \left[\frac{|D^k \bar{f}_{0|J}(\tilde{x}|y_J, z_J)|}{\bar{f}_{0|J}(\tilde{x}|y_J, z_J)}\right]^{\frac{(2+\varepsilon\beta_{J^c}^{-1} d_{J^c}^{-1})}{\sum_{i\in J^c} k_i/\beta_i}} \bar{f}_{0|J}(\tilde{x}|y_J, z_J) d\tilde{x} < \bar{F}, \tag{6.11}$$

$$\int \left[\frac{\tilde{L}(\tilde{x})}{\bar{f}_{0|J}(\tilde{x}|y_J, z_J)}\right]^{2+\varepsilon\beta_{J^c}^{-1} d_{J^c}^{-1}} \bar{f}_{0|J}(\tilde{x}|y_J, z_J) d\tilde{x} < \bar{F}. \tag{6.12}$$

This assumption is mostly relevant for the case of the unbounded support and the proposed condition suggests that the envelope function $\tilde{L}$ should be comparable to $\bar{f}_{0|J}$.

Assumption 6.5. *There exists a positive and finite* $\bar{y}$ *such that for any* $(y_I, y_J) \in \mathcal{Y}$, $z \in \mathcal{Z}$, $w \in \mathcal{W}$ *and* $x \in \mathcal{X}$

$$\sup_{\tilde{y}_I \in A_{y_I} \cap \{||\tilde{y}_I|| \leq \bar{y}\}} f_0(\tilde{y}_I, y_J, x|z, w) \leq \bar{f}$$

$$\int_{A_{y_I} \cap \{||\tilde{y}_I|| \leq \bar{y}\}} f_0(\tilde{y}_I, y_J, x|z, w) d\tilde{y}_I \geq \int_{A_{y_I} \cap \{||\tilde{y}_I|| > \bar{y}\}} f_0(\tilde{y}_I, y_J, x|z, w) d\tilde{y}_I$$

The second inequality in the assumption always holds for A_{y_I} contained within the unit cube. When A_{y_I} is a rectangle with at least one infinite side, an interpretation of this assumption is that the tail probabilities for the latent variable $\tilde{y}_I$ conditional on (x, y_J, z, w) decline uniformly in (x, y_J, z, w). A simple sufficient condition for this is a bounded support for $\tilde{y}_I$.

Assumption 6.6. *We assume that g_0 satisfies*

$$\int e^{\kappa||w||^2} g_0(w, \tilde{z}) d\tilde{z} dw \leq B < \infty$$

for some constant $\kappa > 0$ and $B > 0$.

This assumption of sub-Gaussian tails for the data generating distribution of continuous covariates w allows us to handle an unbounded support as in Norets and Pati (2017).

Assumption 6.7. *For some small $\nu > 0$,*

$$N_J = o(n^{1-\nu}). \tag{6.13}$$

As some parts of the proof require $\log(1/\epsilon_n)$ to be of order $\log n$ this condition is imposed to exclude the case of N_J implying very slow (non-polynomial) rates.

7. Proofs and Intermediate Results for Posterior Contraction Rates.

Let

$$t_{J0} = \begin{cases} \frac{d_{J^c}[1+1/(\beta_{J^c} d_{J^c})+1/\tau]+\max\{\tau_1,1,\tau_2/\tau\}}{2+1/\beta_{J^c}} & \text{if } J^c \neq \emptyset \\ \max\{\tau_1, 1\}/2 & \text{if } J^c = \emptyset \end{cases} \tag{7.1}$$

where (τ, τ_1, τ_2) are defined in Section 6.

Theorem 7.1. *Suppose the assumptions from Sections 6.1 and 6.2 hold for a given $J \in \mathcal{A}$. Let*

$$\epsilon_n = \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{7.2}$$

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

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

As in Section 2, when $J^c = \emptyset$, β_{J^c} can be defined to be infinity and $\beta_{J^c}/(2\beta_{J^c}+1) = 1/2$ in (7.2).

Theorem 7.1 provides a valid upper bound on the posterior contraction rate under the assumptions for a fixed J. Theorem 3.1 imposes the same assumptions for every $J \in \mathcal{A}$; hence, the smallest bound over J from Theorem 7.1 applies, and Theorem 3.1 is immediately implied by Theorem 7.1.

7.1. *Proof of Theorem 7.1.* Let us introduce some additional notation,

$$\begin{aligned}
p_0(z,w) &= \int_{A_z} g_0(\tilde{z},w)d\tilde{z} \\
p_0(y,x|z,w) &= \frac{p_0(y,x,z,w)}{p_0(z,w)} \\
f_0(\tilde{y}_I,y_J,x,z,w) &= \int_{A_{y_J}}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)g_0(\tilde{z},w)d\tilde{y}_J d\tilde{z} \\
f_0(\tilde{y}_I,y_J,x|z,w) &= \frac{f_0(\tilde{y}_I,y_J,x,z,w)}{p_0(z,w)}
\end{aligned}$$

To prove Theorem 7.1, we use the sufficient conditions for posterior contraction from Theorem 2.1. in Ghosal and van der Vaart (2001). As was previously noted in Shen and Ghosal (2016) and Norets and Pati (2017), the results in Ghosal and van der Vaart (2001) for joint distributions do not require any substantive modifications for the case of conditional distributions as long as the expected total variation distance, d_{TV}, is used. Let ϵ_n and $\tilde{\epsilon}_n$ be positive sequences with $\tilde{\epsilon}_n \leq \epsilon_n$, $\epsilon_n \to 0$, and $n\tilde{\epsilon}_n^2 \to \infty$, and c_1, c_2, c_3, and c_4 be some positive constants. Let ρ be the expected total variation or Hellinger distance and suppose $\mathcal{F}_n \subset \mathcal{F}$ is a sieve with the following bound on the metric entropy $M_e(\epsilon_n, \mathcal{F}_n, \rho)$

$$\log M_e(\epsilon_n, \mathcal{F}_n, \rho) \leq c_1 n\epsilon_n^2, \tag{7.3}$$

$$\Pi(\mathcal{F}_n^c) \leq c_3 \exp\{-(c_2+4)n\tilde{\epsilon}_n^2\}. \tag{7.4}$$

Suppose also that the prior thickness condition holds

$$\Pi(\mathcal{K}(p_0,\tilde{\epsilon}_n)) \geq c_4 \exp\{-c_2 n\tilde{\epsilon}_n^2\}, \tag{7.5}$$

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

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

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

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

The choice of the sieve and verification of the conditions (7.3) and (7.4) are similar to a number of comparable results in the literature on posterior contraction rates for mixture models. The details with the adjustments to the present set-up are given in Lemma 8.6 in the Appendix. The prior thickness condition requires a bit more effort to verify and therefore we formulate and prove it as a separate theorem. Parts of the proof employ the results obtained in the corresponding proof of Theorem 4.2. in Norets and Pelenis (2021).

THEOREM 7.2. *Suppose the assumptions from Sections 6.1 and 6.2 hold for a given $J \in \mathcal{A}$. Let $t_J > t_{J0}$, where t_{J0} is defined in (7.1), and*

$$\tilde{\epsilon}_n = \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}. \tag{7.6}$$

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

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq \exp\{-Cn\tilde{\epsilon}_n^2\}. \tag{7.7}$$

PROOF. By Lemma 8.1 for $p(\cdot|\cdot, \theta, m)$ defined in (3.2)

$$\begin{aligned}
&d_h^2\left(p(y, x|z, w, \theta, m)p_0(z, w), p_0(y, x|z, w)p_0(z, w)\right) \\
&\leq 4\eta d_h^2\left(\int_{A_y} p(\tilde{y}, x|z, w, \theta, m)p(z, w|\theta, m)d\tilde{y}, \int_{A_y}\int_{A_z} f_0(\tilde{y}, x|\tilde{z}, w)\bar{g}_0(\tilde{z}, w)d\tilde{z}d\tilde{y}\right) \\
&= 4\eta d_h^2\left(p(y, x|z, w, \theta, m)p(z, w|\theta, m), p_0(y, x|z, w)\bar{p}_0(z, w)\right).
\end{aligned}$$

With this inequality, we can exploit approximation results derived for joint discrete-continuous distributions.

Define $\beta = d_{J^c}\left[\sum_{k\in J^c}\beta_k^{-1}\right]^{-1}$, $\beta_{\min} = \min_{j\in J^c}\beta_j$, and $\sigma_n = [\tilde{\epsilon}_n/\log(1/\tilde{\epsilon}_n)]^{1/\beta}$. For ε defined in (6.11)-(6.12), b and τ defined in (6.7), and a sufficiently small $\delta > 0$, let $a_0 = \{(8\beta + 4\varepsilon + 8 + 8\beta/\beta_{\min})/(b\delta)\}^{1/\tau}$, $a_{\sigma_n} = a_0\{\log(1/\sigma_n)\}^{1/\tau}$, and $b_1 > \max\{1, 1/2\beta\}$ satisfying $\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{5/4} \leq \tilde{\epsilon}_n$. The proofs of Theorems 4 and 6 in Shen et al. (2013) imply the following claim for each $(y_J, z_J) = k \in \mathcal{Y}_J \times \mathcal{Z}_J$ under the assumptions of Section 6.2.

