Empirical Bayes conditional density estimation

Catia Scricciolo

1 Department of Decision Sciences, Bocconi University, Milan, Italy, catia.scricciolo@unibocconi.it

Abstract

The problem of nonparametric estimation of the conditional density of a response given explanatory variables is classical and prominent in many practical prediction problems, where the conditional density provides a more detailed description of the association between the predictor(s) and the response than does the regression function. The problem has applications across different fields like economy, actuarial sciences, medicine. We investigate empirical Bayes adaptive estimation of conditional densities establishing that an empirical Bayes selection of the prior hyper-parameter values can still lead to estimators that are comparable to the frequentist ones in being optimal from the minimax point of view over classes of locally Hölder densities and in performing a “dimension reduction”, thus ultimately retrieving the correct dimension of the relevant covariates, if some of the explanatory variables introduced in the model contain no information about the response and are therefore irrelevant to the purpose of estimating its conditional density.

1 Introduction

The problem of estimating the conditional density of a response, given predictors, is classical and of the most 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 example, does the conditional expectation that can only capture partial aspects. The conditional density contains information on how the different features of the response distribution, like skewness, shape and so on, change with covariates. Conditional density estimation for predictive purposes have applications across different fields like economy, actuarial sciences, medicine.

Nonparametric estimation of a collection of conditional densities over the covariate space is complicated by the curse of dimensionality. Classical references on
nonparametric conditional density estimation adopting a frequentist approach to inference are Efromovich (2007, 2010) and Hall et al. (2004). See also the recent contribution by Bertin et al. (2013). Popular Bayesian methods for conditional density estimation 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 article presents results on posterior contraction rates for conditional density estimation over classes of locally Hölder densities by finite mixtures of Gaussian kernels with covariate-dependent mixing weights having a special structure. The derived Bayesian estimation procedure converges at a rate that automatically adapts to the unknown smoothness level of the sampling conditional density and retrieves the correct dimension of the relevant covariates when some of the explanatory variables considered in the model are irrelevant, thus ultimately performing a “dimension reduction”.

The aim of this note is to investigate whether an empirical Bayes conditional density estimation procedure corresponding to an infinite mixture of Gaussian kernels, with the same predictor-dependent mixing weights as in Norets and Pati (2014), may have a performance at par with that of the procedure in Norets and Pati (2014) in terms of rate adaptation and dimensionality reduction. Interestingly, under the same set of assumptions about the data generating process, the performance of the conditional density estimation procedure of an empirical Bayesian who considers a data-driven choice of the prior hyper-parameters matches that of an “honest” Bayesian.

The organization of the paper is as follows. Section 1.1 introduces the notation used throughout the paper. Section 2 presents the main results about empirical Bayes posterior concentration at optimal minimax $L^1$-rate for locally Hölder conditional densities, with dimension reduction in the presence of irrelevant covariates in the model.

1.1 Notation

- $\mathbb{N}_0 = \{0, 1, \ldots\}$.
- $\mathbb{R}_+$: the set of (strictly) positive real numbers.
- $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ for $d_x \in \mathbb{N}$: covariate space.
- $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ for $d_y \in \mathbb{N}$: response space.
• \( Z = \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \): sample space.

• \( \mathcal{F} = \{ f : Z \to [0, \infty) \mid \text{Borel-measurable and } \forall x \in \mathcal{X}, \int_{\mathcal{Y}} f(y|x)\,dy = 1 \} \): space of conditional probability densities with respect to Lebesgue measure.

• \( \phi_\sigma \) denotes a centered multivariate normal density with covariance matrix \( \sigma^2 I \), for \( I \) the identity matrix whose dimension will be clear from the context.

• \( \delta_z \) stands for point mass at \( z \).

• Let \( Q \) be a fixed probability measure on \( (\mathcal{X}, \mathcal{B}(\mathcal{X})) \), with \( \mathcal{B}(\mathcal{X}) \) the Borel \( \sigma \)-field on \( \mathcal{X} \), that possesses Lebesgue density \( q \). Denote by \( \mathbb{E}_Q \) expectation with respect to \( Q \).

