STATISTICA, anno LXXV, n. 1, 2015

EMPIRICAL BAYES CONDITIONAL DENSITY ESTIMATION

Catia Scricciolo [1]
Department of Decision Sciences, Bocconi University, Milan, Italy

1. Introduction

The problem of estimating the conditional density of a response, given a set of predictors, is classical and of primary importance in real data analysis since the conditional density provides a more comprehensive description of the association between the response and the predictors than, for instance, does the conditional expectation or regression function which can only capture partial aspects of it. The conditional density contains information on how the different features of the response distribution, like skewness, shape and so on change with the covariates. Conditional density estimation for predictive purposes have applications across different fields like economy, actuarial sciences and medicine.

Nonparametric estimation of a collection of conditional densities over a covariate space presents two main features: (a) the multivariate curve may have different regularity levels along different directions, (b) the function may depend only on a subset of the covariates. The goal is to estimate a multivariate function of the relevant predictors, while discarding the remaining ones, and obtain procedures that simultaneously adopt to the unknown dimension of the predictor and to the possibly anisotropic regularity of the function. Classical references on nonparametric conditional density estimation taking a frequentist approach are Efromovich (2007, 2010) and Hall *et al.* (2004); see also the recent contribution by Bertin *et al.* (2015). The problem of conditional density estimation has been studied taking a Bayesian nonparametric approach only recently and popular methods are based on generalized stick-breaking process mixture models for which supporting results, in terms of frequentist asymptotic properties of posterior distributions, have been given by Pati *et al.* (2013) and Norets and Pati (2014). The former article provides sufficient conditions for posterior consistency in conditional density estimation for a broad class of predictor-dependent mixtures of Gaussian kernels. The latter presents results on posterior contraction rates for conditional density estimation over classes of locally (isotropic) Hölder smooth densities using *finite* mixtures of Gaussian kernels, with covariate-dependent mixing weights having a special structure. The entailed density estimation procedure converges at a rate

[1] Corresponding Author. E-mail: catia.scricciolo@unibocconi.it

that automatically adapts to the unknown dimension of the set of relevant covariates, thus ultimately performing a dimension reduction, and to the regularity level of the sampling conditional density.

In this note, the focus is on defining procedures for conditional density estimation that attain minimax rates (up to log-factors) of posterior concentration adopting to both the dimension of the set of relevant covariates and to the regularity level of the function. We consider a procedure based on *infinite* mixtures of Gaussian kernels, with the same predictor-dependent mixing weights as in Norets and Pati (2014), and show that it can have a performance on par with that of the procedure proposed by the above cited authors in terms of rate adaptation to the predictor dimension and to the (isotropic) regularity level. Under the same set of assumptions on the data generating process and the prior law, the performance of the conditional density estimation procedure of an empirical Bayesian, who considers an automatic data-driven selection of the prior hyper-parameters, matches with that of an "honest" Bayesian. We deal in detail with the isotropic case; extension of the result to the anisotropic case follows along the same lines.

The organization of the article is as follows. Section 1.1 sets up the notation. Section 2 presents the main results on adaptive empirical Bayes posterior concentration at minimax-optimal L^1-rates (up to log-factors) for locally Hölder smooth conditional densities, with contextual adaptive dimension reduction in the presence of irrelevant covariates. Final remarks and comments are gathered in Section 3. The statement of a theorem invoked in the proof of the main result is reported in the Appendix for easy reference.

1.1. Notation

Let $\mathbb{N}_0 = \{0, 1, \dots\}$ be the set of non-negative integers and $\mathbb{R}_+$ that of positive real numbers. For any $a, b \in \mathbb{R}$, we denote by $a \wedge b$ their minimum and by $a \vee b$ their maximum. We write "$\lesssim$" and "$\gtrsim$" for inequalities valid up to a constant multiple which is universal or inessential for our purposes. For a generic sequence $\{a_n\}$, we use the notation $a_n = o(1)$ $(n \to \infty)$ to mean that $a_n \to 0$ as $n \to \infty$. For sequences $\{a_n\}$ and $\{b_n\}$, by writing $a_n = O(b_n)$ $(n \to \infty)$ we mean that $b_n \neq 0$ and there exists a constant $K > 0$ so that $|a_n/b_n| < K$ for every $n \in \mathbb{N}$.

For $d_x \in \mathbb{N}$, let $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ be the covariate space. For $d_y \in \mathbb{N}$, let $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ be the response space and, for $d := d_x + d_y$, let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d$ be the sample space.

For any $k \in \mathbb{N}$, if $E \subset \mathbb{R}^k$ and $x \in \mathbb{R}^k$, the translate of E is the set $E + x := \{z + x : z \in E\}$. If $\xi, \vartheta \in \mathbb{R}^k$, the Euclidean distance between ξ and ϑ is $\|\xi - \vartheta\| := \{\sum_{j=1}^k (\xi_j - \vartheta_j)^2\}^{1/2}$.

Let

$$\mathcal{F} := \left\{ f : \mathcal{Z} \to [0, \infty) \,\middle|\, \text{Borel-measurable and, } \forall\, x \in \mathcal{X},\ \int_{\mathcal{Y}} f(y|x)\mathrm{d}y = 1 \right\}$$

be the space of conditional probability densities with respect to Lebesgue measure m on $\mathcal{Y}$. The same symbol m will also be used to denote Lebesgue measure on $\mathcal{Z}$. A centered multivariate normal density with covariance matrix $\sigma^2 I$, for I the identity matrix whose dimension is clear from the context, is denoted by ϕ_σ. The symbol δ_z stands for point mass at z.

Let Q be a fixed probability measure on the measurable space $(\mathcal{X},\, \mathcal{B}(\mathcal{X}))$, with $\mathcal{B}(\mathcal{X})$ the Borel σ-field on $\mathcal{X}$, that possesses Lebesgue density q.

Given any real number $p \geq 1$ and Borel-measurable function $g : \mathcal{Z} \to \mathbb{R}$, for every $x \in \mathcal{X}$, we introduce the notation $\|g\|_{p,x} := (\int_{\mathcal{Y}} |g(x,\, y)|^p \mathrm{d}y)^{1/p}$ that is useful to define global distances between conditional densities. For any pair of (conditional) densities $f_1,\, f_2 \in \mathcal{F}$, let the q-integrated L^1-distance be defined as $\|f_2 - f_1\|_1 := \int_{\mathcal{X}} \|f_2 - f_1\|_{1,x} q(x)\mathrm{d}x$ and, analogously, the squared q-integrated Hellinger distance as $h^2(f_2,\, f_1) := \int_{\mathcal{X}} \|f_2^{1/2} - f_1^{1/2}\|_{2,x}^2 q(x)\mathrm{d}x$. For (conditional) densities $f,\, f_0 \in \mathcal{F}$, the q-integrated Kullback-Leibler divergence of f from f_0 is defined as $\mathrm{KL}(f_0;\, f) := \int_{\mathcal{X}\times\mathcal{Y}} f_0 q \log(f_0 q/fq)\mathrm{d}m$, m being here the Lebesgue measure on $\mathcal{Z}$, which coincides with the Kullback-Leibler divergence of fq from $f_0 q$. Analogously, the q-integrated second moment of $\log(f_0 q/fq)$ is defined as $\mathrm{V}_2(f_0;\, f) := \int_{\mathcal{X}\times\mathcal{Y}} f_0 q |\log(f_0 q/fq)|^2 \mathrm{d}m$ and coincides with the second moment of $\log(f_0 q/fq)$ with respect to $f_0 q$.

The ϵ-covering number of a semi-metric space $(M,\, d)$, denoted by $N(\epsilon,\, M,\, d)$, is the minimal number of d-balls of radius ϵ needed to cover the set M.

2. Main Results

Let $Z^{(n)} = (Z_1,\, \ldots,\, Z_n)$ be a random sample of independent and identically distributed (i.i.d.) observations $Z_i = (X_i,\, Y_i) \in \mathcal{Z}$, $i = 1,\, \ldots,\, n$, from a probability measure P_0 on the measurable space $(\mathcal{Z},\, \mathcal{B}(\mathcal{Z}))$, where $\mathcal{B}(\mathcal{Z})$ is the Borel σ-field on $\mathcal{Z}$, that possesses Lebesgue density $f_0 q$ that is referred to as the *true joint data generating density*, with $f_0 \in \mathcal{F}$ the conditional density of the response Y, given the predictor X, and q the marginal density of X, called the *design density*, which is fixed and, for theoretical investigation, does not need to be known or estimated. The problem is to estimate the conditional density f_0 when no parametric assumption is formulated on it, taking an empirical Bayes approach that employs an automatic data-driven selection of the prior hyper-parameters. For a recent overview of empirical Bayes methods, the reader may refer to Petrone *et al.* (2014a). Even if the proposed empirical Bayes procedure simultaneously leads to adaptation with respect to both aspects (a) and (b) illustrated in the Introduction, the two issues are treated separately for ease of exposition: we first deal with adaptive estimation over classes of locally Hölder smooth conditional densities when the dimension of the predictor is correctly specified and then show adaptive dimension reduction in the case where fewer covariates are relevant. Adaptive dimension reduction clearly plays a key role in view of the curse of dimensionality. In Section 2.2, it is shown that, when the response is independent of (some of) the covariates introduced in the model, the empirical Bayes posterior asymptotically performs a dimension reduction, thus contracting at a rate that results from the combination of the dimension of the subset of relevant explanatory variables and the possibly anisotropic regularity level of the curve as a function of the selected covariates.

2.1. Empirical Bayes posterior concentration for conditional density estimation

In this section, we consider empirical Bayes posterior contraction rates for estimating conditional densities when the dimension of the predictor is correctly specified.