There exists a partition $\{U_{j|k}, j = 1, \ldots, K\}$ of $\{\tilde{x} \in \tilde{\mathcal{X}} : ||\tilde{x}|| \leq 2a_{\sigma_n}\}$, such that for $j = 1, \ldots, N$, $U_{j|k}$ is contained within an ellipsoid with center $\mu^\star_{j|k}$ and radii $\{\sigma_n^{\beta/\beta_i}\tilde{\epsilon}_n^{2b_1}, i \in J^c\}$

$$U_{j|k} \subset \left\{\tilde{x} : \sum_{i=1}^{d_{J^c}}\left[(\tilde{x}_i - \mu^\star_{j|k,i})/(\sigma_n^{\beta/\beta_{d_J+i}}\tilde{\epsilon}_n^{2b_1})\right]^2 \leq 1\right\};$$

for $j = N+1, \ldots, K$, $U_{j|k}$ is contained within an ellipsoid with radii $\{\sigma_n^{\beta/\beta_i}, i \in J^c\}$, and $1 \leq N < K \leq C_1\sigma_n^{-d_{J^c}}\{\log(1/\tilde{\epsilon}_n)\}^{d_{J^c}+d_{J^c}/\tau}$, where $C_1 > 0$ does not depend on n and y_J.

Furthermore, by Lemma 5.10 in Norets and Pelenis (2021), there exists a constant $B_0 > 0$ such that for all $(y_J, z_J) \in \mathcal{Y}_J \times \mathcal{Z}_J$

$$\bar{F}_{0|J}\left(||\tilde{X}|| > a_{\sigma_n}|y_J, z_J\right) \leq B_0 \sigma_n^{4\beta+2\varepsilon} \underline{\sigma}_n^8, \tag{7.8}$$

where

$$\underline{\sigma}_n = \min_{i \in J^c} \sigma_n^{\beta/\beta_i}.$$

For $m = N_J K$ we define $\theta^\star$ and $S_{\theta^\star}$ as

$$\begin{aligned}
\theta^\star = \Big\{ & \{\mu_1^\star, \dots, \mu_m^\star\} = \Big\{(k, \mu_{j|k}^\star),\ j = 1, \dots, K,\ k \in \mathcal{Y}_J \times \mathcal{Z}_J \Big\}, \\
& \{\alpha_1^\star, \dots, \alpha_m^\star\} = \Big\{\alpha_{jk}^\star = \alpha_{j|k}^\star \bar{\pi}_{0J}(k),\ j = 1, \dots, K,\ k \in \mathcal{Y}_J \times \mathcal{Z}_J \Big\}, \\
& \sigma_J^{\star 2} = \{\sigma_i^{\star 2} = 1/[64 N_i^2 \beta \log(1/\sigma_n)],\ i \in J\} \\
& \sigma_{J^c}^\star = \{\sigma_i^\star = \sigma_n^{\beta/\beta_i},\ i \in J^c\}, \Big\}
\end{aligned}$$

$$\begin{aligned}
S_{\theta^\star} = \Big\{ & \{\mu_1, \dots, \mu_m\} = \{(\mu_{jk,J}, \mu_{jk,J^c}),\ j = 1, \dots, K,\ k \in \mathcal{Y}_J \times \mathcal{Z}_J\}, \\
& \mu_{jk,J^c} \in U_{j|k}, \quad \mu_{jk,i} \in \left[k_i - \frac{1}{4N_i}, k_i + \frac{1}{4N_i}\right],\ i \in J, \\
& \sigma_i^2 \in \left(0, \sigma_i^{\star 2}\right),\ i \in J, \\
& \sigma_i^2 \in \left(\sigma_i^{\star 2}(1 + \sigma_n^{2\beta})^{-1}, \sigma_i^{\star 2}\right),\ i \in J^c, \\
& (\alpha_1, \dots, \alpha_m) = \{\alpha_{jk},\ j = 1, \dots, K,\ k \in \mathcal{Y}_J \times \mathcal{Z}_J\} \in \Delta^{m-1}, \\
& \sum_{r=1}^m |\alpha_r - \alpha_r^\star| \leq 2\sigma_n^{2\beta}, \qquad \min_{j \leq K, k \in \mathcal{Y}_J \times \mathcal{Z}_J} \alpha_{jk} \geq \frac{\sigma_n^{2\beta + d_{J^c}}}{2m^2} \Big\}.
\end{aligned}$$

If the assumptions from Section 6.2 hold, then it is shown in equation (4.27) in Norets and Pelenis (2021) that for $m = N_J K$ and $\theta \in S_{\theta^*}$

$$d_h^2\left(p(y, x|z, w, \theta, m) p(z, w|\theta, m), p_0(y, x|z, w) \bar{p}_0(z, w)\right) \leq \sigma_n^{2\beta}. \tag{7.9}$$

Next, let us consider a lower bound on the ratio $p(y, x|z, w, \theta, m)/p_0(y, x|z, w)$ for $\theta \in S_{\theta^\star}$ and $m = N_J K$. In Lemma 8.3 in the Appendix we show that for any $(x, w) \in \mathcal{X} \times \mathcal{W}$ with $\|(x, w)\| \leq a_{\sigma_n}$,

$$\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} \geq C_2 \frac{\sigma_n^{2\beta}}{2m^2} \prod_{i \in J^c(w, z_I)} \sigma_n^{\frac{\beta}{\beta_i}} = \lambda_n. \tag{7.10}$$

for some constant $C_2 > 0$; and for any $(x, w) \in \mathcal{X} \times \mathcal{W}$ with $\|(x, w)\| > a_{\sigma_n}$,

$$\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} \geq \exp\left\{-\frac{8\,\|(x, w)\|^2}{\underline{\sigma}_n^2} - C_3 \log n\right\} \tag{7.11}$$

for some constant $C_3 > 0$.

Consider all sufficiently large n such that $\lambda_n < e^{-1}$ and (7.10) and (7.11) hold. Then, for any $\theta \in S_{\theta^\star}$,

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\mathcal{X}\times\mathcal{W}}\left(\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\right)^2 \mathbf{1}\left\{\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} < \lambda_n\right\} p_0(y, z, x, w)dwdx \\
&= \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\mathcal{X}\times\mathcal{W}\times\mathcal{Y}_I\times\mathcal{Z}_I}\left(\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\right)^2 \\
&\mathbf{1}\left\{\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} < \lambda_n, ||(x', w')'|| > a_{\sigma_n}, \tilde{y}_I \in A_{y_I}, \tilde{z}_I \in A_{z_I}\right\} f_0(y_J, z_J, \tilde{z}_I, \tilde{y}_I, w, x)d\tilde{x} \\
&\leq \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\{\tilde{x}:||(x', w')'|| > a_{\sigma_n}\}}\left(\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\right)^2 \mathbf{1}\left\{\tilde{y}_I \in A_{y_I}, \tilde{z}_I \in A_{z_I}\right\} f_0(y_J, z_J, \tilde{x})d\tilde{x} \\
&\leq \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\{\tilde{x}:||(x', w')'|| > a_{\sigma_n}\}}\left[\frac{128}{\underline{\sigma}_n^4}||\tilde{x}||^4 + 2(C_3\log n)^2\right] f_{0|J}(\tilde{x}|y_J, z_J)d\tilde{x}\bar{\pi}_{0J}(y_J, z_J) \\
&\leq \frac{128}{\underline{\sigma}_n^4}\sum_{y_J\in\mathcal{Y}_J}\sum_{z_J\in\mathcal{Z}_J} E_{0|y_J,z_J}\left(\left\|\tilde{X}\right\|^8\right)^{1/2}\left(F_{0|y_J,z_J}\left(\left\|\tilde{X}\right\| > a_{\sigma_n}\right)\right)^{1/2}\bar{\pi}_{0J}(y_J, z_J) \\
&+ 2(C_3\log n)^2 B_0\sigma_n^{4\beta+2\varepsilon}\underline{\sigma}_n^8 \\
&\leq C_4\sigma_n^{2\beta+\varepsilon}
\end{aligned} \tag{7.12}$$

for some constant $C_4 > 0$ and all sufficiently large n, where the last inequality holds by the tail condition in (6.7), (7.8), and $(\log n)^2\sigma_n^{2\beta+\varepsilon}\underline{\sigma}_n^8 \to 0$.

Furthermore, for n large enough such that $\lambda_n < e^{-1}$,

$$\begin{aligned}
&\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\mathbf{1}\left\{\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} < \lambda_n\right\} \\
&\leq \left(\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\right)^2\mathbf{1}\left\{\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} < \lambda_n\right\}
\end{aligned}$$

and, therefore,

$$\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\mathcal{X}\times\mathcal{W}}\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\mathbf{1}\left\{\frac{p(y, x|z, w, \theta, m)}{p_0(y, x|z, w)} < \lambda_n\right\} p_0(y, z, x, w)dwdx \leq C_4\sigma_n^{2\beta+\varepsilon}. \tag{7.13}$$

Inequalities (7.9), (7.12), and (7.13) combined with Lemma 8.2 imply

$$E_0\left(\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\right) \leq A\tilde{\epsilon}_n^2,\ E_0\left(\left[\log\frac{p_0(y, x|z, w)}{p(y, x|z, w, \theta, m)}\right]^2\right) \leq A\tilde{\epsilon}_n^2$$

for any $\theta \in S_{\theta^\star}$, $m = N_J K$, and some positive constant A (details are provided in Lemma 8.4 in the Appendix).