• Let \( p \geq 1 \) and \( g : Z \to \mathbb{R} \). For any \( x \in \mathcal{X} \), introduce the notation \( \|g\|_{p,x} := (\int_{\mathcal{Y}} |g(x, y)|^p \, dy)^{1/p} \). Let the \( q \)-integrated \( L^1 \)-distance between any \( f_1, f_2 \in \mathcal{F} \) be defined as \( \|f_2 - f_1\|_1 := \mathbb{E}_Q[\|f_2 - f_1\|_{1,x}] = \int_{\mathcal{X}} \|f_2 - f_1\|_{1,x} q(x) \, dx \). Analogously, let the \( q \)-integrated Hellinger distance between any \( f_1, f_2 \in \mathcal{F} \) be defined as \( h(f_2, f_1) := (\mathbb{E}_Q[\|f_2^{1/2} - f_1^{1/2}\|_{2,x}^2])^{1/2} = \int_{\mathcal{X}} \|f_2^{1/2} - f_1^{1/2}\|_{2,x}^2 q(x) \, dx \). Also, for any densities \( f, f_0 \in \mathcal{F} \), let the \( q \)-integrated Kullback-Leibler divergence of \( f \) from \( f_0 \) be defined as \( \text{KL}(f_0; f) := \int_{\mathcal{X}} q \int_{\mathcal{Y}} f_0 \log(f_0/f) \) which coincides with the Kullback-Leibler divergence of \( f q \) from \( f_0 q \).

• For real numbers \( a \) and \( b \), 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 which is universal or inessential for our purposes.

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

2 Main Results

Let \( Z^{(n)} = (Z_1, \ldots, Z_n) \) denote the observation at the \( n \)th stage consisting of independent and identically distributed (i.i.d.) replicates \( Z_i = (X_i, Y_i) \in Z \), \( i = 1, \ldots, n \), from a probability measure \( P_0 \) on \( (Z, \mathcal{B}(Z)) \) that possesses Lebesgue density \( f_0 q \), \( f_0 \in \mathcal{F} \), which is referred to as the true joint data generating density. The marginal density \( q \) is fixed and, for theoretical investigation, does not need to be known or estimated. The problem is to estimate the conditional probability density \( f_0 \) of the response \( Y \), given the predictor \( X \), when no parametric assumption is made on its form, taking an empirical Bayes approach that employs a data-driven choice of the
prior hyper-parameter values. For a recent account on empirical Bayes methods, the reader may refer to Petrone et al. (2014a). Empirical Bayes posterior concentration at $L^1$-minimax adaptive rates, up to a logarithmic factor, over classes of locally Hölder smooth conditional densities, is investigated in Section 2.1. Furthermore, in Section 2.2 it is shown that the empirical Bayes posterior asymptotically performs a dimension reduction when the true conditional density $f_0$ does not depend on all of the covariates introduced in the model, thus contracting at a rate which is only affected by the actual dimension of the explanatory variables.

### 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 specification

A prior can be induced on the space $\mathcal{F}$ of Lebesgue conditional densities by a law $\Pi_X$ on a collection of mixing probability measures $\mathcal{M}_X = \{P_x \in \mathcal{M}(\Theta), x \in \mathcal{X}\}$, where $\mathcal{M}(\Theta)$ denotes the space of all probability measures on $\Theta \subseteq \mathcal{Y}$, using a mixture of Gaussian kernels for the conditional density

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

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

$$P_x = \sum_{j=1}^\infty p_j(x)\delta_{\theta_j(x)}, \quad (2.1)$$

with random weights $p_j(x) \geq 0$, $j = 1, 2, \ldots$, such that $\sum_{j=1}^\infty p_j(x) = 1$ almost surely, and random support points $\{\theta_j(x)\}_{j=1}^\infty$ that are independent and identically distributed according to some probability measure $G$ on $\Theta$. Following Pati et al. (2013), we single out two relevant special cases.

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

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

where $\beta'x$ denotes the usual inner product on $\mathbb{R}^{d_x}$ and the mixing measure $P_x$ corresponding to $F_x$ is such that $P_x = \sum_{j=1}^{\infty} p_j(x)\delta_{\theta_j}$ almost surely, with the
\( \beta_j \overset{\text{iid}}{\sim} G \). 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 in constructing the stick-breaking weights has been considered in [Rodriguez and Dunson (2011)] who exhibit applications of the probit stick-breaking process model to real data.