Prior law specification. A prior distribution can be induced on the space $\mathcal{F}$ of conditional densities by a law $\Pi_{\mathcal{X}}$ on a collection of mixing probability measures $\mathcal{M}_{\mathcal{X}} = \{P_x \in \mathcal{M}(\Theta),\, x \in \mathcal{X}\}$, where $\mathcal{M}(\Theta)$ denotes the space of all probability measures on some subset $\Theta \subseteq \mathcal{Y}$, using a mixture of d_y-dimensional Gaussian kernels to model the conditional density

$$f(\cdot|x) = (F_x * \phi_\sigma)(\cdot) = \int_\Theta \phi_\sigma(\cdot - \theta)\mathrm{d}F_x(\theta), \quad x \in \mathcal{X},$$

where, for every $x \in \mathcal{X}$, F_x is the cumulative distribution function corresponding to a probability measure P_x which is assumed to be (almost surely) discrete

$$P_x = \sum_{j=1}^\infty p_j(x)\delta_{\theta_j(x)},$$

with random weights $p_j(x) \geq 0$, $j \in \mathbb{N}$, such that $\sum_{j=1}^\infty p_j(x) = 1$ almost surely, and random support points $\{\theta_j(x)\}$ that are i.i.d. replicates drawn from a probability measure G_x on Θ. Following Pati *et al.* (2013), we single out two relevant special cases.

- *Predictor-dependent mixtures of Gaussian linear regressions* (MGLR$_x$): the conditional density is modeled as a mixture of Gaussian linear regressions

$$f(\cdot|x) = \int_{\mathbb{R}^{d_x}} \phi_\sigma(\cdot - \beta' x)\mathrm{d}F_x(\beta), \quad x \in \mathcal{X},$$

  where $\beta' x$ denotes the usual inner product in $\mathbb{R}^{d_x}$ and the mixing measure P_x corresponding to F_x is such that $P_x = \sum_{j=1}^\infty p_j(x)\delta_{\beta_j}$ almost surely, with the vectors of regression coefficients $\beta_j \overset{\text{iid}}{\sim} G$. For a particular structure of the random weights $p_j(x)$'s, probit stick-breaking mixtures of Gaussian kernels are obtained. Probit transformation of Gaussian processes for constructing the stick-breaking weights has been considered in Rodríguez and Dunson (2011), who exhibit applications to real data of the probit stick-breaking process model.

- *Gaussian mixtures of fixed-p dependent processes*: if $p_j(x) \equiv p_j$ for all $x \in \mathcal{X}$, we obtain mixtures of Gaussian kernels with fixed weights. Versions of fixed-p dependent Dirichlet process mixtures of Gaussian densities (fixed p-DDP) have been applied to ANOVA, survival analysis and spatial modeling.

We consider a variant of the prior proposed in Norets and Pati (2014). Let ν be a probability measure on $\mathcal{X}$ and G a probability measure on $\mathcal{Y}$. For $(\lambda,\, \tau) \in \mathcal{Y} \times \mathbb{R}_+$, with abuse of notation, let $G_\tau(\cdot - \lambda)$ denote the probability measure on $\mathcal{Y}$ with

Lebesgue density $\tau^{-1}(\mathrm{d}G/\mathrm{d}m)((\cdot-\lambda)/\tau)$. Given $(\mu_j^x,\, \mu_j^y) \in \mathscr{Z}$, $j \in \mathbb{N}$, and $\sigma \in \mathbb{R}_+$, for every $x \in \mathcal{X}$, let

$$p_{j,\sigma}(x) := \frac{p_j \phi_\sigma(x-\mu_j^x)}{\sum_{q=1}^{\infty} p_q \phi_\sigma(x-\mu_q^x)}, \quad j \in \mathbb{N}. \tag{1}$$

We propose the following prior specification:

$$\begin{aligned} Y_i|(X_i = x_i),\, (F_x)_{x\in\mathcal{X}},\, \sigma \,&\sim\, (F_{x_i} * \phi_\sigma)(\cdot) = \sum_{j=1}^{\infty} p_{j,\sigma}(x_i)\phi_\sigma(\cdot - \mu_j^y), \\ \sum_{j=1}^{\infty} p_j \delta_{(\mu_j^x,\, \mu_j^y)} \,&\sim\, \mathrm{DP}(c_0\nu \times G_\tau(\cdot - \lambda)) \text{ independent of } \sigma \sim \mathrm{IG}(\alpha,\, \beta), \end{aligned}$$

where $c_0 \in \mathbb{R}_+$ is a finite constant and $\alpha,\, \beta \in \mathbb{R}_+$ are the shape and scale parameters of an inverse-gamma prior distribution, respectively. In this case, F_x corresponds to the probability measure $P_x = \sum_{j=1}^{\infty} p_{j,\sigma}(x)\delta_{\mu_j^y}$. For later use, note that, defined the mapping $g : x \mapsto \sum_{q=1}^{\infty} p_q\phi_\sigma(x - \mu_q^x)$ and modeled the conditional density f as $\sum_{j=1}^{\infty} p_{j,\sigma}(x)\phi_\sigma(\cdot - \mu_j^y)$, the density product fg is a mixture of d-dimensional Gaussian densities

$$f(y|x)g(x) = \sum_{j=1}^{\infty} p_j \phi_\sigma(x - \mu_j^x)\phi_\sigma(y - \mu_j^y). \tag{2}$$

By the stick-breaking representation of a Dirichlet process (DP), the random weights $p_j = V_j \prod_{h=1}^{j-1}(1 - V_h)$, $j \in \mathbb{N}$, with $V_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1,\, c_0)$, and the locations $\mu_j^y \overset{\text{iid}}{\sim} G_\tau(\cdot - \lambda)$. The last assertion is equivalent to $\mu_j^y = \lambda + \zeta_j$, with $\zeta_j \overset{\text{iid}}{\sim} \tau^{-1}(\mathrm{d}G/\mathrm{d}m)(\cdot/\tau)$, $j \in \mathbb{N}$. The overall prior can be rewritten as

$$\begin{aligned} Y_i|(X_i = x_i),\, (F_x)_{x\in\mathcal{X}},\, \sigma \,&\sim\, \sum_{j=1}^{\infty} p_{j,\sigma}(x_i)\phi_\sigma(\cdot - \lambda - \zeta_j) \\ \sum_{j=1}^{\infty} p_j \delta_{(\mu_j^x,\, \zeta_j)} \,&\sim\, \mathrm{DP}(c_0\nu \times G_\tau) \text{ independent of } \sigma \sim \mathrm{IG}(\alpha,\, \beta). \end{aligned} \tag{3}$$

For the vector $\gamma = (\beta,\, \lambda,\, \tau^2)$ of prior hyper-parameters, let Π_γ stand for the product prior law $\mathrm{DP}(c_0\nu \times G_\tau(\cdot - \lambda)) \times \mathrm{IG}(\alpha,\, \beta)$. Let $\Pi_\gamma(B|Z^{(n)})$ denote the posterior probability of any Borel set B of $(\mathcal{F},\, d)$, where d can be either the q-integrated Hellinger or L^1-distance. For any estimator $\hat{\gamma}_n = (\hat{\beta}_n,\, \hat{\lambda}_n,\, \hat{\tau}_n^2)$ of γ based on $Z^{(n)}$, the empirical Bayes posterior law $\Pi_{\hat{\gamma}_n}(\cdot|Z^{(n)})$ is obtained by plugging $\hat{\gamma}_n$ into the posterior distribution

$$\Pi_{\hat{\gamma}_n}(\cdot|Z^{(n)}) = \Pi_\gamma(\cdot|Z^{(n)})|_{\gamma=\hat{\gamma}_n}.$$

We study empirical Bayes posterior concentration rates relative to d at an ordinary smooth conditional density f_0, namely, we assess the order of magnitude of the radius $M\epsilon_n$ of a shrinking ball centered at f_0 so that

$$P_0^n \Pi_{\hat{\gamma}_n}(f \in \mathcal{F} :\, d(f,\, f_0) > M\epsilon_n|Z^{(n)}) \to 0, \tag{4}$$

where $P_0^n\varphi$ is used to abbreviate expectation $\int_{\mathcal{Z}^n}\varphi \mathrm{d}P_0^n$ under the n-fold product measure P_0^n. We consider the case where the true conditional density f_0, regarded as a mapping from $\mathcal{Z}$ to $\mathbb{R}_+\cup\{0\}$, satisfies a Hölder condition in the sense of the following definition, for which we introduce some more notation. For any $\beta\in\mathbb{R}_+$, let $\langle\beta\rangle:=\max\{i\in\mathbb{N}_0:\, i<\beta\}$ be the largest non-negative integer strictly smaller than β. For a d-dimensional multi-index $k=(k_1,\,\ldots,\,k_d)\in\mathbb{N}_0^d$, define $k.=k_1+\ldots+k_d$ and let D^k denote the mixed partial derivative operator $\partial^{k.}/\partial z_1^{k_1}\ldots\partial z_d^{k_d}$.

Definition 1. *For any $\beta\in\mathbb{R}_+$, $\tau\geq 0$ and function $L:\,\mathcal{Z}\to\mathbb{R}_+\cup\{0\}$, let the class $C^{\beta,L,\tau}(\mathcal{Z})$ consist of functions $f:\,\mathcal{Z}\to\mathbb{R}$ that have finite mixed partial derivatives D^kf of all orders $k.\leq\langle\beta\rangle$ and, for every $k\in\mathbb{N}_0^d$ such that $k.=\langle\beta\rangle$, the mixed partial derivatives of order k are locally (uniformly) Hölder continuous with exponent $\beta-\langle\beta\rangle$ in $\mathcal{Z}$ with envelope L,*

$$|(D^kf)(z+\Delta)-(D^kf)(z)|\leq L(z)e^{\tau\|\Delta\|^2}\|\Delta\|^{\beta-\langle\beta\rangle},\qquad\forall\, z,\,\Delta\in\mathcal{Z}.\tag{5}$$

This function class has been previously considered by Shen *et al.* (2013), who constructively showed that Lebesgue probability density functions in $C^{\beta,L,\tau}(\mathbb{R}^d)$ satisfying additional regularity conditions can be approximated by convolutions with the Gaussian kernel ϕ_σ with an L^1-error of the order σ^β. The construction of the mixing density in the approximation can be viewed as a multivariate extension of the results in Kruijer *et al.* (2010, § 3), the main difference being that condition (5) is weaker than the one employed in Kruijer *et al.* (2010), where it is assumed that $\log f_0\in C^{\beta,L,0}(\mathbb{R})$.