Since the definition of S_{θ^*} is adapted from the corresponding definition in Norets and Pelenis (2021), Lemma 5.16 in the Appendix of Norets and Pelenis (2021) delivers that for all sufficiently large n, $s = 1 + 1/\beta + 1/\tau$, and some $C_5 > 0$,

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq \Pi(m = N_J K, \theta \in S_{\theta^\star}) \geq \exp\left[-C_5 N_J \tilde{\epsilon}_n^{-d_{J^c}/\beta}\{\log(n)\}^{d_{J^c}s+\max\{\tau_1,1,\tau_2/\tau\}}\right].$$

The right hand side in the inequality above is bounded below by $\exp\{-Cn\tilde{\epsilon}_n^2\}$ for any $C > 0$, $\tilde{\epsilon}_n = \left[\frac{N_J}{n}\right]^{\beta/(2\beta+d_{J^c})}(\log n)^{t_J}$, any $t_J > (d_{J^c}s+\max\{\tau_1, 1, \tau_2/\tau\})/(2+d_{J^c}/\beta)$, and all sufficiently large n. As the inequality in the definition of t_J is strict the theorem is immediately implied. When $J = \emptyset$ and $N_J = 1$, the theorem can be proved by the same argument if we add an artificial discrete coordinate with only one possible value to the vector of observables.

□

7.2. *Extension to Markov processes.* Our model can be used for specifying a prior on Markov transition probabilities as one could just set $(z_t, w_t) = (y_{t-1}, x_{t-1})$. General sufficient conditions for posterior contraction rates for Markov transition probabilities were obtained in Ghosal and van der Vaart (2007a); however, they appear to be too strong for models based mixtures of normals. Martin and Hong (2012) provide very weak sufficient conditions for convergence rates of predictive distributions in the context of ergodic Markov processes. Specifically, their theoretical results in Section 7 and their Proposition 5 imply that $n^{-1}\sum_{i=1}^n E\left[K_{Y_{i-1}}(f_{\theta^*}, \hat{f}_{i-1})\right] = O_P(\epsilon_n^2)$, where K is the Kullback-Leibler divergence, θ^* is the "true" value of the parameter and $\hat{f}_{i-1}$ is the predictive distribution with respect to the posterior density Π_{i-1} for an ergodic Markov process $(Y_n : n \geq 0)$. A prior thickness condition for ϵ_n and $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ are sufficient for this result. Thus, our prior thickness results in Theorem 7.2 also deliver convergence rates for predictive distributions when our prior is used for modeling transition probability of an ergodic Markov process.

8. Conclusions and Future Work. In this paper, we propose and analyze a Bayesian nonparametric model for conditional discrete-continuous distributions. The model possesses outstanding asymptotic properties: it can fully exploit the smoothness in continuous and discrete variables if it is present in the data and delivers (up to a log term) optimal posterior contraction rates. The model is feasible to estimate by MCMC. In our applications, it performers better than standard classical parametric and nonparametric methods. Thus, it is an attractive alternative to workhorse limited dependent variable models such as probit, ordered probit and Poisson regression.

Discovering different model specifications for conditional discrete-continuous distributions that deliver optimal adaptive posterior contraction rates and that are feasible to estimate is an interesting direction for future work. More extensive simulation studies and applications of the model proposed in this paper are also of interest.

References.

AERTS, M., I. AUGUSTYNS, AND P. JANSSEN (1997): "Local Polynomial Estimation of Contingency Table Cell Probabilities," *Statistics*, 30, 127–148.

AITCHISON, J. AND C. G. G. AITKEN (1976): "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

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

BURDA, M., M. HARDING, AND J. HAUSMAN (2008): "A Bayesian Mixed Logit-Probit Model for Multinomial Choice," *Journal of Econometrics*, 147, pp. 232–246.

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

CAMERON, C. AND P. TRIVEDI (2013): *Regression analysis of count data*, Cambridge New York, NY: Cambridge University Press.

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

CHAMBERLAIN, G. AND K. HIRANO (1999): "Predictive Distributions Based on Longitudinal Earnings Data," *Annales d'Economie et de Statistique*, 211–242.

CHIB, S. AND E. GREENBERG (2010): "Additive cubic spline regression with Dirichlet process mixture errors," *Journal of Econometrics*, 156, 322–336.

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

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

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

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

GEWEKE, J. AND M. KEANE (2007): "Smoothly mixing regressions," *Journal of Econometrics*, 138, 252–290.

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

GHOSAL, S. AND A. VAN DER VAART (2007a): "Convergence rates of posterior distributions for noniid observations," *The Annals of Statistics*, 35, 192–223.

——— (2007b): "Posterior convergence rates of Dirichlet mixtures at smooth densities," *The Annals of Statistics*, 35, 697–723.

——— (2017): *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Ghosal, S. and A. W. van der Vaart (2001): "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *The Annals of Statistics*, 29, 1233–1263.

Hall, B. H., Z. Griliches, and J. A. Hausman (1986): "Patents and R and D: Is There a Lag?" *International Economic Review*, 27, 265–283.

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

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

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

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

Jensen, M. J. and J. M. Maheu (2014): "Estimating a semiparametric asymmetric stochastic volatility model with a Dirichlet process mixture," *Journal of Econometrics*, 178, 523–538.

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

Jung, R. C., R. Liesenfeld, and J.-F. Richard (2011): "Dynamic Factor Models for Multivariate Count Data: An Application to Stock-Market Trading Activity," *Journal of Business & Economic Statistics*, 29, 73–85.

Kruijer, W., J. Rousseau, and A. van der Vaart (2010): "Adaptive Bayesian density estimation with location-scale mixtures," *Electronic Journal of Statistics*, 4, 1225–1257.

Li, Q. and J. S. Racine (2008): "Nonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous Data," *Journal of Business and Economic Statistics*, 26, 423–434.

Martin, R. and L. Hong (2012): "On convergence rates of Bayesian predictive densities and posterior distributions," *arXiv:1210.0103*.

McCulloch, R. and P. Rossi (1994): "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64, 207–240.

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

——— (2021): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models," *Econometric Theory*, 37, 49–81.

Norets, A. and D. Pati (2017): "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 980–1012.

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

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

——— (2021): "Adaptive Bayesian Estimation of Mixed Discrete-Continuous Distributions under Smoothness and Sparsity," *Econometrica*, forthcoming.

Peng, F., R. A. Jacobs, and M. A. Tanner (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.

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

Scricciolo, C. (2006): "Convergence rates for Bayesian density estimation of infinite-dimensional exponential families." *Annals of Statatistics*, 34, 2897–2920.

——— (2015): "Empirical Bayes Conditional Density Estimation," *Statistica*, 75, 37–55.

Shen, W. and S. Ghosal (2016): "Adaptive Bayesian density regression for high-dimensional data," *Bernoulli*, 22, 396–420.

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

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

Wood, S., W. Jiang, and M. Tanner (2002): "Bayesian mixture of splines for spatially adaptive nonparametric regression," *Biometrika*, 89, 513–528.

Appendix.

Lemma 8.1. *Let $p_0(y,x|z,w)$ and $p(y,x|z,w,\theta,m)$ be conditional discrete continuous distributions. Let g be a density on $\mathcal{Z}\times\mathcal{W}$ and g_0 be a density on $\tilde{\mathcal{Z}}\times\mathcal{W}$ g_0 satisfying $\eta\bar{g}_0(\tilde z,w)\geq g_0(\tilde z,w)$ for all $(\tilde z,w)$. Then*

$$\begin{aligned}&d_h^2\left(p(y,x|z,w,\theta,m)p_0(z,w),p_0(y,x|z,w)p_0(z,w)\right)\\&\leq 4\eta d_h^2\left(p(y,x|z,w,\theta,m)g(z,w),p_0(y,x|z,w)\bar{p}_0(z,w)\right).\end{aligned}$$