b) **Gaussian mixtures of fixed-\( p \) dependent processes**: if \( p_j(x) \equiv p_j \) for all \( x \in \mathcal{X} \), with \( p_j \geq 0 \), \( j \in \mathbb{N} \), and \( \sum_{j=1}^{\infty} p_j = 1 \) almost surely, 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 adopt a variant of the prior proposed in [Norets and Pati (2014)]. Let \( \nu \) be a probability measure on \( \mathcal{X} \) and \( G \) a probability measure on \( \mathcal{Y} \) with Lebesgue density \( g \). For \( \psi = (\lambda, \tau) \in \mathcal{Y} \times \mathbb{R}_+ \), let \( G_\psi \) denote the probability measure corresponding to the density \( \tau^{-1} g((\cdot - \lambda)/\tau) \). We suggest the following model-based prior specification:

\[
Y_i | (X_i = x_i), (F_x)_{x \in \mathcal{X}}, \sigma \sim (F_{x_i} \ast \phi_\sigma)(y) = \sum_{j=1}^{\infty} p_j(x_i) \phi_\sigma(y - \mu_j^y),
\]

with \( p_j(x_i) := \frac{p_j \phi_\sigma(x_i - \mu_j^x)}{\sum_{m=1}^{\infty} p_m \phi_\sigma(x_i - \mu_m^x)}, \ j \in \mathbb{N} \),

\[
\sum_{j=1}^{\infty} p_j \delta(\mu_j^x, \mu_j^y) \sim \text{DP}(c_0 \nu \otimes G_\psi) \quad \text{independent of} \quad \sigma \sim \text{IG}(\alpha, \beta),
\]

where \( c_0 \) is a finite positive constant and \( \alpha, \beta > 0 \) are the shape and scale parameters, respectively, of an inverse-gamma distribution. Equivalently, by the stick-breaking representation of a Dirichlet process, the weights \( p_j = V_j \prod_{k=1}^{j-1} (1 - V_k), \ j \in \mathbb{N} \), with \( V_j \overset{\text{iid}}{\sim} \text{Beta}(1, c_0) \), and the locations \( \mu_j^y \overset{\text{iid}}{\sim} G_\psi \). The last assertion is equivalent to \( \mu_j^y = \lambda + \zeta_j \), with \( \zeta_j \overset{\text{iid}}{\sim} g_\tau, \ j \in \mathbb{N} \), where \( g_\tau(\cdot) = \tau^{-1} g(\cdot/\tau) \). The overall prior can be rewritten as

\[
Y_i | (X_i = x_i), (F_x)_{x \in \mathcal{X}}, \sigma \sim \sum_{j=1}^{\infty} p_j(x_i) \phi_\sigma(y - \lambda - \zeta_j)
\]

\[
\sum_{j=1}^{\infty} p_j \delta(\mu_j^x, \zeta_j) \sim \text{DP}(c_0 \nu \otimes G_\tau) \quad \text{indep. of} \quad \sigma \sim \text{IG}(\alpha, \beta),
\]

where \( G_\tau \) is the probability measure with density \( g_\tau \). For \( \gamma = (\lambda, \tau, \beta) \), let \( \Pi_\gamma \) denote the product prior \( \text{DP}(c_0 \nu \otimes G_\psi) \otimes \text{IG}(\alpha, \beta) \). The posterior probability of any Borel set \( B \) on \( \mathcal{F} \)

\[
\Pi_\gamma(B|Z^{(n)}) \propto \int_B \prod_{i=1}^{n} (F_{X_i} \ast \phi_\sigma)(Y_i)q(X_i)\Pi_\gamma(d(F_{X_i}, \sigma)).
\]
For an estimate $\hat{\gamma} = (\hat{\lambda}, \hat{\tau}, \hat{\beta})$ of $\gamma$ based on $Z^{(n)}$, the empirical Bayes posterior $\Pi_{\hat{\gamma}}(\cdot|Z^{(n)})$ is obtained by plugging $\hat{\gamma}$ into the fully Bayes posterior

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

We study the empirical Bayes posterior concentration relative to $d$, the $q$-integrated Hellinger or $L^1$-distance, at an ordinary smooth conditional density $f_0$, namely, we assess the order of magnitude of the radius $M_{\epsilon_n}$ of a ball centered at $f_0$ so that

$$
\mathbb{E}_0^n \Pi_{\gamma}(f \in F : d(f, f_0) > M_{\epsilon_n}|Z^{(n)}) \to 0,
$$

where $\mathbb{E}_0^n$ denotes expectation under the $n$-fold product measure $P_0^n$. We deal with the case where $f_0$ is locally Hölder smooth in the sense of the following definition for which we introduce some more notation. For any $\beta_0 > 0$, let $\beta_0 := \max\{i \in \mathbb{N}_0 : i < \beta_0\}$ be the largest integer strictly smaller than $\beta_0$. For a 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 $\partial^{k_1}_{z_1} \ldots \partial^{k_d}_{z_d}$.