If ϵ_n is (an upper bound on) the posterior contraction rate and the convergence in (4) is at least as fast as ϵ_n^2, then ϵ_n is (an upper bound on) the rate of convergence relative to d of the estimator $\hat{f}_n(\cdot|x)=\int_{\mathcal{F}}f(\cdot|x)\Pi_{\hat{\gamma}_n}(\mathrm{d}f|Z^{(n)})$. Since the convergence rate of an estimator cannot be faster than the minimax rate over the density function class considered, the posterior contraction rate cannot be faster than the minimax rate. So, if the posterior distribution achieves the minimax rate, then also $\{\hat{f}_n(\cdot|x)\}_{x\in\mathcal{X}}$ has minimax-optimal convergence rate and is adaptive.

In order to state the main result on empirical Bayes posterior contraction rates at locally Hölder smooth densities, we report below the assumptions on the "true" joint data generating density f_0q and the prior law Π_γ.

2.1.1. Assumptions on the joint data generating density and on the prior law

Assumptions on f_0q

(i) $\mathcal{X}=[0,\,1]^{d_x}$;

(ii) q is bounded;

(iii) $f_0\in C^{\beta,L,\tau}(\mathcal{Z})$. For some $\eta\in\mathbb{R}_+$, $\int_{\mathcal{Z}}(|L|/f_0)^{2+\eta/\beta}f_0\mathrm{d}m<\infty$ and

$$\int_{\mathcal{Z}}(|D^kf_0|/f_0)^{(2\beta+\eta)/k}f_0\mathrm{d}m<\infty\qquad\text{for all }k.\leq\langle\beta\rangle;$$

(iv) there exist constants $B_0,\,\tau\in\mathbb{R}_+$ such that, for every $x\in\mathcal{X}$,

$$f_0(y|x)\lesssim\exp\left(-B_0\|y\|^\tau\right)\qquad\text{for large }\|y\|.$$

Assumption on Π_γ

(v) the base probability measure $\nu\times G$ of the Dirichlet process possesses Lebesgue density and there exist constants p, $C_0 \in \mathbb{R}_+$ so that

$$\frac{\mathrm{d}G}{\mathrm{d}m}(y) \propto \exp\left(-C_0\|y\|^p\right) \qquad \text{for large } \|y\|.$$

Assumption (ii) is verified as soon as the design density is continuous on the closed unit interval, see the comments following the statement of Theorem 2 concerning its role in the proof. Assumption (iii) requires Hölder type regularity of f_0 in addition to integrability conditions, which jointly with assumption (iv), are used to approximate $f_0\mathbf{1}_{\mathcal{X}}$ with a finite d-dimensional Gaussian mixture having a sufficiently restricted number of support points, see Theorem 3, Proposition 1 and Theorem 4 of Shen *et al.* (2013).

We now state the main result.

Theorem 2. *Suppose there exists a set* $K_n \subset \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+$ *such that* $P_0^n(\hat{\gamma}_n \in K_n^c) = o(1)$. *Under assumptions* ($i$)*-*($v$)*, the empirical Bayes posterior distribution corresponding to the prior in* (3) *contracts at a rate* $\epsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t$ *for a suitable constant* $t > 0$.

We give a few comments on Theorem 2 before presenting its proof. The empirical Bayes posterior distribution corresponding to the prior described in (3) contracts at a rate $n^{-\beta/(2\beta+d)}(\log n)^t$ which differs from the minimax L^1-rate attached to the class of locally Hölder densities $C^{\beta,L,\tau}(\mathcal{Z})$ for at most a logarithmic factor. The quality of the estimation improves with increasing regularity level β and deteriorates with increasing dimension d. Furthermore, the rate automatically adapts to the unknown regularity level β of the "true" conditional density f_0, whatever $\beta \in \mathbb{R}_+$, see, *e.g.*, Scricciolo (2015) for an overview of the main schemes for Bayesian adaptation. This implies existence of empirical Bayes procedures for conditional density estimation that attain minimax-optimal rates, up to logarithmic terms, over the full scale of locally Hölder densities and perform as well as adaptive Bayesian procedures like the one entailed by the hierarchical prior of finite Dirichlet mixtures of Gaussian densities proposed by Norets and Pati (2014).

The problem presents two main difficulties:

(a) data-dependence of the prior law due to an automatic data-driven selection of the prior hyper-parameters;

(b) dependence of f_0 on the covariates, which gives account for dependence of the convergence rate on the dimension d of the sample space $\mathcal{Z}$.

Concerning (a), data-dependence of the prior can be dealt with resorting to the same key idea as in Petrone *et al.* (2014b) and Donnet *et al.* (2014), which is based on a prior measure change aimed at transferring data-dependence from the prior law to the likelihood, as long as a parameter transformation can be identified.

Concerning (b), dependence of f_0 on the covariates can be dealt with regarding f_0 as a d-multivariate *joint* density with respect to *Lebesgue* measure on $[0,\,1]^{d_x} \times \mathcal{Y}$. Indeed, f_0 is a *joint* density, but with respect to the measure $Q \times m$ on $\mathcal{Z}$, which prevents immediate use of Gaussian mixtures for its approximation. A device due to Norets and Pati (2014) based on the inequality

$$h(f,\, f_0) \lesssim \|(fg)^{1/2} - (f_0\mathbf{1}_{[0,\,1]^{d_x}})^{1/2}\|_2,$$

which relates the q-integrated Hellinger distance between the conditional densities f and f_0 to the Hellinger distance between the joint densities fg and $f_0\mathbf{1}_{[0,\,1]^{d_x}}$, where $f(y|x)g(x) = \sum_{j=1}^{\infty} p_j \phi_\sigma(x - \mu_j^x)\phi_\sigma(y - \mu_j^y)$ by virtue of equality (2), takes advantage of the special structure of the mixing weights $p_{j,\sigma}(x)$ in model (1) for the conditional density f to approximate the joint Lebesgue density $f_0 1_{[0,\,1]^{d_x}}$ by mixtures of d-dimensional Gaussian densities. Thus, the problem of approximating the "true" joint data generating density $f_0 q$ with fq is translated into the problem of approximating $f_0 1_{[0,\,1]^{d_x}}$ with mixtures of d-dimensional Gaussian densities.

Proof. We appeal to Theorem 5 reported in the Appendix which is an adapted version of Theorem 1 in Donnet *et al.* (2014).

We first define the parameter transformation for the change of prior law. For sequences $\underline{b}_n \downarrow 0$, $\bar{b}_n \uparrow \infty$, $\underline{l}_n \downarrow -\infty$, $\bar{l}_n \uparrow \infty$, $\underline{t}_n \downarrow 0$ and $\bar{t}_n \uparrow \infty$, consider a set $K_n = [\underline{b}_n,\, \bar{b}_n) \times [\underline{l}_n,\, \bar{l}_n) \times [\underline{t}_n^2,\, \bar{t}_n^2) \subseteq \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+$ such that $P_0^n(\hat{\gamma}_n \in K_n^c) = o(1)$. For a sequence $u_n \downarrow 0$ to be suitably defined later on, consider a u_n-covering of K_n by Euclidean open balls of radius u_n. To the aim, let $v_n,\, w_n,\, z_n$ be positive infinitesimal sequences to be chosen as later on prescribed. Consider

- a covering of $[\underline{b}_n,\, \bar{b}_n)$ with intervals $B_r = [b_r,\, b_{r+1})$, where $b_r := \underline{b}_n(1 + z_n)^{r-1}$ for $r = 1,\, \ldots,\, \lceil \log(\bar{b}_n/\underline{b}_n)/\log(1 + z_n)\rceil$,
- a v_n-covering of $[\underline{l}_n,\, \bar{l}_n)$ with intervals $L_k = [l_k,\, l_{k+1})$, where $l_k := \underline{l}_n + (k-1)v_n$ for $k = 1,\, \ldots,\, \lceil(\bar{l}_n - \underline{l}_n)/v_n + 1\rceil$,
- a covering of $[\underline{t}_n^2,\, \bar{t}_n^2]$ with intervals $T_s = [t_s^2,\, t_{s+1}^2)$, where $t_s^2 := \underline{t}_n^2(1 + w_n)^{s-1}$ for $s = 1,\, \ldots,\, \lceil 2\log(\bar{t}_n/\underline{t}_n)/\log(1 + w_n)\rceil$.

For any $b \in B_r$, let $\pi_r := b/b_r$. We have $1 \le \pi_r < 1 + z_n$. For any $t^2 \in T_s$, let $\rho_s := (t^2/t_s^2)^{1/2}$. We have $1 \le \rho_s < (1 + w_n)^{1/2}$. Fix $\gamma' = (b_r,\, l_k,\, t_s^2)$. For any $\gamma = (b,\, l,\, t^2) \in B_r \times L_k \times T_s$, the Euclidean distance $\|\gamma - \gamma'\| = [(b - b_r)^2 + (l - l_k)^2 + (t^2 - t_s^2)^2]^{1/2} \le [(1 + z_n)^2 z_n^2 \bar{b}_n^2 + v_n^2 + (1 + w_n)^2 w_n^2 \bar{t}_n^4]^{1/2} =: u_n$. In order to have $u_n = o(1)$, it suffices that $w_n = o(\bar{t}_n^{-2})$ and $z_n = o(\bar{b}_n^{-1})$. The u_n-covering number N_n of K_n relative to the Euclidean distance is

$$N_n = O\left(\frac{\log(\bar{b}_n/\underline{b}_n)}{\log(1 + z_n)} \times \frac{\bar{l}_n - \underline{l}_n}{v_n} \times \frac{\log(\bar{t}_n/\underline{t}_n)}{\log(1 + w_n)}\right),$$

with $v_n,\, w_n,\, z_n$ that need to be chosen so that $N_n = o(e^{n\epsilon_n^2})$ as postulated by requirement [**A1**].