Proof. Let $\bar{p}_0(z,w)=\int_{A_z}\bar{g}_0(\tilde z,w)d\tilde z$. Then

$$\begin{aligned}&d_h^2\left(p(y,x|z,w,\theta,m)p_0(z,w),p_0(y,x|z,w)p_0(z,w)\right)\\&=d_h^2\left(\int_{A_y}p(\tilde y,x|z,w,\theta,m)\int_{A_z}g_0(\tilde z,w)d\tilde z d\tilde y,\int_{A_y}\int_{A_z}f_0(\tilde y,x|\tilde z,w)g_0(\tilde z,w)d\tilde z d\tilde y\right)\\&\leq\eta d_h^2\left(\int_{A_y}p(\tilde y,x|z,w,\theta,m)\int_{A_z}\bar{g}_0(\tilde z,w)d\tilde z d\tilde y,\int_{A_y}\int_{A_z}f_0(\tilde y,x|\tilde z,w)\bar{g}_0(\tilde z,w)d\tilde z d\tilde y\right)\\&\leq\eta d_h^2\left(\int_{A_y}p(\tilde y,x|z,w,\theta,m)\bar{p}_0(z,w)d\tilde y,\int_{A_y}\int_{A_z}f_0(\tilde y,x|\tilde z,w)\bar{g}_0(\tilde z,w)d\tilde z d\tilde y\right)\\&\leq 2\eta\left[d_h^2\left(\int_{A_y}p(\tilde y,x|z,w,\theta,m)\bar{p}_0(z,w)d\tilde y,\int_{A_y}p(\tilde y,x|z,w,\theta,m)g(z,w)d\tilde y\right)\right.\end{aligned}$$

$$+ d_h^2\left(\int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}, \int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}\right)\Bigg] \leq 2\eta(I+II),$$

where

$$I = d_h^2\left(\int_{A_y} p(\tilde{y},x|z,w,\theta,m)\bar{p}_0(z,w)d\tilde{y}, \int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}\right)$$
$$II = d_h^2\left(\int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}, \int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}\right).$$

Note that

$$\begin{aligned}
I &= d_h^2\left(\int_{A_y} p(\tilde{y},x|z,w,\theta,m)\bar{p}_0(z,w)d\tilde{y}, \int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}\right)\\
&= d_h^2\left(\bar{p}_0(z,w), g(z,w)\right) = \sum_{z\in\mathcal{Z}}\int_{\mathcal{W}}\left(\sqrt{\bar{p}_0(z,w)} - \sqrt{g(z,w)}\right)^2 dw\\
&= 2\left(1 - \sum_{z\in\mathcal{Z}}\int_{\mathcal{W}}\sqrt{\bar{p}_0(z,w)g(z,w)}dw\right) \leq II,
\end{aligned}$$

where the final inequality follows from

$$\begin{aligned}
II &= \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\mathcal{W}\times\mathcal{X}}\Bigg(\int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y} + \int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}\\
&\quad -2\sqrt{\int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}\int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}}\Bigg)dwdx\\
&= 2\Bigg(1 - \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\mathcal{W}\times\mathcal{X}}\sqrt{\int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}}\\
&\qquad\cdot\sqrt{\int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}}dwdx\Bigg)\\
&= 2\Bigg(1 - \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\int_{\mathcal{W}\times\mathcal{X}}\sqrt{\int_{A_y} p(\tilde{y},x|z,w,\theta,m)d\tilde{y}\frac{\int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}}{\bar{p}_0(z,w)}}\\
&\qquad\cdot\sqrt{g(z,w)\bar{p}_0(z,w)}\Bigg)dwdx\\
&\geq 2\left(1 - \sum_{z\in\mathcal{Z}}\int_{\mathcal{W}}\sqrt{g(z,w)\bar{p}_0(z,w)}\right)dw
\end{aligned}$$

as for all z,w

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}}\int_{\mathcal{X}}\sqrt{\int_{A_y} p(\tilde{y},x|z,w,\theta,m)d\tilde{y}\frac{\int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}}{\bar{p}_0(z,w)}}dx\\
&\qquad\leq \frac{1}{2}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X}}\left(p(y,x|z,w,\theta,m)) + \frac{\int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}}{\bar{p}_0(z,w)}\right)dx = 1.
\end{aligned}$$

Combining the inequalities above

$$
\begin{aligned}
& d_h^2\left(p(y,x|z,w,\theta,m)p_0(z,w), p_0(y,x|z,w)p_0(z,w)\right) \leq 4\eta II \\
&= 4\eta d_h^2\left(\int_{A_y} p(\tilde{y},x|z,w,\theta,m)g(z,w)d\tilde{y}, \int_{A_y}\int_{A_z} f_0(\tilde{y},x|\tilde{z},w)\bar{g}_0(\tilde{z},w)d\tilde{z}d\tilde{y}\right) \\
&= 4\eta d_h^2\left(p(y,x|z,w,\theta,m)g(z,w), p_0(y,x|z,w)\bar{p}_0(z,w)\right).
\end{aligned}
$$

□

LEMMA 8.2. *There is a $\lambda_0 \in (0,1)$ such that for any $\lambda \in (0,\lambda_0)$ and any two conditional densities $p, q \in \mathcal{F}$, a probability measure P on $\mathcal{Z}$ that has a conditional density equal to p, and d_h defined with the distribution on $\mathcal{X}$ implied by P,*

$$
P\log\frac{p}{q} \leq d_h^2(p,q)\left(1+2\log\frac{1}{\lambda}\right) + 2P\left\{\left(\log\frac{p}{q}\right)1\left(\frac{q}{p}\leq\lambda\right)\right\},
$$

$$
P\left(\log\frac{p}{q}\right)^2 \leq d_h^2(p,q)\left(12+2\left(\log\frac{1}{\lambda}\right)^2\right) + 8P\left\{\left(\log\frac{p}{q}\right)^2 1\left(\frac{q}{p}\leq\lambda\right)\right\},
$$

PROOF. The proof is exactly the same as the proof of Lemma 4 of Shen et al. (2013), which in turn, follows the proof of Lemma 7 in Ghosal and van der Vaart (2007b). □

LEMMA 8.3. *Under the assumptions and notation of Section 7, for any $(y,z,x,w) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X} \times \mathcal{W}$, some constants $C_1, C_2 > 0$ and all sufficiently large n,*

$$
\frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} \geq C_1 \frac{\sigma_n^{2\beta}}{2m^2} \prod_{i\in J^c(w,z_I)} \sigma_n^{\frac{\beta}{\beta_i}} = \lambda_n.
$$

when $\|(x,w)\| \leq a_{\sigma_n}$ and

$$
\frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} \geq \exp\left\{-\frac{8\|(x,w)\|^2}{\underline{\sigma}_n^2} - C_2\log n\right\}
$$

when $\|(x,w)\| > a_{\sigma_n}$,

PROOF. For n large enough so that $a_{\sigma_n} > \bar{y}$ and by Assumption 6.5

$$
\begin{aligned}
& \frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} = \frac{\int_{A_{y_I}} f(y_J,\tilde{y}_I,x|z,w,\theta,m)d\tilde{y}_I}{\int_{A_{y_I}} f_0(y_J,\tilde{y}_I,x|z,w)d\tilde{y}_I} \\
& \geq \frac{\int_{A_{y_I}\bigcap\{||\tilde{y}_I||\leq a_{\sigma_n}\}} f(y_J,\tilde{y}_I,x|z,w,\theta,m)d\tilde{y}_I}{2\int_{A_{y_I}\bigcap\{||\tilde{y}_I||\leq a_{\sigma_n}\}} f_0(y_J,\tilde{y}_I,x|z,w)d\tilde{y}_I} \\
& \geq \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I\in A_{y_I}\bigcap\{||\tilde{y}_I||\leq a_{\sigma_n}\}} f(y_J,\tilde{y}_I,x|z,w,\theta,m)
\end{aligned}
$$

$$= \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I \in A_{y_I} \bigcap \{||\tilde{y}_I|| \le a_{\sigma_n}\}} \frac{f(y_J, \tilde{y}_I, x, z, w, \theta, m)}{p(z, w, \theta, m)}$$
$$= \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I \in A_{y_I} \bigcap \{||\tilde{y}_I|| \le a_{\sigma_n}\}} \frac{\int_{A_{y_J} \times A_Z} \sum_{j=1}^m \alpha_j \phi_j(\tilde{y}, x, \tilde{z}, w) d\tilde{y}_J d\tilde{z}}{\int_{A_Z} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}, w) d\tilde{z}}$$
$$\ge \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I \in A_{y_I} \bigcap \{||\tilde{y}_I|| \le a_{\sigma_n}\}} \frac{\int_{A_{y_J} \times A_Z} \alpha_{j^*} \phi_{j^*}(\tilde{y}_J) \phi_{j^*}(\tilde{y}_I) \phi_{j^*}(x) \phi_{j^*}(\tilde{z}_J) \phi_{j^*}(\tilde{z}_I) \phi_{j^*}(w) d\tilde{y}_J d\tilde{z}}{\int_{A_Z} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}_J) \phi_j(\tilde{z}_I) \phi_j(w) d\tilde{z}}.$$

Notation: ϕ_j is dependent on it's arguments contained within θ and $j \in \{1, \ldots, m\}$. To derive the bounds on the ratio we will consider two cases conditional on whether $||(x, w)|| \le a_{\sigma_n}$ or not.