**Definition 2.1.** For any $\beta_0 > 0$, $\tau_0 \geq 0$ and function $L : Z \to [0, \infty)$, the class of locally $\beta_0$-Hölder functions with envelope $L$, denoted by $C^{\beta_0, L, \tau_0}(Z)$, is the set of all functions $f : Z \to \mathbb{R}$ that have finite mixed partial derivatives $D^k f$ of all orders $k. \leq \beta_0$ and are such that, for every $k \in \mathbb{N}_0^d$, with $k. = \beta_0$,

$$
|D^k f(z + \Delta) - D^k f(z)| \leq L(z) \epsilon^{\tau_0} \|\Delta\|^2 \|\Delta\|^{\beta_0 - 2}, \quad z, \Delta \in Z.
$$

This class has been considered by Shen et al. (2013) who constructively show that functions in $C^{\beta_0, L, \tau_0}(Z)$ can be approximated by Gaussian mixtures $\phi_\sigma$ with an $L^1$-error of the order $\sigma^{\beta_0}$. A precise statement of the result on empirical Bayes posterior contraction rates for locally Hölder densities requires some assumptions listed below on the joint data generating density $f_0 q$ and on the prior law $\Pi_{\gamma}$.

**Assumptions on $f_0 q$:**

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

(ii) $q$ is bounded above;

(iii) $f_0 \in C^{\beta_0, L, \tau_0}(Z)$ and, for some $\eta > 0$,

$$
\int_Z (|L|/f_0)^{2+\eta/\beta_0} f_0 < \infty \quad \text{and} \quad \int_Z (|D^k f_0|/f_0)^{(2\beta_0+\eta)/k} f_0 < \infty \quad \text{for all} \quad k. \leq \beta_0;
$$
(iv) there exists a constant \( c > 0 \) such that, for every \( x \in \mathcal{X} \), \( f_0(y|x) \lesssim \exp \left( -c \|y\|^{\tau} \right) \) for all \( \|y\| \) large enough.

**Assumption on \( \Pi_\gamma \):**

(v) the base probability measure \( \nu \otimes G \) possesses Lebesgue positive density and there exist constants \( a, C > 0 \) such that \( 1 - G([-y, y]|_{\mathbb{R}^d}) \lesssim \exp \left( -Cy^a \right) \) for all sufficiently large \( y \in \mathbb{R}_+ \).

We now state the main result.

**Theorem 2.1.** Suppose there exists a set \( K_n \subset \mathcal{Y} \times \mathbb{R}^2_+ \) such that \( P_n(\hat{\gamma} \in K_n^c) = o(1) \). Under assumptions (i)-(v), the empirical Bayes posterior distribution corresponding to the prior in (2.2) contracts at rate \( \epsilon_n = n^{-\beta_0/(2\beta_0+d)}(\log n)^t \), with \( d := d_x + d_y \) and \( t > 0 \) a suitable constant.

A few comments may be of interest. The empirical Bayes posterior corresponding to the prior in (2.2) contracts at a rate \( n^{-\beta_0/(2\beta_0+d)}(\log n)^t \) which differs for at most a logarithmic factor from the minimax \( L^1 \)-rate associated with the class of locally \( \beta_0 \)-Hölder densities. Moreover, the rate automatically adapts to the unknown regularity level \( \beta_0 \) of \( f_0 \), despite lack of knowledge of it to be exploited in the definition of the prior (see, e.g. Scricciolo (2014) for a recent 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 a logarithmic term, and perform as well as fully Bayes adaptive procedures like the one derived from the hierarchical prior of finite Dirichlet mixtures of Gaussian densities proposed in Norets and Pati (2014).