Fix $\gamma' = (b_r,\, l_k,\, t_s^2) \in B_r \times L_k \times T_s$ and consider any $\gamma = (b,\, l,\, t^2) \in B_r \times L_k \times T_s$. If $\sigma' \sim \mathrm{IG}(\alpha,\, b_r)$ then $\pi_r\sigma' \sim \mathrm{IG}(\alpha,\, b)$. For $z_j' = (\mu_j^x,\, \zeta_j')$, if $F' =$

$\sum_{j=1}^{\infty} p_j \delta_{z_j'} \sim \mathrm{DP}(c_0\nu \times G_{t_s})$ then $F = \sum_{j=1}^{\infty} p_j \delta_{(\mu_j^x,\, l+\rho_s\zeta_j')} \sim \mathrm{DP}(c_0\nu \times G_t(\cdot - l))$, where l denotes a d_y-dimensional vector with components all equal to l. Throughout, we use the same symbol l to denote either the scalar or the vector, the correct interpretation being clear from the context.

Let $\theta = (F,\, \sigma)$. For every $x \in \mathcal{X}$, let $f_\theta(\cdot|x) = \sum_{j=1}^{\infty} p_{j,\sigma}(x)\phi_\sigma(\cdot - \mu_j^y)$. The transformation $\psi_{\gamma',\gamma}(\theta)$ gives rise to the following density

$$f_{\psi_{\gamma',\gamma}(\theta)}(\cdot|x) = \sum_{j=1}^{\infty} p_{j,\pi_r\sigma'}(x)\phi_{\pi_r\sigma'}(\cdot - l - \rho_s\zeta_j').$$

We now identify a set B_n such that

$$\inf_{\gamma\in K_n} \Pi_\gamma(B_n) \gtrsim e^{-Cn\epsilon_n^2} \tag{6}$$

for some constant $C > 0$. Preliminarily, note that, by Lemma 7.1 of Norets and Pati (2014), in virtue of assumption (ii), the squared q-integrated Hellinger distance between f_θ and f_0 can be thus bounded above:

$$h^2(f_\theta,\, f_0) \leq 4\|q\|_\infty \, \|(f_\theta g)^{1/2} - (f_0\mathbf{1}_{\mathcal{X}})^{1/2}\|_2^2,$$

where $\|q\|_\infty := \sup_{x\in\mathcal{X}} q(x)$ and the Lebesgue density g is such that $f_\theta(y|x)g(x) = \sum_{j=1}^{\infty} p_j\phi_{\mu_j,\sigma}(x,\, y)$, that is, $g(x) = \sum_{q=1}^{\infty} p_q\phi_{\mu_q^x,\sigma}(x)$. This allows us to use d-dimensional Gaussian mixtures $\sum_{j=1}^{\infty} p_j\phi_{\mu_j,\sigma}(x,\, y)$ to approximate the density $f_0(y|x)\mathbf{1}_{\mathcal{X}}(x)$ defined on $\mathcal{Z}$. The set B_n is the same as the one described in Theorem 3.1 of Norets and Pati (2014). Let $\sigma_n = (\epsilon_n|\log\epsilon_n|^{-1})^{1/\beta}$ and $a_{\sigma_n} = a_0|\log\sigma_n|^{1/\tau}$, with $a_0 = [(8\beta + 4\eta + 16)/(B_0\delta)]^{1/\tau}$ for a sufficiently small $\delta > 0$. Find $b_1 > \max\{1,\, 1/(2\beta)\}$ so that $\epsilon_n^{b_1}|\log\epsilon_n|^{5/4} < \epsilon_n$. As in the proof of Theorem 3.1 in Norets and Pati (2014), which is an adaptation of that of Theorem 4 in Shen *et al.* (2013), the following facts hold. First, there exists a partition $U_1,\, \ldots,\, U_K$ of $\{z \in \mathcal{Z} :\, \|z\| \leq a_{\sigma_n}\}$ such that, for $j = 1,\, \ldots,\, N$, with $1 \leq N < K$, the ball U_j is centered at $z_j = (x_j,\, y_j)$ and has diameter $\sigma_n\epsilon_n^{2b_1}$, while, for $j = N{+}1,\, \ldots,\, K$, each set U_j has diameter bounded above by σ_n. This can be realized with $1 \leq N < K = O(\sigma_n^{-d}|\log\epsilon_n|^{d(1+1/\tau)})$. Further extend this to a partition $U_1,\, \ldots,\, U_M$ of $\mathbb{R}^d$, for $M = O(\epsilon_n^{-d/\beta}|\log\epsilon_n|^{ds})$, with $s = 1 + 1/\beta + 1/\tau$, such that $1 \geq \inf_{(l,\, t)\in K_n}(c_0\nu \times G_t(\cdot - l))(U_j) \gtrsim (\sigma_n\epsilon_n^{2b_1})^d$ for all $j = 1,\, \ldots,\, M$, provided that $\bar{l}_n = O(a_{\sigma_n})$, $\bar{t}_n = O(a_{\sigma_n}^p)$ and $a_{\sigma_n} = O(\underline{t}_n|\log\epsilon_n|^{1/p})$. Second, by virtue of assumptions (iii) and (iv), there exists $\theta^* = (F^*,\, \sigma_n)$, where $F^* = \sum_{j=1}^{N} p_j^*\delta_{\mu_j^*}$, with $\mu_j^* = z_j$ for $j = 1,\, \ldots,\, N$, so that $f_{\theta*}(y|x)g(x) = \sum_{j=1}^{N} p_j^*\phi_{\mu_j^*,\sigma_n}(x,\, y)$ and $\|(f_{\theta^*}g)^{1/2} - (f_0\mathbf{1}_{\mathcal{X}})^{1/2}\|_2 = O(\sigma_n^\beta)$. Third, $P_0(\|Z\| > a_{\sigma_n}) = O(\sigma_n^{4\beta+2\eta+8})$.

Let $\mathcal{M}(\mathbb{R}^d)$ denote the class of all probability measures on $\mathbb{R}^d$. Define $p_j^* = 0$ for $j = N + 1,\, \ldots,\, M$. Let $B_n = \mathcal{P}_n \times \mathcal{S}_n$ be the set with

$$\mathcal{P}_n = \left\{F \in \mathcal{M}(\mathbb{R}^d) : \sum_{j=1}^{M} |F(U_j) - p_j^*| \leq 2\epsilon_n^{2db_1}, \quad \min_{j=1,\,\ldots,\,M} F(U_j) \geq \epsilon_n^{4db_1}/2\right\}$$

and $\mathcal{S}_n = [\sigma_n(1 + \sigma_n^{2\beta})^{-1/2},\, \sigma_n]$. Note that $M\epsilon_n^{2db_1} \leq \epsilon_n^{2d(b_1-1/2\beta)}|\log\epsilon_n|^{ds} \leq 1$ and $\inf_{(l,\, t)\in\mathbb{R}\times\mathbb{R}_+} \min_{1\leq j\leq M}(c_0\nu \times G_t(\cdot - l))(U_j)^{1/2} \gtrsim \epsilon_n^{2db_1}(\epsilon_n^{b_1-1/2\beta}|\log\epsilon_n|)^{-d} \gtrsim$

$\epsilon_n^{2db_1}$. For every $\theta = (F,\, \sigma) \in B_n$, the q-integrated Hellinger distance $h(f_\theta,\, f_0) = O(\sigma_n^\beta)$. Proceeding as in Theorem 3.1 of Norets and Pati (2014), we obtain that $\max\{\mathrm{KL}(f_0;\, f_\theta),\, \mathrm{V}_2(f_0;\, f_\theta)\} = O(n\epsilon_n^2)$. We now evaluate the probability of the set $B_n = \mathcal{P}_n \times \mathcal{S}_n$. By applying Lemma 10 of Ghosal and van der Vaart (2007),

$$\inf_{(l,\, t)\in K_n} \mathrm{DP}_{c_0\nu\times G_t(\cdot - l)}(\mathcal{P}_n) \gtrsim \exp\left(-M|\log\epsilon_n|\right) \gtrsim \exp\left(-c_1\epsilon_n^{-d/\beta}|\log\epsilon_n|^{ds+1}\right).$$

Also, for the probability of the set $\mathcal{S}_n$ under the $\mathrm{IG}(\alpha,\, b)$, which is denoted by $P_b(\mathcal{S}_n)$, we have

$$\begin{aligned}\inf_{b\in K_n} P_b(\mathcal{S}_n) &= \inf_{b\in K_n}\int_{\sigma_n^{-1}}^{\sigma_n^{-1}(1+\sigma_n^{2\beta})^{1/2}} \frac{b^\alpha}{\Gamma(\alpha)} e^{-b\sigma}\sigma^{\alpha-1}\,\mathrm{d}\sigma \\ &\gtrsim \underline{b}_n^\alpha \exp\left(-\sqrt{2}\bar{b}_n/\sigma_n\right)\sigma_n^{-\alpha}[(1+\sigma_n^{2\beta})^{\alpha/2}-1] \gtrsim \exp\left(-c_2\bar{b}_n/\sigma_n\right)\end{aligned}$$

for a suitable constant $c_2 > 0$, provided that $\bar{b}_n = O(\log^a n)$, with $a > 0$, and $\underline{b}_n^{-1} = O(\sigma_n^{-1})$. Consequently,

$$\inf_{\gamma\in K_n} \mathrm{DP}_{c_0\nu\times G_t(\cdot - l)}(\mathcal{P}_n) \times P_b(\mathcal{S}_n) \gtrsim \exp\left(-c_3\epsilon_n^{-d/\beta}|\log\epsilon_n|^{(ds+1)\vee a}\right) \gtrsim \exp\left(-c_3 n\epsilon_n^2\right),$$

provided that, for $\epsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t$, the exponent $t \geq [(ds+1)\vee a]/(2+1/\beta)$. To complete verification of condition [**A1**], we show that, for some constant $c_4 > 0$,

$$\sup_{\gamma'\in K_n}\sup_{\theta\in B_n} P_0^n\Big(\inf_{\gamma:\, \|\gamma-\gamma'\|\leq u_n} \ell_n(\psi_{\gamma',\gamma}(\theta)) < -c_4 n\epsilon_n^2\Big) = o(N_n^{-1}).$$