For $||(x, w)|| \le a_{\sigma_n}$ choose j^* such that for all $i \in I(z)$

$$\int_{A_{y_J}} \phi_{j^*}(\tilde{y}_J) d\tilde{y}_J \ge \frac{1}{2}$$
$$\int_{A_{z_J}} \phi_{j^*}(\tilde{z}_J) d\tilde{z}_J \ge \frac{1}{2}$$
$$\text{if } A_{Z_i} \subseteq \begin{cases} \left(-\infty, \frac{1}{2N_i}\right], & Proj(U_{j^*}) \subset (-\infty, 0) \\ (0, 1), & z_i \in Proj(U_{j^*}) \\ \left(1 - \frac{1}{2N_i}, +\infty\right], & Proj(U_{j^*}) \subset (1, \infty) \end{cases} \tag{8.1}$$

As $||(x, w)|| \le a_{\sigma_n}$ and $\tilde{y}_I \le a_{\sigma_n}$, then there exists an ellipsoid $U^*_{j|k}$ such that it contains $(x, w, \tilde{y}_I)$. Furthermore, by the construction of ellipsoid $U^*_{j|k}$

$$\phi_{j^*}(\tilde{y}_I) \phi_{j^*}(x) \phi_{j^*}(w) \ge (2\pi)^{-1/2} \prod_{i \in J^c(x, w, \tilde{y}_I)} \sigma_i^{-1} \exp\{-1\}$$

For $A_{z_i} \subset [0, 1]$ we consider two cases with $\sigma_i \ge 1/2N_i$ and $\sigma_i < 1/2N_i$. When $\sigma_i \ge 1/2N_i$, then for the chosen j^* and all j

$$\int_{A_{z_i}} \phi_{j^*}(\tilde{z}_i) d\tilde{z}_i \ge e^{-1} \frac{\lambda(A_{z_i})}{\sqrt{2\pi}\sigma_i} \text{ and } \int_{A_{z_i}} \phi_j(\tilde{z}_i) d\tilde{z}_i \le \frac{\lambda(A_{z_i})}{\sqrt{2\pi}\sigma_i}.$$

When $\sigma_i < 1/2N_i$, then for the chosen j^* and all j

$$\int_{A_{z_i}} \phi_j(\tilde{z}_i) d\tilde{z}_i \le 1$$
$$\int_{A_{z_i}} \phi_{j^*}(\tilde{z}_i) d\tilde{z}_i = \int_{z_i - \frac{1}{2N_i}}^{z_i + \frac{1}{2N_i}} \phi(\tilde{z}_i, \mu_{j^*}, \sigma_i) d\tilde{z}_i = \int_{(z_i - \frac{1}{2N_i} - \mu_{j^*})/\sigma_i}^{(z_i + \frac{1}{2N_i} - \mu_{j^*})/\sigma_i} \phi(\tilde{z}_i, 0, 1) d\tilde{z}_i$$
$$= \int_{\Delta - \frac{1}{2N_i}\sigma_i}^{\Delta + \frac{1}{2N_i}\sigma_i} \phi(\tilde{z}_i, 0, 1) d\tilde{z}_i \ge \int_0^1 \phi(\tilde{z}_i, 0, 1) d\tilde{z}_i \approx 0.34,$$

where last inequality is true since $\Delta = (z_i - \mu_{j^*})/\sigma_i < 1$ by design of the ellipsoid U_{j^*} and $\frac{1}{2N_i}\sigma_i > 1$.

For $A_{z_i} \not\subset [0,1]$ and for the chosen j^* and all j

$$\int_{A_{z_i}} \phi_j(\tilde{z}_i)d\tilde{z}_i \leq 1$$

$$\int_{A_{z_i}} \phi_{j^*}(\tilde{z}_i)d\tilde{z}_i \geq \int_0^\infty \phi(\tilde{z}_i,0,1)d\tilde{z}_i = 0.5$$

as $\mu_{j^*} \in A_{z_i}$. In all these cases we obtain that for all j

$$\frac{\int_{A_{z_i}} \phi_j(\tilde{z}_i)d\tilde{z}_i}{\int_{A_{z_i}} \phi_{j^*}(\tilde{z}_i)d\tilde{z}_i} \leq \max\{e^{-1}, 0.34, 0.5\} = 0.5.$$

Then, combining the above results, we obtain that for $||(x,w)|| \leq a_{\sigma_n}$ the ratio is bounded by

$$\begin{aligned}
&\frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} \\
&\geq \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I \in A_{y_I} \bigcap \{||\tilde{y}_I|| \leq a_{\sigma_n}\}} \frac{\int_{A_{y_J} \times A_z} \alpha_{j^*}\phi_{j^*}(\tilde{y}_J)\phi_{j^*}(\tilde{y}_I)\phi_{j^*}(x)\phi_{j^*}(\tilde{z}_J)\phi_{j^*}(\tilde{z}_I)\phi_{j^*}(w)d\tilde{y}_J d\tilde{z}}{\int_{A_z} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}_J)\phi_j(\tilde{z}_I)\phi_j(w)d\tilde{z}} \\
&= \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I \in A_{y_I} \bigcap \{||\tilde{y}_I|| \leq a_{\sigma_n}\}} \frac{\int_{A_{y_J} \times A_{z_J}} \alpha_{j^*}\phi_{j^*}(\tilde{y}_J)\phi_{j^*}(\tilde{y}_I)\phi_{j^*}(x)\phi_{j^*}(\tilde{z}_J)\phi_{j^*}(w)d\tilde{y}_J d\tilde{z}_J}{\int_{A_z} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}_J)\frac{\phi_j(\tilde{z}_I)}{\int_{A_{z_I}} \phi_{j^*}(\tilde{z}_I)}\phi_j(w)d\tilde{z}} \\
&\geq C_1^* \frac{\min_j \alpha_j \prod_{i \in J^c(x,w,\tilde{y}_I)} \sigma_i^{-1}}{\int_{A_{z_j}} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}_J)\phi_j(w)d\tilde{z}} \geq C_1 \frac{\min_j \alpha_j \prod_{i \in J^c(x,w,\tilde{y}_I)} \sigma_i^{-1}}{\prod_{i \in J^c(w)} \sigma_i^{-1} \int_{A_{z_J}} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}_J)d\tilde{z}} \\
&\geq C_1 \min_j \alpha_j \prod_{i \in J^c(x,\tilde{y}_I)} \sigma_i^{-1} \geq C_1 \min_j \alpha_j \prod_{i \in J^c(x,\tilde{y}_I)} \sigma_n^{-\frac{\beta}{\beta_i}} \geq C_1 \frac{\sigma_n^{2\beta+d_{J^c}}}{2m^2} \prod_{i \in J^c(x,\tilde{y}_I)} \sigma_n^{-\frac{\beta}{\beta_i}} \\
&= C_1 \frac{\sigma_n^{2\beta}}{2m^2} \prod_{i \in J^c(w,\tilde{z}_I)} \sigma_n^{\frac{\beta}{\beta_i}} = \lambda_n.
\end{aligned}$$

Therefore, for sufficiently large n and $||(x,w)|| \leq a_{\sigma_n}$

$$\frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} \geq C_1 \frac{\sigma_n^{2\beta}}{2m^2} \prod_{i \in J^c(w,z_I)} \sigma_n^{\frac{\beta}{\beta_i}} = \lambda_n.$$

For $||(x,w)|| > a_{\sigma_n}$, we will derive a comparable bound for the ratio. First note, that by construction of ellipsoids $U_{j|k}$ for any $j \leq K$ and any $k \in \mathcal{Y}_J \times \mathcal{Z}_J$, $||(x',w') - \mu_{jk,J^c(w,x)}||^2 \leq ||\tilde{x} - \mu_{jk}||^2 \leq 12||(x',w')||^2$, where $\tilde{x} = (\tilde{y}_I', \tilde{z}_I', x', w')'$ with $||\tilde{y}_I|| \leq \bar{y} < a_{\sigma_n}$ and $\tilde{z}_I = 0$. Therefore, for sufficiently large n such that $1 + \sigma_n^{2\beta} < 8/6$, and, thus, $\sigma_i^2 > \underline{\sigma}_n^2 6/8$,

$$\phi_j(\tilde{y}_I)\phi_j(x)\phi_j(w) \geq C_2^* \prod_{i \in J^c(x,w,\tilde{y}_I)} \sigma_n^{-\frac{\beta}{\beta_i}} \exp\left\{-\frac{8||(x',w')||^2}{\underline{\sigma}_n^2}\right\},$$

where $\underline{\sigma}_n = \min_{i \in J^c} \sigma_n^{\beta/\beta_i}$. Then, for n large enough

$$f(y_J, z_J, \tilde{y}_I, \tilde{z}_I, x, w|\theta, m) = \Sigma_{k \in \mathcal{Y}_J \times \mathcal{Z}_J} \Sigma_{j=1}^K \alpha_{jk} \int_{A_{y_J} \times A_{z_J}} \phi_{jk}(\tilde{y}_J)\phi_{jk}(\tilde{z}_J)d\tilde{y}_J d\tilde{z}_J$$

$$
\begin{aligned}
&\cdot \phi_j(\tilde{y}_I)\phi_j(x)\phi_j(w)\phi_j(\tilde{z}_I) \\
&\geq C_2^* \prod_{i\in J^c(x,w,\tilde{y}_I)} \sigma_n^{-\frac{\beta}{\beta_i}} \exp\left\{-\frac{8||(x',w')||^2}{\underline{\sigma}_n^2}\right\} \\
&\cdot \Sigma_{j=1}^K \Sigma_{k\in\mathcal{Y}_J\times\mathcal{Z}_J} \alpha_{jk} \int_{A_{y_J}\times A_{z_J}} \phi_{jk}(\tilde{y}_J)\phi_{jk}(\tilde{z}_J) d\tilde{y}_J d\tilde{z}_J \phi_j(\tilde{z}_I) \\
&\geq C_2^* \prod_{i\in J^c(x,w,\tilde{y}_I)} \sigma_n^{-\frac{\beta}{\beta_i}} \exp\left\{-\frac{8||(x',w')||^2}{\underline{\sigma}_n^2}\right\} \min \alpha_{jk} K \phi_j(\tilde{z}_I).
\end{aligned}
$$