The problem at study presents two main difficulties: (a) data-dependence of the prior due to an empirical Bayes selection of the prior hyper-parameter values, (b) dependence of \( f_0 \) on the covariates, which gives account for the dependence of the rate on the dimension \( d = d_x + d_y \) of the sample space \( \mathcal{Z} \). Data-dependence of the prior can be dealt with using the same idea as in Petrone et al. (2014b) and Donnet et al. (2014), which relies on a change of the prior measure aimed at transferring data-dependence from the prior to the likelihood, when a mapping can be identified. Dependence of the conditional density on the covariates can be dealt with regarding \( f_0 \) as a \((d_x + d_y)\)-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 \( Q \otimes \lambda \) on \( \mathcal{Z} \), which prevents immediate use of (Lebesgue) Gaussian mixture densities for approximating it. A tricky device due to Norets and Pati (2014) allows to exploit the approximation of the joint
Lebesgue density $f_01_{[0,1]}$ by a finite mixture of $(d_x + d_y)$-multivariate Gaussian densities taking advantage of the special structure of the mixing weights $p_j(x)$ in the model: the mixture $\sum_{j=1}^{\infty} p_j(x)\phi_\sigma(x - \mu_j^x)\phi_\sigma(y - \mu_j^y)$ can in fact be bounded above by a multiple $\sigma^{d_x} \sum_{j=1}^{\infty} p_j(x)\phi_\sigma(y - \mu_j^y)$ of the model density in (2.2).

**Proof.** The proof appeals to Theorem 1 of Donnet et al. (2014) and uses intermediate results by Norets and Pati (2014).

**Remark 2.1.** Theorem 2.1 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 as well considered. In order to identify the mapping for the change of prior measure, it suffices to note that, for $\alpha \in \mathbb{N}$, if $\alpha_r \overset{iid}{\sim} \text{Gamma}(1, 1)$, $r = 1, \ldots, \alpha$, then $\beta/(\sigma_1 + \ldots + \sigma_\alpha) \sim \text{IG}(\alpha, \beta)$.

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

In this section, we consider the case where a $d_x$-dimensional explanatory variable is considered in the model, but not all of the covariates are indeed relevant to the response whose conditional distribution may depend only on some of them, say $d_0^x \leq d_x$, which, without loss of generality, can be thought of as the first $d_0^x$ of the whole collection employed in the model specified in (2.2). Besides rate adaptation, another appealing feature of the empirical Bayes procedure herein considered is automatic dimensionality reduction in the presence of irrelevant covariates, at par with the fully Bayes posterior distribution corresponding to the prior proposed by Norets and Pati (2014). The posterior automatically selects the model having the correct subset of covariates among all possible competing models.

**Theorem 2.2.** Suppose that the true conditional density $f_0$ depends on the first $d_0^x$ covariates and satisfies assumptions (iii)-(iv) presented in Section 2.1. Under the same conditions as in Theorem 2.1, the empirical Bayes posterior distribution corresponding to the prior in (2.2) contracts at rate $e_n^{-1} = n^{-\beta_0/(2\beta_0 + d^0)}(\log n)^t$, with $d^0 := d_0^x + d_y$ and $t > 0$ a suitable constant.

The proof follows the same trail as that of Theorem 2.1, the only relevant difference stemming from the prior concentration rate which turns out to depend on the dimension $d_0^x$ of the covariates of $f_0$ because, for all the locations of the approximating Gaussian mixture, when $k > d_0^x$, the components $\mu_{jk} = 0$ so that in effect the mixture does not depend on the covariates $x_k$, with $k = d_0^x + 1, \ldots, d_x$. 

8
Acknowledgements

Bocconi University is gratefully acknowledged for providing financial support.

References

Bertin, K., Lacour, C., and Rivoirard, V. (2013). Adaptive pointwise estimation of conditional density function. http://arxiv.org/pdf/1312.7402.pdf.

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

Efromovich, S. (2007). Conditional density estimation in a regression setting. The Annals of Statistics, 35(6):2504–2535.

Efromovich, S. (2010). Oracle inequality for conditional density estimation and an actuarial example. Annals of the Institute of Statistical Mathematics, 62(2):249–275.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association, 99(468):1015–1026.

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

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

Petrone, S., Rizzelli, S., Rousseau, J., and Scricciolo, C. (2014a). Empirical Bayes methods in classical and Bayesian inference. METRON, 72(2):201–215.

Petrone, S., Rousseau, J., and Scricciolo, C. (2014b). Bayes and empirical Bayes: do they merge? Biometrika, 101(2):285–302.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Analysis, 6(1):145–177.

Scricciolo, C. (2014). Bayesian adaptation. Journal of Statistical Planning and Inference (in press).
Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.