Fix $\gamma' = (b_r,\, l_k,\, t_s^2) \in B_r \times L_k \times T_s$ and consider any $\gamma = (b,\, l,\, t^2) \in B_r \times L_k \times T_s$. For every $\theta \in B_n$,

$$\begin{aligned}\inf_{\gamma:\, \|\gamma-\gamma'\|\leq u_n} f_{\psi_{\gamma',\gamma}(\theta)}(y|x) &\geq \inf_{\gamma:\, \|\gamma-\gamma'\|\leq u_n}\sum_{j=1}^M \mathbf{1}_{\|\zeta_j'\|\leq a_{\sigma_n}} p_{j,\pi_r\sigma'}(x)\phi_{\pi_r\sigma'}(y-l-\rho_s\zeta_j') \\ &\geq T_n(y)(1+z_n)^{-2}e^{-12d_x z_n/\sigma_n^2} \\ &\qquad\qquad \times\sum_{j=1}^M \mathbf{1}_{\|\zeta_j'\|\leq a_{\sigma_n}} p_{j,\sigma'}(x)\phi_{\sigma'}(y-l_k-\zeta_j'),\end{aligned}$$

where

$$T_n(y) := \exp\Big(-\frac{1}{(\pi_r\sigma')^2}[w_n^2 a_{\sigma_n}^2 + d_y v_n^2 + (w_n a_{\sigma_n} + v_n){d_y}^{1/2}(a_{\sigma_n} + \|y-l_k\|)]\Big).$$

Over the set $\mathcal{Y}_0^n = \{(y_1,\, \ldots,\, y_n) \in (\mathbb{R}^{d_y})^n : \sum_{i=1}^n\sum_{j=1}^{d_y}(y_{ij} - \mathbb{E}_0[Y_j])^2 \leq d_y n\tau_n^2\}$, where $\tau_n = O(\log^\kappa n)$ for $\kappa > 0$,

$$T_n(y) \geq \exp\Big(-\frac{4}{\sigma_n^2}(1+d_y^{1/2})m_n[a_{\sigma_n} + 4\max\{d_y^{1/2}\bar{l}_n/2,\, \tau_n\}]\Big),$$

with $m_n := \max\{w_n a_{\sigma_n},\, d_y^{1/2}v_n\}$. Set $c_n(x;\, \sigma') := \sum_{j=1}^M \mathbf{1}_{\|\zeta_j'\|\leq a_{\sigma_n}} p_{j,\sigma'}(x)$, we have $c_n(x;\, \sigma') \geq e^{-8d_x^{1/2}\epsilon_n^2}\sum_{j=1}^M \mathbf{1}_{\|\zeta_j'\|\leq a_{\sigma_n}} p_j \geq e^{-8d_x^{1/2}\epsilon_n^2}(1-2\epsilon_n^{2db_1}) > e^{-8d_x^{1/2}\epsilon_n^2}\epsilon_n^2$.

Let F' be the distribution obtained by re-normalizing $\sum_{j=1}^{M} \mathbf{1}_{\|\zeta_j'\| \le a_{\sigma_n}} p_j \delta_{(\mu_j^x,\, l_k + \zeta_j')}$. For $\theta' = (F',\, \sigma')$, on the event $\mathcal{Y}_0^n$, for a suitable constant $C' > 0$,

$$\begin{aligned}
\inf_{\gamma:\, \|\gamma - \gamma'\| \le u_n} & \ell_n(f_{\psi_{\gamma',\gamma}(\theta)}) \\
&\ge \sum_{i=1}^{n} \log \frac{f_{\theta'}(y_i|x_i)}{f_0(y_i|x_i)} - 2n \log(1 + z_n) + \sum_{i=1}^{n} \log c_n(x_i;\, \sigma') \\
&\qquad - \frac{4n}{\sigma_n^2}[(1 + d_y^{1/2}) m_n (a_{\sigma_n} + 4 \max\{d_y^{1/2} \bar{l}_n/2,\, \tau_n\}) + 3 d_x z_n] \\
&\ge \sum_{i=1}^{n} \log \frac{f_{\theta'}(y_i|x_i)}{f_0(y_i|x_i)} - C' n \epsilon_n^2,
\end{aligned}$$

provided that $z_n = O(\sigma_n^2 \epsilon_n^2)$ and $m_n = O(\sigma_n^2 \epsilon_n^2 (\max\{a_{\sigma_n},\, \bar{l}_n,\, \tau_n\})^{-1})$. Also, we have $1 - P_0^n(\mathcal{Y}_0^n) = O((n\tau_n^4)^{-1})$ and need that $(n\tau_n^4)^{-1} = o(N_n^{-1})$.

We show that the requirements of condition [**A2**] are satisfied. We start by describing a set $\mathcal{F}_n$ of conditional densities such that, for some constant $\zeta > 0$,

$$\log N(\zeta \epsilon_n,\, \mathcal{F}_n,\, h) = O(n\epsilon_n^2). \tag{7}$$

We consider the same sieve $\{\mathcal{F}_n\}$ as in Theorem 4.1 of Norets and Pati (2014). For $H_n = \lfloor n\epsilon_n^2/(\log n) \rfloor$, $\underline{p}_n = e^{-nH_n}$, $\underline{\sigma}_n = \epsilon_n^{1/\beta}$, $\bar{\sigma}_n = e^{Tn\epsilon_n^2}$ for some constant $T > 0$, and $\bar{\mu}_n = (\log n)^{\tau_1}$ for some $\tau_1 > 0$, let

$$\begin{aligned}
\mathcal{F}_n := \Bigg\{ \Bigg(\sum_{j=1}^{\omega} p_{j,\sigma}(x) \phi_\sigma(\cdot - \mu_j^y) \Bigg)_{x \in \mathcal{X}} &: p_j \ge \underline{p}_n,\ \mu_j^y \in [-\bar{\mu}_n,\, \bar{\mu}_n]^{d_y},\ j = 1,\, \ldots,\, \omega, \\
&\qquad \omega \le H_n,\ \sigma \in [\underline{\sigma}_n,\, \bar{\sigma}_n] \Bigg\}.
\end{aligned}$$

For every fixed $\gamma' \in K_n$, let $\mathcal{F}_n(\gamma') := \bigcup_{\gamma:\, \|\gamma - \gamma'\| \le u_n} \psi_{\gamma',\gamma}^{-1}(\mathcal{F}_n)$, where $\psi_{\gamma',\gamma}^{-1}(\mathcal{F}_n)$ denotes the preimage of the set $\mathcal{F}_n$ under the transformation $\psi_{\gamma',\gamma}$. We show that condition (a) is satisfied. Fix any $\gamma' = (b_r,\, l_k,\, t_s^2) \in K_n$. Proceeding as in Theorem 4.1 of Norets and Pati (2014),

$$\begin{aligned}
\sup_{\gamma:\, \|\gamma - \gamma'\| \le u_n} \sup_{\theta \in \mathcal{F}_n(\gamma')} \sup_{x \in \mathcal{X}} & \|f_\theta(\cdot|x) - f_{\psi_{\gamma',\gamma}(\theta)}(\cdot|x)\|_1 \\
&\lesssim \frac{1}{\sigma'(1 \wedge \pi_r)} \sum_{j=1}^{d_y} [|l - l_k| + \sigma'|1 - \pi_r|] + \frac{1}{\underline{\sigma}_n^2}|1 - \pi_r| \\
&\lesssim \frac{v_n}{\underline{\sigma}_n} + \frac{(1 + z_n) z_n \bar{b}_n}{\underline{\sigma}_n^2 \underline{b}_n} \lesssim \epsilon_n
\end{aligned}$$

as long as $v_n = O(\underline{\sigma}_n \epsilon_n)$ and $z_n = O(\underline{\sigma}_n^2 \underline{b}_n \epsilon_n / \bar{b}_n)$.

Regarding condition (b_1), it follows from (6) that $\sup_{\gamma \in K_n} \Pi_\gamma(\mathcal{F}_n(\gamma))/\Pi_\gamma(B_n) \lesssim e^{Kn\epsilon_n^2/2}$ for a suitable constant $K > 0$ arising from condition (b_3).