Next, pick j^* so that equation (8.1) is satisfied and by definition $\alpha^* \geq \min \alpha_{jk}$. Then, similarly to the previous case,

$$
\begin{aligned}
&\frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} \\
&\geq \frac{\bar{f}^{-1}}{2} \inf_{\tilde{y}_I\in A_{y_I}\bigcap\{||\tilde{y}_I||\leq a_{\sigma_n}\}} \frac{\int_{A_{y_J}\times A_Z} \alpha_{j^*}\phi_{j^*}(\tilde{y}_J)\phi_{j^*}(\tilde{y}_I)\phi_{j^*}(x)\phi_{j^*}(\tilde{z}_J)\phi_{j^*}(\tilde{z}_I)\phi_{j^*}(w) d\tilde{y}_J d\tilde{z}}{\int_{A_Z} \sum_{j=1}^m \alpha_j \phi_j(\tilde{z}_J)\phi_j(\tilde{z}_I)\phi_j(w) d\tilde{z}} \\
&\geq C_2^* \frac{\sigma_n^{2\beta}}{2m^2} \prod_{i\in J^c(w,\tilde{z}_I)} \sigma_n^{\frac{\beta}{\beta_i}} K \exp\left\{-\frac{8||(x',w')||^2}{\underline{\sigma}_n^2}\right\} \geq \exp\left\{-\frac{8||(x',w')||^2}{\underline{\sigma}_n^2} - C_2 \log n\right\}
\end{aligned}
$$

as for n large enough such that $\left|\log\left(K\frac{\sigma_n^{2\beta+\sigma_{i\in J^c(w,\tilde{z}_I)}{}^{\beta/\beta_i}}}{m^2}\right)\right| \leq \log n$. Therefore, for sufficiently large n and $||(x,w)|| > a_{\sigma_n}$

$$
\frac{p(y,x|z,w,\theta,m)}{p_0(y,x|z,w)} \geq \exp\left\{-\frac{8\,\|(x,w)\|^2}{\underline{\sigma}_n^2} - C_2 \log n\right\}.
$$

$\square$

LEMMA 8.4. *Under the assumptions and notation of Section 7, for $\lambda_n < \lambda_0$, where λ_0 is defined in Lemma 8.2,*

$$
\begin{aligned}
E_0\left(\log\frac{p_0(y,x|z,w)}{p(y,x|z,w,\theta,m)}\right) &\leq A\tilde{\epsilon}_n^2, \\
E_0\left(\left[\log\frac{p_0(y,x|z,w)}{p(y,x|z,w,\theta,m)}\right]^2\right) &\leq A\tilde{\epsilon}_n^2.
\end{aligned}
$$

PROOF.

$$
\begin{aligned}
&E_0\left(\left[\log\frac{p_0(y,x|z,w)}{p(y,x|z,w,\theta,m)}\right]^2\right) \\
&\leq d_H^2(p_0(\cdot|\cdot),p(\cdot|\cdot,\theta,m))\left(12+2\left(\log\frac{1}{\lambda_n}\right)^2\right) + 8P\left\{\left(\log\frac{p_0(\cdot|\cdot)}{p(\cdot|\cdot,\theta,m)}\right)^2 \mathbf{1}\left\{\frac{p(\cdot|\cdot,\theta,m)}{p_0(\cdot|\cdot)} < \lambda_n\right\}\right\} \\
&\lesssim \sigma_n^{2\beta}(12+2\log(1/\lambda_n)^2) + \sigma_n^{2\beta+\epsilon} \lesssim \log(1/\lambda_n)^2\sigma_n^{2\beta},
\end{aligned}
$$

where first inequality is derived using Lemma 8.2 and penultimate inequality is derived using inequalities (7.9) and (7.13). Similarly,

$$
\begin{aligned}
&E_0\left(\log \frac{p_0(y,x|z,w)}{p(y,x|z,w,\theta,m)}\right) \\
&\le d_H^2(p_0(\cdot|\cdot),p(\cdot|\cdot,\theta,m))\left(1+2\left(\log\frac{1}{\lambda_n}\right)\right)+2P\left\{\left(\log\frac{p_0(\cdot|\cdot)}{p(\cdot|\cdot,\theta,m)}\right)\mathbf{1}\left\{\frac{p(\cdot|\cdot,\theta,m)}{p_0(\cdot|\cdot)}<\lambda_n\right\}\right\} \\
&\lesssim \sigma_n^{2\beta}(1+2\log(1/\lambda_n))+\sigma_n^{2\beta+\epsilon}\lesssim \log(1/\lambda_n)\sigma_n^{2\beta}.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\log(1/\lambda_n)\sigma_n^{2\beta} \le \log(1/\lambda_n)^2\sigma_n^{2\beta} &= \log\left(\frac{2N_JK^2}{\sigma_n^{2\beta}}\prod_{i\in J^c(w,z_I)}\sigma_n^{-\frac{\beta}{\beta_i}}\right)^2\tilde{\epsilon}_n^2(\log(\tilde{\epsilon}_n^{-1}))^{-2} \\
&\le \left(\frac{\log\left[2N_J^2(C_1\sigma_n^{-d_{J^c}}\{\log(\tilde{\epsilon}_n^{-1})\}^{d_{J^c}+d_{J^c}/\tau})^2\sigma_n^{-2\beta}\prod_{i\in J^c(w,z_I)}\sigma_n^{-\frac{\beta}{\beta_i}}\right]}{\log(\tilde{\epsilon}_n^{-1})}\right)^2\tilde{\epsilon}_n^2,
\end{aligned}
$$

where the term multiplying $\tilde{\epsilon}_n^2$ on the right hand side is bounded by Assumption 6.7 ($N_J = o(n^{1-\nu})$) and definitions of $\tilde{\epsilon}_n$ and σ_n. □

LEMMA 8.5. *Under the assumptions and notation of Section 7, for $H\in\mathbb{N}$, $0<\underline{\sigma}<\overline{\sigma}$, and $\overline{\mu}>0$, let us define a sieve*

$$
\mathcal{F}=\{p(y,x|\theta,m):\ m\le H,\ \mu_j\in[-\overline{\mu},\overline{\mu}]^d, j=1,\dots,m,\sigma_i\in[\underline{\sigma},\overline{\sigma}], i=1,\dots,d\}. \tag{8.2}
$$

For $0<\epsilon<1$ and $\underline{\sigma}\le 1$,

$$
\begin{aligned}
M_e(\epsilon,\mathcal{F},d_{TV}) \le & H\cdot\left\lceil\frac{16\overline{\mu}(d_y+d_x)}{\underline{\sigma}\epsilon}\right\rceil^{H)(d_y+d_x)}\cdot\left\lceil\frac{384(d_w+d_z)\overline{\mu}^2}{\underline{\sigma}^2\epsilon}\right\rceil^{H(d_w+d_z)} \\
&\cdot H\left\lceil\frac{\log(\underline{\alpha}^{-1})}{\log(1+\epsilon/[12H])}\right\rceil^{H-1}\cdot\left\lceil\frac{\log(\overline{\sigma}/\underline{\sigma})}{\log(1+\underline{\sigma}^2\epsilon/[768(\overline{\mu}^x)^2\max\{d_x+d_y,d_z+d_w\}])}\right\rceil.
\end{aligned}
$$

For all sufficiently large H, large $\overline{\sigma}$ and small $\underline{\sigma}$,

$$
\begin{aligned}
\Pi(\mathcal{F}^c) \le & H^2 d\exp\{-a_{13}\overline{\mu}^{\tau_3}\}+\exp\{-a_{10}H(\log H)^{\tau_1}\} \\
&+da_1\exp\{-a_2\underline{\sigma}^{-2a_3}\}+da_4\exp\{-2a_5\log\overline{\sigma}\}.
\end{aligned}
$$

PROOF. The proof is similar to proofs of related results in Norets and Pati (2017), Shen et al. (2013), and Ghosal and van der Vaart (2001) among others.

For a fixed value of m, define set $S^m_{\mu^{y,x}}$ to contain centers of $|S^m_{\mu^{y,x}}| = \lceil 16\overline{\mu}(d_y+d_x)/(\underline{\sigma}\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu},\overline{\mu}]$. Then define set $S^m_{\mu^w}$ to contain centers of $|S^m_{\mu^w}| =$

$\lceil 384\overline{\mu}^2(d_w)/(\underline{\sigma}^2\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu},\overline{\mu}]$. Similarly, define set $S^m_{\mu^z}$ to contain centers of $|S^m_{\mu^z}| = \lceil 384\overline{\mu}^2(d_z)/(\underline{\sigma}^2\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu},\overline{\mu}]$

Define set S^m_σ as in Theorem 4.1. by Norets and Pati (2017), for $N_\alpha = \lceil \log(\underline{\alpha}^{-1})/\log(1+\epsilon/(12m))\rceil$ define

$$Q_\alpha = \{\gamma_j,\ j=1,\dots,N_\alpha:\ \gamma_1=\underline{\alpha},\ (\gamma_{j+1}-\gamma_j)/\gamma_j = \epsilon/(12m),\ j=1,\dots,N_\alpha-1\}$$

and let $S^m_\alpha = \{(\tilde\alpha_1,\dots,\tilde\alpha_m)\in\Delta^{m-1}:\ \tilde\alpha_{j_k}\in Q_\alpha,\ 1\le j_1<j_2<\dots<j_{m-1}\le m\}$.

Define

$$S_\sigma = \{\sigma^l, l=1,\dots,N_\sigma = \lceil \log(\overline{\sigma}/\underline{\sigma})/(\log(1+\underline{\sigma}^2\epsilon/(768(\overline{\mu})^2\max\{d_x+d_y,d_z+d_w\})\rceil, \sigma^1=\underline{\sigma},$$
$$(\sigma^{l+1}-\sigma^l)/\sigma^l = \underline{\sigma}^2\epsilon/(768(\overline{\mu})^2\max\{d_x+d_y,d_z+d_w\})\}.$$

Let us show that

$$S_{\mathcal{F}} = \{p(y,x|z,w,\theta,m):\ m\le H,\ \alpha\in S^m_\alpha,\ \sigma_i\in S_\sigma,\ \mu^y_{ji_y}\in S^m_{\mu^{y,x}},\ \mu^x_{ji_x}\in S^m_{\mu^{y,x}},\ \mu^w_{ji_w}\in S^m_{\mu^w},$$
$$\mu^w_{ji_z}\in S^m_{\mu^z},\ j\le m,\ i\le d, i_y\le d_y, i_x\le d_x, i_w\le d_w, i_z\le d_z\}$$

is an ϵ-net for $\mathcal{F}$ in d_{TV}. For a given $p(\cdot|\theta,m)\in\mathcal{F}$ with $\sigma^{l_i}\le\sigma_i\le\sigma^{l_i+1}$, $i=1,\dots,d$ find $\tilde\alpha\in S^m_\alpha$, $\tilde\mu^y_{ji_y}\in S^m_{\mu^{y,x}}$, $\tilde\mu^x_{ji_x}\in S^m_{\mu^{y,x}}$, $\tilde\mu^w_{ji_w}\in S^m_{\mu^w}$, $\tilde\mu^z_{ji_z}\in S^m_{\mu^z}$ and $\tilde\sigma_i=\sigma_{l_i}\in S_\sigma$ such that for all $j=1,\dots,m$, $i=1,\dots,d$, $i_y=1,\dots,d_y$, $i_x=1,\dots,d_x$, $i_w=1,\dots,d_w$ and $i_z=1,\dots,d_z$

$$\frac{\alpha_j-\tilde\alpha_j}{\alpha_j}\le\frac{\epsilon}{12},\ \frac{|\sigma_i-\tilde\sigma_i|}{\sigma_i}\le\frac{\underline{\sigma}^2\epsilon}{768\overline{\mu}^2\max\{d_x+d_y,d_z+d_w\}},\ |\mu^y_{ji_y}-\tilde\mu^y_{ji_y}|\le\frac{\underline{\sigma}\epsilon}{16(d_y+d_x)},$$
$$|\mu^x_{ji_x}-\tilde\mu^x_{ji_x}|\le\frac{\underline{\sigma}\epsilon}{16(d_y+d_x)},\ |\mu^w_{ji_w}-\tilde\mu^w_{ji_w}|\le\frac{\underline{\sigma}^2\epsilon}{384\overline{\mu}d_w},\ |\mu^z_{ji_z}-\tilde\mu^z_{ji_z}|\le\frac{\underline{\sigma}^2\epsilon}{384\overline{\mu}d_z}.$$

Applying Lemma 5.4 and equation (5.12) of Norets and Pelenis (2021) for each (z,w) we obtain that

$$d_{TV}(p(y,x|z,w,\theta,m),p(y,z|z,w,\tilde\theta,m))\le d_{TV}(f(\tilde y,x|z,w,\theta,m),f(\tilde y,x|z,w,\tilde\theta,m)).$$

Similarly to the proof of Theorem 4.1 in Norets and Pati (2017) for each $(z,w)\in\mathcal{Z}\times\mathcal{W}$

$$\int|p(\tilde y,x|z,w,\theta,m)-p(\tilde y,x|z,w,\tilde\theta,m)|dy\le 2\max_{j=1,\dots,m}||\phi_{\mu^{y,x}_j,\sigma}-\phi_{\tilde\mu^{y,x}_j,\tilde\sigma}||_1$$
$$+2\left(\max_j\frac{|K_j-\tilde K_j|}{K_j}+\max_j\frac{|\alpha_j-\tilde\alpha_j|}{\alpha_j}+\max_j\frac{|K_j-\tilde K_j||\alpha_j-\tilde\alpha_j|}{\alpha_jK_j}\right)$$

where

$$K_j=\prod_{i=1}^{d_w}\exp\left\{-\frac{(w_i-\mu^w_{ji})^2}{2(\sigma^z_i)^2}\right\}\prod_{i=1}^{d_z}\int_{A_{z_i}}\phi_i(\tilde z_i)d\tilde z_i.$$

As in Theorem 4.1. in Norets and Pati (2017) note that $||\phi_{\mu_j^{y,x},\sigma} - \phi_{\tilde{\mu}_j^{y,x},\tilde{\sigma}}||_1 \leq \frac{\epsilon}{4}$. Then note that

$$\frac{|K_j - \tilde{K}_j|}{K_j} \leq \frac{|K_j^w - \tilde{K}_j^w|}{K_j^w} + \sum_{i=1}^{d_z} \frac{|K_{ji}^z - \tilde{K}_{ji}^z|}{K_{ji}^z}, \text{ where}$$

$$K_j^w = \prod_{i=1}^{d_w} \exp\left\{-\frac{(w_i - \mu_{ji}^w)^2}{2(\sigma_i^z)^2}\right\} \text{ and } K_{ji}^z = \int_{A_{z_i}} \phi_i(\tilde{z}_i) d\tilde{z}_i.$$

The proof of Corrolary 5.1 in Norets and Pati (2017) delivers that

$$\int \frac{|K_j^w - \tilde{K}_j^w|}{K_j^w} g_0(w) dw \leq \frac{\epsilon}{24}$$

For K_{ji}^z we consider two separate cases. First, if $A_{z_i} \subset [0,1]$, then we show that

$$\frac{|K_{ji}^z - \tilde{K}_{ji}^z|}{K_{ji}^z} = \left|1 - \frac{\tilde{K}_{ji}^z}{K_{ji}^z}\right| = \sup_{\tilde{z}_i \in A_{z_i}} \left|1 - \frac{\phi(\tilde{z}_i, \tilde{\mu}_{ji}^z, \tilde{\sigma}_i)}{\phi(\tilde{z}_i, \mu_{ji}^z, \sigma_i)}\right| \leq \frac{\epsilon}{24 d_z}.$$