To check condition (b_2), for every $\gamma' = (b_r,\, l_k,\, t_s^2) \in K_n$ and any $\theta \in \mathcal{F}_n(\gamma')$, we find an upper bound on $\sup_{\gamma:\, \|\gamma - \gamma'\| \le u_n} f_{\psi_{\gamma',\gamma}(\theta)}(\cdot|x)$ by a function (not necessarily a density) $\bar{f}(\cdot|x)$. For some constant $c_0 > 0$, let $a_n = c_0(\log n)^{1/\tau}$. For $\|y\| \leq a_n/2$, if $\|\zeta_j'\| > a_n$ and $d_y^{1/2}\bar{l}_n \leq a_n/4$ then $\|y - l_k - \zeta_j'\| > \|\zeta_j'\|/4$. Setting $r_n^2 := [1 - 16d_y^{1/2}(v_n \vee w_n)]^{-1}$, for every $\omega \leq H_n$,

$$\begin{aligned}
&f_{\psi_{\gamma',\gamma}(\theta)}(y|x)\mathbf{1}_{\|y\|\leq a_n/2}(y)\\
&\quad\leq \sum_{j=1}^{\omega} p_{j,\sigma}(x)\phi_\sigma(y - l_k - \zeta_j')\\
&\qquad\quad \times \exp\left(\frac{1}{\sigma^2}\max\{v_n,\, w_n\}(d_y^{1/2} + \|\zeta_j'\|)\|y - l_k - \zeta_j'\|\right)\mathbf{1}_{\|y\|\leq a_n/2}(y)\\
&\quad\leq \max\{e^{(3/2+d_y^{1/2})^2(v_n\vee w_n)(a_n\vee\bar{l}_n)a_n/\sigma^2},\, r_n\}\\
&\qquad\quad \times \sum_{j=1}^{\omega} p_{j,\sigma}(x)[\mathbf{1}_{\|\zeta_j'\|\leq a_n}\phi_{l_k+\zeta_j',\sigma}(y) + \mathbf{1}_{\|\zeta_j'\|>a_n}\phi_{l_k+\zeta_j',r_n\sigma}(y)]\mathbf{1}_{\|y\|\leq a_n/2}(y)\\
&\quad\leq \max\{e^{(3/2+d_y^{1/2})^2(v_n\vee w_n)(a_n\vee\bar{l}_n)a_n/(\pi_r\sigma')^2},\, r_n\}\\
&\qquad\quad \times e^{6d_xz_n/(\sigma')^2}\mathbf{1}_{\|y\|\leq a_n/2}(y)\\
&\qquad\quad \times \sum_{j=1}^{\omega} p_{j,\sigma'}(x)[\mathbf{1}_{\|\zeta_j'\|\leq a_n}\phi_{l_k+\zeta_j',\pi_r\sigma'}(y) + \mathbf{1}_{\|\zeta_j'\|>a_n}\phi_{l_k+\zeta_j',r_n\pi_r\sigma'}(y)]\\
&\quad=: \bar{f}(y|x),
\end{aligned}$$

where in the third inequality we have used the fact that $p_{j,\sigma}(x) \leq e^{6d_xz_n/(\sigma')^2}p_{j,\sigma'}(x)$. Note that $\pi_r\sigma' \in [\underline{\sigma}_n,\, \bar{\sigma}_n]$ and $l_k+\zeta_j' \in [-\bar{\mu}_n,\, \bar{\mu}_n]^{d_y}$ for $j = 1,\, \ldots,\, \omega$, with $\omega \leq H_n$. Set the positions

$$c' := \max\{e^{(3/2+d_y^{1/2})^2(v_n\vee w_n)(a_n\vee\bar{l}_n)a_n/(\pi_r\sigma')^2},\, r_n\} \times e^{6d_xz_n/(\sigma')^2}$$

and

$$\begin{aligned}
c(x) := \sum_{j=1}^{\omega} p_{j,\sigma'}(x)\bigg[&\mathbf{1}_{\|\zeta_j'\|\leq a_n}\int_{\|y\|\leq a_n/2}\phi_{l_k+\zeta_j',\pi_r\sigma'}(y)\,\mathrm{d}y\\
&+ \mathbf{1}_{\|\zeta_j'\|>a_n}\int_{\|y\|\leq a_n/2}\phi_{l_k+\zeta_j',r_n\pi_r\sigma'}(y)\,\mathrm{d}y\bigg],
\end{aligned}$$

and observed that $c(x) \leq 1$ for all $x \in \mathcal{X}$, under the constraints $z_n = O((n\epsilon_n)^{-2})$ and $v_n \vee w_n = O(((a_n \vee \bar{l}_n)a_nn\epsilon_n^2)^{-1})$, the normalizing constant of $\prod_{i=1}^n \bar{f}(y_i|x_i)$ can be thus bounded above

$$\begin{aligned}
\prod_{i=1}^{n}[c' \times c(x_i)] &< \Big(\max\{e^{(3/2+d_y^{1/2})^2(v_n\vee w_n)(a_n\vee\bar{l}_n)a_n/\underline{\sigma}_n^2},\, r_n\} \times e^{6d_xz_n(1+z_n)^2/\underline{\sigma}_n^2}\Big)^n\\
&\lesssim \exp\big(C_3n(v_n \vee w_n)(a_n \vee \bar{l}_n)a_n(n\epsilon_n^2)^2 + 48d_xnz_n(n\epsilon_n^2)^2\big) \lesssim e^{C_3'n\epsilon_n^2}
\end{aligned}$$

for suitable constants $C_3,\, C_3' > 0$. Let $\mathcal{Y}_1 = \{y \in \mathcal{Y} :\ \|y\| \leq a_n/2\}$. We are allowed to consider the restriction to $(\mathcal{X}\times\mathcal{Y}_1)^n$ since, by virtue of assumption (iv), $P_0^n((\mathcal{X}\times$

$\mathcal{Y}_1^c)^n) = (\int_0^1 \int_{\|y\|>a_n/2} f_0(y|x)q(x)\,\mathrm{d}x\mathrm{d}y)^n \lesssim e^{-B_0 n(a_n/2)^\tau} \lesssim e^{-B_0 n\epsilon_n^2}$. Recalling that, in the present setting, $\mathrm{d}Q_{\theta,\gamma'}/\mathrm{d}m = \sup_{\gamma:\,\|\gamma-\gamma'\|\leq u_n} f_{\psi_{\gamma',\gamma}(\theta)}(\cdot|x)q(x)$, in order to show that condition (b_2) is satisfied, we need to prove that

$$\sup_{\gamma'\in K_n}\int_{\mathcal{F}_n^c(\gamma')} Q^n_{\theta,\gamma'}(\mathcal{Z}^n)\frac{\Pi_{\gamma'}(\mathrm{d}\theta)}{\Pi_{\gamma'}(B_n)} = o(N_n^{-1}e^{-C_2 n\epsilon_n^2}).$$

By inequality (6), it suffices to show that

$$\sup_{\gamma'\in K_n}\int_{\mathcal{F}_n^c(\gamma')} Q^n_{\theta,\gamma'}(\mathcal{Z}^n)\Pi_{\gamma'}(\mathrm{d}\theta) = O(e^{-En\epsilon_n^2}) \tag{8}$$

for some constant $E > (C_2 \vee c_3)$, where c_3 plays the role of C in (6). The integral in (8) can be thus split up:

$$\begin{aligned}
&\sup_{\gamma'\in K_n}\int_{\mathcal{F}_n^c(\gamma')} Q^n_{\theta,\gamma'}(\mathcal{Z}^n)\Pi_{\gamma'}(\mathrm{d}\theta)\\
&\qquad= \sup_{\gamma'\in K_n}\Bigg[\int_{F\in\mathcal{M}(\mathbb{R}^d)}\Bigg(\int_{\sigma'<\underline{\sigma}_n} + \int_{\sigma'>\bar{\sigma}_n/2}\Bigg) Q^n_{\theta,\gamma'}((\mathcal{X}\times\mathcal{Y}_1)^n)\Pi_{\gamma'}(\mathrm{d}\theta)\\
&\qquad\qquad\qquad + \int_{F\in\mathcal{F}_n^c(\gamma')}\int_{\underline{\sigma}_n/2}^{\bar{\sigma}_n} Q^n_{\theta,\gamma'}((\mathcal{X}\times\mathcal{Y}_1)^n)\Pi_{\gamma'}(\mathrm{d}\theta)\Bigg]\\
&\qquad=: S_1 + S_2 + S_3.
\end{aligned}$$

To deal with the term S_1, we partition $(0,\,\underline{\sigma}_n) = \bigcup_{j=0}^\infty[\underline{\sigma}_n 2^{-(j+1)},\,\underline{\sigma}_n 2^{-j})$. For every $j\in\mathbb{N}_0$, let $u_{n,j} = e_n(\underline{\sigma}_n 2^{-j})$, with $e_n = o(1)$ so that $u_{n,j} < u_n$. For every $\gamma' = (b_r,\,l_k,\,t_s^2)\in K_n$, consider a $u_{n,j}$-covering of $\{\gamma : \|\gamma-\gamma'\|\leq u_n\}$ with centering points γ_i, for $i = 1,\,\ldots,\,N_j$, with $N_j \leq (u_n/u_{n,j})^3$. For a suitable constant $A > 0$,

$$\begin{aligned}
&\sup_{\gamma'\in K_n}\int_{F\in\mathcal{M}(\mathbb{R}^d)}\int_{\sigma'<\underline{\sigma}_n} Q^n_{\theta,\gamma'}((\mathcal{X}\times\mathcal{Y}_1)^n)\Pi_{\gamma'}(\mathrm{d}\theta)\\
&= O\Bigg(\sum_{j=0}^\infty \exp\Big(nu_{n,j}[(3/2+d_y^{1/2})^2(a_n\vee\bar{l}_n)a_n + 6d_x]/(\underline{\sigma}_n 2^{-(j+1)})^2 + nu_{n,j}\Big)\\
&\qquad\qquad\qquad\qquad \times \max_{1\leq i\leq N_j} P_{b_i}([\underline{\sigma}_n 2^{-(j+1)},\,\underline{\sigma}_n 2^{-j})\Bigg)\\
&= O\Bigg(\sum_{j=0}^\infty \exp\Big(2ne_n[(3/2+d_y^{1/2})^2(a_n\vee\bar{l}_n)a_n + 6d_x]/(\underline{\sigma}_n 2^{-(j+1)}) + ne_n\underline{\sigma}_n 2^{-j}\Big)\\
&\qquad\qquad\qquad\qquad \times \exp(-(\underline{b}_n/\underline{\sigma}_n)2^j)2^{(\alpha-1)j}\sum_{i=1}^{N_j}(b_i/\underline{\sigma}_n)^{\alpha-1}\Bigg)\\
&= O\Bigg(u_n(\underline{b}_n/\underline{\sigma}_n)^{\alpha-1}\exp(ne_n\underline{\sigma}_n + u_n - \log(e_n\underline{\sigma}_n))\\
&\qquad\qquad \sum_{j=0}^\infty e^{-(2^j\{[\underline{b}_n - 2ne_n[(3/2+d_y^{1/2})^2(a_n\vee\bar{l}_n)a_n+6d_x]/\underline{\sigma}_n - 1\} + j(1-\alpha)\log 2)}\Bigg)\\
&= O(e^{-An\epsilon_n^2})
\end{aligned}$$

provided that $e_n = o((n\underline{\sigma}_n)^{-1})$, $\underline{b}_n \gtrsim (\log n)^{-\upsilon}$ for some $\upsilon > 0$ and $e_n = O(n^{-1}(a_n \vee \bar{l}_n)^{-1}a_n^{-1})$.