To obtain the above result note that for any $\tilde{z}_i \in [0,1]$

$$\begin{aligned}
\left|1 - \frac{\phi(\tilde{z}_i, \tilde{\mu}_{ji}^z, \tilde{\sigma}_i)}{\phi(\tilde{z}_i, \mu_{ji}^z, \sigma_i)}\right| &\leq \left|1 - \frac{\sigma_i}{\tilde{\sigma}_i}\right| + \frac{\sigma_i}{\tilde{\sigma}_i}\left|1 - \exp\left\{\frac{(\tilde{z}_i - \mu_{ji}^z)^2}{2\sigma_i^2} - \frac{(\tilde{z}_i - \tilde{\mu}_{ji}^z)^2}{2\tilde{\sigma}_i^2}\right\}\right| \\
&\leq 2\frac{\tilde{\sigma}_i - \sigma_i}{\sigma_i} + 4\left|\frac{(\tilde{z}_i - \mu_{ji}^z)^2}{2\sigma_i^2} - \frac{(\tilde{z}_i - \tilde{\mu}_{ji}^z)^2}{2\tilde{\sigma}_i^2}\right| \\
&\leq 2\frac{\tilde{\sigma}_i - \sigma_i}{\sigma_i} + 4\left|\frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\tilde{\sigma}_i^2}\right)(\tilde{z}_i - \mu_{ji}^z)^2 + \frac{1}{2\tilde{\sigma}_i^2}\left((\tilde{z}_i - \mu_{ji}^z)^2 - (\tilde{z}_i - \tilde{\mu}_{ji}^z)^2\right)\right| \\
&\leq 2\frac{\tilde{\sigma}_i - \sigma_i}{\sigma_i} + 4\left|\frac{\frac{\sigma_i - \tilde{\sigma}_i}{\sigma_i}}{\underline{\sigma}^2}\right| \left|\tilde{z}_i - \mu_{ji}^z\right|^2 + \frac{4}{2\underline{\sigma}^2}\left(\left|(\mu_{ji}^z)^2 - (\tilde{\mu}_{ji}^z)^2\right| + 2|\tilde{z}_i|\left|\mu_{ji}^z - \tilde{\mu}_{ji}^z\right|\right) \\
&\leq 2\frac{\underline{\sigma}^2\epsilon}{768\overline{\mu}^2 \max\{d_x + d_y, d_z + d_w\}} \\
&\quad + 4\left(\frac{\underline{\sigma}^2\epsilon}{768\overline{\mu}^2 \max\{d_x + d_y, d_z + d_w\}}\frac{4\bar{\mu}^2}{\underline{\sigma}^2} + \frac{1}{\underline{\sigma}^2}\left(\left|\mu_{ji}^z - \tilde{\mu}_{ji}^z\right|(\bar{\mu} + 1)\right)\right) \\
&\leq \frac{\epsilon}{386 d_z} + \frac{\epsilon}{92 d_z} + \frac{8\bar{\mu}}{\underline{\sigma}^2}\frac{\underline{\sigma}^2\epsilon}{384\overline{\mu} d_z} < \frac{\epsilon}{24 d_z},
\end{aligned}$$

where we have used that $\sigma_i/\tilde{\sigma}_i \leq 2$ and that $|1 - e^x| \leq 2|x|$ for $|x| < 1$.

Second, let, without loss of generality, $A_{z_i} = [1 - 1/2N_i, +\infty)$ and let $a = (1 - 1/2N_i - \mu_{ji}^z)/\sigma_i$ and $\tilde{a} = (1 - 1/2N_i - \tilde{\mu}_{ji}^z)/\tilde{\sigma}_i$. Also, suppose, without loss of generality, that $\tilde{a} > a$. Then

$$\left|1 - \frac{\tilde{K}_{ji}^z}{K_{ji}^z}\right| = \frac{\int_a^{\tilde{a}} \phi(t,0,1)dt}{\int_{\tilde{a}}^{\infty} \phi(t,0,1)dt} = \frac{|\tilde{a} - a|\phi(\tilde{\tilde{a}},0,1)}{\int_{\tilde{a}}^{\infty} \phi(t,0,1)dt}$$

for some $\tilde{\tilde{a}} \in [a, \tilde{a}]$ by the mean value theorem. For $\tilde{a} < 1$

$$\frac{|\tilde{a} - a|\phi(\tilde{\tilde{a}},0,1)}{\int_{\tilde{a}}^{\infty} \phi(t,0,1)dt} \leq \frac{|\tilde{a} - a|}{\sqrt{2\pi}(1 - \Phi(1))}.$$

For $\tilde{a} \geq 1$

$$\frac{|\tilde{a}-a|\phi(\tilde{\tilde{a}},0,1)}{\int_{\tilde{a}}^{\infty}\phi(t,0,1)dt} \leq \frac{|\tilde{a}-a|\phi(\tilde{\tilde{a}},0,1)}{\phi(\tilde{a},0,1)}\left(\tilde{a}+\sqrt{\tilde{a}+4}\right) \leq |\tilde{a}-a|4\tilde{a}\frac{\phi(\tilde{\tilde{a}},0,1)}{\phi(\tilde{a},0,1)}.$$

Note that $\tilde{a} \leq 2\bar{\mu}/\underline{\sigma}$ and that

$$\frac{\phi(\tilde{\tilde{a}},0,1)}{\phi(\tilde{a},0,1)} = \frac{\phi(\tilde{\tilde{\mu}}_{ji}^{z},1-1/2N_i,\tilde{\sigma}_i)}{\phi(\tilde{\mu}_{ji}^{z},1-1/2N_i,\tilde{\sigma}_i)} \leq \exp\{\frac{\epsilon}{24}\} \leq 2$$

for some $\tilde{\tilde{\mu}}_{ji}^{z} \in [\tilde{\mu}_{ji}^{z}, \mu_{ji}^{z}]$ using the result in Equation (4.2) from Norets and Pati (2017). In both cases we find that

$$\left|1-\frac{\tilde{K}_{ji}^{z}}{K_{ji}^{z}}\right| = \frac{|\tilde{a}-a|\phi(\tilde{\tilde{a}},0,1)}{\int_{\tilde{a}}^{\infty}\phi(t,0,1)dt} \leq |\tilde{a}-a|8\frac{\bar{\mu}}{\underline{\sigma}}.$$

Furthermore,

$$\begin{aligned}
|\tilde{a}-a| &\leq \left|(1-\frac{1}{2N}-\mu_{ji}^{z})\frac{\tilde{\sigma}_i-\sigma_i}{\tilde{\sigma}_i\sigma_i}\right| + \left|\frac{\mu_{ji}^{z}-\tilde{\mu}_{ji}^{z}}{\tilde{\sigma}_i}\right| \leq \frac{2\bar{\mu}}{\underline{\sigma}}\frac{\tilde{\sigma}_i-\sigma_i}{\sigma_i} + \frac{|\mu_{ji}^{z}-\tilde{\mu}_{ji}^{z}|}{\underline{\sigma}} \\
&\leq \frac{\underline{\sigma}\epsilon}{384\bar{\mu}\max\{d_x+d_y,d_z+d_w\}} + \frac{\underline{\sigma}\epsilon}{384\overline{\mu}d_z} \leq \frac{\underline{\sigma}\epsilon}{192\bar{\mu}\max\{d_x+d_y,d_z+d_w\}}
\end{aligned}$$

and, therefore,

$$\left|1-\frac{\tilde{K}_{ji}^{z}}{K_{ji}^{z}}\right| \leq |\tilde{a}-a|8\frac{\bar{\mu}}{\underline{\sigma}} \leq \frac{\epsilon}{24\max\{d_x+d_y,d_z+d_w\}} < \frac{\epsilon}{24d_z}.$$

Combining all the above results we obtain that $d_{TV}(p(y,x|z,w,\theta,m),p(y,z|z,w,\tilde{\theta},m)) \leq \epsilon$ as desired. This concludes the proof for the covering number.

The upper bound on $\Pi(\mathcal{F}^c)$ is obtained in the same way as in the proof of Theorem 4.1 in Norets and Pati (2017) with the only difference being that the dimension d appears in front of some of the terms in the bound due to coordinate specific scale parameters and slightly different choice of the prior tail condition (6.6).

□

LEMMA 8.6. *Consider* $\epsilon_n = (N_J/n)^{\beta_{Jc}/(2\beta_{Jc}+1)}(\log n)^{t_J}$ *and* $\tilde{\epsilon}_n = (N_J/n)^{\beta_{Jc}/(2\beta_{Jc}+1)}(\log n)^{\tilde{t}_J}$ *with* $t_J > \tilde{t}_J + \max\{0,(1-\tau_1)/2\}$ *and* $\tilde{t}_J > t_{J0}$*, where* t_{J0} *is defined in* (7.1)*. Define* $\mathcal{F}_n$ *as in* (8.2) *with* $\epsilon = \epsilon_n, H = n\epsilon_n^2/(\log n)$, $\underline{\alpha} = e^{-nH}$, $\underline{\sigma} = n^{-1/(2a_3)}$, $\overline{\sigma} = e^n$, *and* $\overline{\mu} = n^{1/\tau_3}$*. Then, for some constants* $c_1, c_3 > 0$ *and every* $c_2 > 0$, $\mathcal{F}_n$ *satisfies* (7.3) *and* (7.4) *for all large* n.

PROOF. From Lemma 8.5,

$$\log M_e(\epsilon_n,\mathcal{F}_n,\rho) \leq c_1 H \log n = c_1 n\epsilon_n^2.$$

Also,

$$\Pi(\mathcal{F}_n^c) \leq H^2 d \exp\{-a_{13} n\} + \exp\{-a_{10} H (\log H)^{\tau_1}\}$$
$$+ d a_1 \exp\{-a_2 n\} + d a_4 \exp\{-2 a_5 n\}.$$

Hence, $\Pi(\mathcal{F}_n^c) \leq e^{-(c_2+4)n\tilde{\epsilon}_n^2}$ for any c_2 if $\epsilon_n^2 (\log n)^{\tau_1 - 1}/\tilde{\epsilon}_n^2 \to \infty$, which holds for $t_J > \tilde{t}_J + \max\{0, (1-\tau_1)/2\}$.

$\square$

Economics Department,
Brown University, Providence, RI 02912
E-mail: andriy_norets@brown.edu

Institute for Advanced Studies Vienna,
Josefstaedter Strasse 39,
Vienna 1080, Austria E-mail: pelenis@ihs.ac.at