Concerning S_2, for a suitable constant $B > 0$

$$\begin{aligned} S_2 \lesssim (\max\{e^{4(3/2+d_y^{1/2})^2(v_n \vee w_n)(a_n \vee \bar{l}_n)a_n e^{-2Tn\epsilon_n^2}},\, r_n\})^n \\ \times e^{24d_x n z_n e^{-2Tn\epsilon_n^2}}(1+z_n)^n \sup_{b \in K_n} P_b((\bar{\sigma}_n/2,\, \infty)) \lesssim e^{-Bn\epsilon_n^2} \end{aligned}$$

because

$$\begin{aligned} \sup_{b \in K_n} P_b((\bar{\sigma}_n/2,\, \infty)) &= \sup_{b \in K_n} \int_0^{4\bar{\sigma}_n^{-2}} \frac{b_r^\alpha}{\Gamma(\alpha)} e^{-b_r\sigma}\sigma^{\alpha-1}\,\mathrm{d}\sigma \\ &\leq (4\bar{b}_n\bar{\sigma}_n^{-2})^{\alpha-1}(1 - e^{-4\bar{b}_n\bar{\sigma}_n^{-2}}) \\ &= (4\bar{b}_n\bar{\sigma}_n^{-2})^{\alpha-1}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k!}(4\bar{b}_n\bar{\sigma}_n^{-2})^k \\ &\lesssim \bar{b}_n e^{-2Tn\epsilon_n^2}\exp(-2\alpha Tn\epsilon_n^2 + \alpha\log\bar{b}_n) \end{aligned}$$

provided that $z_n = O(\epsilon_n^2)$ and $(v_n \vee w_n) = O(n^{-1}(a_n \vee \bar{l}_n)^{-1}a_n^{-1}\epsilon_n^2)$.

Concerning S_3, for any $\epsilon \in (0,\, 1)$ and a suitable constant $D > 0$,

$$\begin{aligned} &\int_{F \in \mathcal{F}_n^c(\gamma')}\int_{\underline{\sigma}_n/2}^{\bar{\sigma}_n} Q_{\theta,\gamma'}^n((\mathcal{X} \times \mathcal{Y}_1)^n)\Pi_{\gamma'}(\mathrm{d}\theta) \\ &\quad\lesssim (\max\{e^{4(3/2+d_y^{1/2})^2(v_n \vee w_n)(a_n \vee \bar{l}_n)a_n/\underline{\sigma}_n^2},\, r_n\})^n e^{24d_x n z_n/\underline{\sigma}_n^2 + nz_n} \\ &\qquad\quad \times (1+z_n)^{-n}\underline{\sigma}_n^{-n}\exp(-ne^{-8d_y^{1/2}(v_n \vee w_n)}c\bar{\mu}_n^2/[2(1+z_n)^2\bar{\sigma}_n^2]) \lesssim e^{-Dn\epsilon_n^2}, \end{aligned}$$

provided that $z_n = O(n^{-1}\underline{\sigma}_n^2\epsilon_n^2)$ and $(v_n \vee w_n) = O(n^{-1}(a_n \vee \bar{l}_n)^{-1}a_n^{-1}\underline{\sigma}_n^2\epsilon_n^2)$, with $a_n < 2d_y^{1/2}\bar{\mu}_n$.

We now check that condition (b_3) is satisfied. We show that there exists a constant $K > 0$ such that, for any fixed $\gamma' = (b_r,\, l_k,\, t_s^2) \in K_n$, for every $\epsilon > 0$ and all $\theta \in \mathcal{F}_n(\gamma')$ such that the q-integrated Hellinger distance $h(f_\theta,\, f_0) > \epsilon$, there exists a test $\phi_n(f_\theta)$ satisfying

$$P_0^n\phi_n(f_\theta) \leq e^{-Kn\epsilon^2} \qquad \text{and} \qquad Q_{\theta,\gamma'}^n[1 - \phi_n(f_\theta)] \leq e^{-Kn\epsilon^2}. \tag{9}$$

By Corollary 1 of Ghosal and van der Vaart (2007), for every $\theta \in \mathcal{F}_n(\gamma')$ such that $h(f_\theta,\, f_0) > M\epsilon_n$, there exists a test ϕ_n, which is the maximum of all tests attached to probability measures that are the centers of balls covering $\{\theta \in \mathcal{F}_n(\gamma') : h(f_\theta,\, f_0) > M\epsilon_n\}$, such that

$$\begin{aligned} P_0^n\phi_n \lesssim N(M\epsilon_n/4,\, \mathcal{F}_n(\gamma'),\, h)e^{-n(M\epsilon_n/4)^2} \\ \text{and} \qquad \sup_{\theta \in \mathcal{F}_n(\gamma')} P_\theta^n(1 - \phi_n) \lesssim e^{-n(M\epsilon_n/4)^2}. \end{aligned}$$

By inequality (7), the requirement on the I type error probability in (9) is satisfied. The second requirement is satisfied provided that, for some constant

$M'' > 0$, we have $h(f_{\psi_{\gamma',\gamma}(\theta)},\, f_0) > M''\epsilon_n$ for all γ such that $\|\gamma - \gamma'\| \leq u_n$. Since $h(f_{\psi_{\gamma',\gamma}(\theta)},\, f_0) \geq 2^{-1}(\|f_\theta - f_0\|_1 - \|f_\theta - f_{\psi_{\gamma',\gamma}(\theta)}\|_1)$, it is enough that $\sup_{x\in\mathcal{X}} \|f_\theta(\cdot|x) - f_{\psi_{\gamma',\gamma}(\theta)}(\cdot|x)\|_1 \leq M'\epsilon_n$ for some constant $M' < M$ so that $M'' = M - M'$. This can be seen to hold as for condition (a). Inequality (8) then follows by combining upper bounds on S_1, S_2 and S_3.

The proof is completed noting that the assertion follows by choosing sequences v_n, w_n and z_n so that all the constraints arisen in the proof are simultaneously satisfied.

Remark 3. *Theorem 2 takes into account only a data-driven choice of the scale parameter of an inverse-gamma prior on the bandwidth, but an empirical Bayes selection of the shape parameter could be considered as well. In order to identify the mapping for the change of prior measure, it suffices to note that, for* $\alpha \in \mathbb{N}$, *if* $\alpha_r \overset{\text{iid}}{\sim} \mathrm{Gamma}(1,\, 1)$, $r = 1,\, \ldots,\, \alpha$, *then* $\beta/(\sigma_1 + \ldots + \sigma_\alpha) \sim \mathrm{IG}(\alpha,\, \beta)$.

2.2. Empirical Bayes dimension reduction in the presence of irrelevant covariates

We now deal with the case where a d_x-dimensional explanatory variable is considered, but not all the covariates are relevant to the response whose conditional distribution may depend only on fewer of them, say $0 \leq d_x^0 \leq d_x$, which, without loss of generality, can be thought of as the first d_x^0 of the whole collection employed in the model specified in (3). Besides rate adaptation, another appealing feature of the empirical Bayes procedure herein considered is automatic dimension reduction in the presence of irrelevant covariates, on par with the posterior distribution corresponding to the prior proposed by Norets and Pati (2014). The posterior automatically selects the model with the subset of relevant covariates among all competing models.

Theorem 4. *Suppose that the true conditional density* f_0 *depends on the first* $d_x^0 \in \mathbb{N}_0$ *covariates and satisfies assumptions* (iii)-(iv) *of Section 2.1.1. Under the same conditions as in Theorem 2, the empirical Bayes posterior distribution corresponding to the prior in* (3) *contracts at a rate* $\epsilon_n = n^{-\beta/(2\beta+d^0)}(\log n)^t$, *with* $d^0 := d_x^0 + d_y$ *and* $t > 0$ *a suitable constant.*

The proof follows the same trail as that of Theorem 2, the only difference arising from the prior concentration rate which turns out to depend on the dimension d_x^0 of the relevant covariates of f_0 because, for all the locations of the approximating Gaussian mixture, when $k > d_x^0$, the components $\mu_{jk}^x = 0$ so that eventually the mixture does not depend on the covariates x_k for $k = d_x^0 + 1,\, \ldots,\, d_x$.

As a simple consequence of Theorem 4, we have that, if $d_x^0 = 0$, then $f_0(y|x) = f_0(y)$ and the response is stochastically independent of the predictor.

3. Final Remarks

In this note, we have proposed an empirical Bayes procedure for conditional density estimation based on infinite mixtures of Gaussian kernels with predictor-dependent mixing weights. We have shown that a data-driven selection of the prior hyper-parameters can lead to inferential answers that are comparable, for large sample sizes, to those of hierarchical posteriors in automatically adapting to the dimension of the set of relevant covariates and to the regularity level of the true sampling conditional density. An empirical Bayes selection of the prior hyper-parameters may lead to pseudo-posterior distributions with the same performance as fully Bayes posteriors, provided the estimator $\hat{\beta}_n$ of the scale parameter of an inverse-gamma prior on the bandwidth takes values in a set $[\underline{b}_n,\, \bar{b}_n)$ such that $P_0^n(\hat{\beta}_n \in [\underline{b}_n,\, \bar{b}_n)^c) = o(1)$. The last requirement imposes restrictions on the sequences $\underline{b}_n$ and $\bar{b}_n$, in particular, on the decay rate at zero of $\underline{b}_n$, which is expectedly more important than the rate at which $\bar{b}_n \uparrow \infty$. If the prior hyper-parameter has an impact on posterior contraction rates, then the choice of the plug-in estimator is crucial and requires special care. This may, for example, rule out the maximum marginal likelihood estimator for β. When the hyper-parameter does not affect posterior contraction rates, as it is the case for the mean λ and variance τ^2 of the Dirichlet base measure, there is flexibility in the choice of the estimator: different choices are indistinguishable in terms of the posterior behavior they induce and empirical Bayes posterior contraction rates are the same as those of any posterior corresponding to a prior with fixed hyper-parameters.

The result of Theorem 4 deals with isotropic Hölder densities, but an extension to anisotropic densities is envisaged. In the anisotropic case, the presented results provide adaptive rates corresponding to the least smooth direction. Sharper rates can be obtained using component-specific bandwidths along the lines of Section 5 in Shen *et al.* (2013) combined with the preceding treatment. Details are omitted.

Acknowledgements

The author would like to thank the Editors and two anonymous referees for their remarks and comments which helped to improve the initial version of the manuscript. Bocconi University is gratefully acknowledged for providing financial support.

Appendix

In this section, an adapted version of Theorem 1 in Donnet *et al.* (2014) is reported for easy reference. Some additional notation is preliminarily introduced.

Let $(\mathcal{X}^{(n)},\, \mathcal{B}_n,\, (P_\theta^{(n)} : \theta \in \Theta))$ be a sequence of statistical experiments, where $\mathcal{X}^{(n)}$ and Θ are Polish spaces endowed with their Borel σ-fields $\mathcal{B}_n$ and $\mathcal{B}(\Theta)$, respectively. Let $d(\cdot,\, \cdot)$ denote a (semi-)metric on Θ. Let $X^{(n)} \in \mathcal{X}^{(n)}$ be the observation at the nth stage from $P_{\theta_0}^{(n)}$, where θ_0 denotes the true parameter. Let $\mu^{(n)}$ be a σ-finite measure on $(\mathcal{X}^{(n)},\, \mathcal{B}_n)$ dominating all probability measures $P_\theta^{(n)}$, for $\theta \in \Theta$. For every $\theta \in \Theta$, let $\ell_n(\theta)$ denote the log-likelihood ratio $\log(p_\theta^{(n)}/p_{\theta_0}^{(n)})$.

We consider a family of prior distributions $(\Pi_\gamma,\, \gamma \in \Gamma)$ on $(\Theta,\, \mathcal{B}(\Theta))$, with $\Gamma \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$. Let $\Pi_\gamma(\cdot|X^{(n)})$ stand for the posterior distribution corresponding to Π_γ. For any measurable function $\hat{\gamma}_n : \mathcal{X}^{(n)} \to \Gamma$, the empirical Bayes posterior law $\Pi_{\hat{\gamma}_n}(\cdot|X^{(n)})$ is obtained by plugging $\hat{\gamma}_n$ into the posterior distribution,

$$\Pi_{\hat{\gamma}_n}(\cdot|X^{(n)}) = \Pi_\gamma(\cdot|X^{(n)})|_{\gamma=\hat{\gamma}_n}.$$

The statement of the theorem follows.

Theorem 5 (Donnet *et al.* (2014)). *Let $\theta_0 \in \Theta$. For every $\gamma,\, \gamma' \in \Gamma$, let $\psi_{\gamma,\gamma'} : \Theta \to \Theta$ be a measurable mapping such that, if $\theta \sim \Pi_\gamma$, then $\psi_{\gamma,\gamma'}(\theta) \sim \Pi_{\gamma'}$. Assume that*

[A1] *there exist sets $K_n \subseteq \Gamma$ with $P^{(n)}_{\theta_0}(\hat{\gamma}_n \in K_n^c) = o(1)$, positive sequences $u_n,\, \epsilon_n \downarrow 0$, with $n\epsilon_n^2 \to \infty$, for which $N_n := N(u_n,\, K_n,\, \|\cdot\|) = o(e^{n\epsilon_n^2})$ and sets $B_n \in \mathcal{B}(\Theta)$ such that, for some constant $C_1 > 0$,*

$$\sup_{\gamma\in K_n}\sup_{\theta\in B_n} P^{(n)}_{\theta_0}\Big(\inf_{\gamma':\ \|\gamma'-\gamma\|\le u_n} \ell_n(\psi_{\gamma,\gamma'}(\theta)) < -C_1 n\epsilon_n^2\Big) = o(N_n^{-1});$$

[A2] *for every $\gamma \in K_n$, there exists a set $\Theta_n(\gamma) \in \mathcal{B}(\Theta)$ such that*

(a) $\sup_{\gamma':\ \|\gamma'-\gamma\|\le u_n} \sup_{\theta\in\Theta_n(\gamma)} d(\theta,\, \psi_{\gamma,\gamma'}(\theta)) \le M'\epsilon_n$ *for some constant $M' > 0$,*

(b) *for constants $\zeta,\, K > 0$ and $C_2 > C_1$,*

(b_1) $\log N(\zeta\epsilon_n,\, \Theta_n(\gamma),\, d) \le Kn\epsilon_n^2/2$ *and* $\sup_{\gamma\in K_n} \dfrac{\Pi_\gamma(\Theta_n(\gamma))}{\Pi_\gamma(B_n)} \le e^{Kn\epsilon_n^2/2}$,

(b_2) *defined $Q^{(n)}_{\theta,\gamma}$ such that* $\mathrm{d}Q^{(n)}_{\theta,\gamma}/\mathrm{d}\mu^{(n)} := \sup_{\gamma':\ \|\gamma'-\gamma\|\le u_n} p^{(n)}_{\psi_{\gamma,\gamma'}(\theta)}$,

$$\sup_{\gamma\in K_n} \int_{\Theta\setminus\Theta_n(\gamma)} Q^{(n)}_{\theta,\gamma}(\mathcal{X}^{(n)}) \frac{\Pi_\gamma(\mathrm{d}\theta)}{\Pi_\gamma(B_n)} = o(N_n^{-1}e^{-C_2 n\epsilon_n^2}),$$

(b_3) *for any $\epsilon > 0$, $\theta \in \Theta_n(\gamma)$ with $d(\theta,\, \theta_0) > \epsilon$, there exists a test $\phi_n(\theta)$ with*

$$P^{(n)}_{\theta_0}\phi_n(\theta) \le e^{-Kn\epsilon^2} \qquad \textit{and} \qquad Q^{(n)}_{\theta,\gamma}[1-\phi_n(\theta)] \le e^{-Kn\epsilon^2}.$$

Then, for a sufficiently large constant $M > 0$,

$$P^{(n)}_{\theta_0}\Pi_{\hat{\gamma}_n}\big(d(\theta,\, \theta_0) > M\epsilon_n|X^{(n)}\big) \to 0.$$

References

K. Bertin, C. Lacour, V. Rivoirard (2015). *Adaptive pointwise estimation of conditional density function.* Annales de l'Institut Henri Poincaré (to appear).

S. Donnet, V. Rivoirard, J. Rousseau, C. Scricciolo (2014). *Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures.* URL `http://arxiv.org/pdf/1406.4406.pdf`.

S. Efromovich (2007). *Conditional density estimation in a regression setting.* The Annals of Statistics, 35, no. 6, pp. 2504–2535.

S. Efromovich (2010). *Oracle inequality for conditional density estimation and an actuarial example.* Annals of the Institute of Statistical Mathematics, 62, no. 2, pp. 249–275.

S. Ghosal, A. van der Vaart (2007). *Posterior convergence rates of Dirichlet mixtures at smooth densities.* The Annals of Statistics, 35, no. 2, pp. 697–723.

P. Hall, J. Racine, Q. Li (2004). *Cross-validation and the estimation of conditional probability densities.* Journal of the American Statistical Association, 99, no. 468, pp. 1015–1026.

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

A. Norets, D. Pati (2014). *Adaptive Bayesian estimation of conditional densities.* URL `http://arxiv.org/pdf/1408.5355.pdf`.

D. Pati, D. B. Dunson, S. T. Tokdar (2013). *Posterior consistency in conditional distribution estimation.* Journal of Multivariate Analysis, 116, pp. 456–472.

S. Petrone, S. Rizzelli, J. Rousseau, C. Scricciolo (2014a). *Empirical Bayes methods in classical and Bayesian inference.* METRON, 72, no. 2, pp. 201–215.

S. Petrone, J. Rousseau, C. Scricciolo (2014b). *Bayes and empirical Bayes: do they merge?* Biometrika, 101, no. 2, pp. 285–302.

A. Rodríguez, D. B. Dunson (2011). *Nonparametric Bayesian models through probit stick-breaking processes.* Bayesian Analysis, 6, no. 1, pp. 145–177.

C. Scricciolo (2015). *Bayesian adaptation.* Journal of Statistical Planning and Inference (in press), 166, 87-101.

W. Shen, S. T. Tokdar, S. Ghosal (2013). *Adaptive Bayesian multivariate density estimation with Dirichlet mixtures.* Biometrika, 100, no. 3, pp. 623–640.

Summary

The problem of nonparametric estimation of the conditional density of a response, given a vector of explanatory variables, is classical and of prominent importance in many prediction problems since the conditional density provides a more comprehensive description of the association between the response and the predictor than, for instance, does the regression function. The problem has applications across different fields like economy, actuarial sciences and medicine. We investigate empirical Bayes estimation of conditional densities establishing that an automatic data-driven selection of the prior hyper-parameters in infinite mixtures of Gaussian kernels, with predictor-dependent mixing weights, can lead to estimators whose performance is on par with that of frequentist estimators in being minimax-optimal (up to logarithmic factors) rate adaptive over classes of locally Hölder smooth conditional densities and in performing an adaptive dimension reduction if the response is independent of (some of) the explanatory variables which, containing no information about the response, are irrelevant to the purpose of estimating its conditional density.

Keywords: Adaptive estimation; Bayesian nonparametrics; Conditional density; Dimension reduction; Hölder spaces; minimax rates of convergence