An Alternate Parameterization for Bayesian Nonparametric / Semiparametric Regression

Joshua C.C. Chan
University of Technology Sydney

Justin L. Tobias
Purdue University

May 2018[1]

Abstract

We present a new procedure for nonparametric Bayesian estimation of regression functions. Specifically, our method makes use of an idea described in Frühwirth-Schnatter and Wagner (2010) to impose linearity exactly (conditional upon an unobserved binary indicator), yet also permits departures from linearity while imposing smoothness of the regression curves. An advantage of this approach is that the posterior probability of linearity is essentially produced as a by-product of the procedure. We apply our methods in both generated data experiments as well as in an illustrative application involving the impact of BMI (body mass index) on labor market earnings.

Keywords: Bayes, Nonparametric, MCMC, shrinkage

JEL classification: C11, I10, J11

[1]All errors are, of course, our own. Email:joshuacc.chan@gmail.com and jltobias@purdue.edu.

1 Introduction

Applied work in economics is increasingly characterized by a need / desire to flexibly represent various relationships: adaptable modeling of error distributions, flexible parameterizations of distributions describing the nature of parameter heterogeneity, and the modeling of regression functions without imposing rigid and potentially inappropriate parametric forms. In this paper we continue in this spirit and focus specifically on methods for modeling regression functions.

We describe a Bayesian approach for the estimation of such relationships, and the methods we introduce share similarities to those previously employed in the literature. One strand of this research - very closely related to what we describe here - introduces priors that impose similarity in local values of the regression functions, thereby producing smoothed, although potentially nonlinear, regression curves. Contributors to this literature include Dale Poirier, who we honor with this volume, as well as a number of other authors (e.g., Koop and Poirier (2004), Koop et al. (2005), Koop and Tobias (2006), Chib and Greenberg (2007), Kline and Tobias (2008) and Chib et al. (2009)). Alternate procedures make use of spline / basis function methods, often taking care to determine the number and location of knots as well as the selection of variables to be included in such representations (e.g., Smith and Kohn (1996), Smith and Kohn (2000), Chib and Greenberg (2010), Kohn et al. (2001), Chib and Greenberg (2013)).

The methods we introduce in this paper can be interpreted as a version of a smoothness prior. However, unlike existing applications of such methods, we separately consider the case of a linear model via the introduction of a latent indicator variable; when this variable equals zero, the model imposes linearity, and when it is one, a traditional smoothness prior is imposed. This idea adapts a similar recommendation made in the novel work of Frühwirth-Schnatter and Wagner (2010) in the context of state-space models.

We argue that there are several advantages to this method. We show in generated data experiments that the methods, with a given prior, can perform well when the true regression curve is either linear or nonlinear. Results are, however, not surprisingly potentially dependent on the prior - strong priors can either oversmooth nonlinear relationships or undersmooth linear (or nearly linear) ones. Unlike other related approaches, we employ a truncated Gaussian prior for our smoothing parameter rather than traditional inverse gamma specifications that can be undesirably (and unwittingly) informative. Finally,

objects of interest - namely the posterior probability of linearity - are directly produced as a by-product of our algorithm. Tests for linearity using other approaches, by comparison, may often require marginal likelihood calculation or, perhaps, the calculation of a Savage-Dickey density ratio. In contrast, our method directly calculates such a statistic within the scheme for posterior simulation and directly reports evidence of linearity / nonlinearity under the priors employed.

The outline of this paper is as follows. Section 2 describes the model, our proposed method for estimating the regression function and the associated posterior simulator. Although this presentation is offered for just a univariate nonparametric regression problem, the techniques described easily adapt to handle, for example, partially linear specifications, additive models, or systems-of-equations analyses with several unknown regression functions. Section 3 presents some generated data experiments, while an illustrative application is given in section 4. The paper concludes with a summary in section 5.

2 The Model, Parameterization and Posterior Simulator

To fix ideas, consider the following univariate nonparametric regression:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

with $(\boldsymbol{\epsilon} \,|\, \mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. We fix ideas on the baseline specification in (1), although we note that the modularity of MCMC implies that methods described below can be easily adapted to more general settings, including partially linear models, limited dependent variable specifications and systems of equations analysis.

Suppose there are $k \leq n$ distinct x_i values in the sample and denote these as $x_1^*, \ldots, x_k^*$ with $x_1^* < \cdots < x_k^*$. Treat each functional value $f(x_i^*)$ as a parameter to be estimated, and let $\theta_i = f(x_i^*)$. Next, stack the k functional values $\{\theta_i\}_{i=1}^k$ into a vector as follows: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) = (f(x_1^*), \ldots, f(x_k^*))'$. Finally, define $\mathbf{D}$ as the $n \times k$ selection matrix that maps each observation to its corresponding functional value $f(x_i^*)$. Each row of $\mathbf{D}$ contains one and only one unit entry and all other values are zero, with the unit entry positioned to select the appropriate $f(x_i^*)$.

With this construction, the nonparametric regression model can be written in typical

regression form:

$$\boldsymbol{y} = \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad (\boldsymbol{\epsilon} \,|\, \mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n). \tag{2}$$

A proper prior can be employed for $\boldsymbol{\theta}$ to smooth the regression curve; we consider a prior that expresses the view that adjacent values of $\boldsymbol{\theta}$ should be "close." To this end, we define $\Delta_j = x_j^* - x_{j-1}^*$, $j = 2, 3, \ldots, k$, and construct a $k \times k$ matrix $\mathbf{G}$ as follows:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \Delta_2^{-1} & -(\Delta_2^{-1} + \Delta_3^{-1}) & \Delta_3^{-1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \Delta_3^{-1} & -(\Delta_3^{-1} + \Delta_4^{-1}) & \Delta_4^{-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \Delta_{k-1}^{-1} & -(\Delta_{k-1}^{-1} + \Delta_k^{-1}) & \Delta_k^{-1} \end{bmatrix}.$$

Note that $\mathbf{G}\boldsymbol{\theta}$ is a $k \times 1$ vector whose first two entries are $[f(x_1)\ f(x_2)]$ and the remaining entries are differences in pointwise slopes of the form $(\mathbf{G}\boldsymbol{\theta})_j = f'(x_j) - f'(x_{j-1})$, $j = 3, 4, \ldots, k$, where $f'(x_j) \equiv [f(x_j) - f(x_{j-1})]/[x_j - x_{j-1}]$. When the final $k-2$ elements of $\mathbf{G}\boldsymbol{\theta}$ are set to zero, the model becomes linear, with the first two elements of $\mathbf{G}\boldsymbol{\theta}$ pinning down the slope and intercept of the line.

We proceed by employing a prior on $\mathbf{G}\boldsymbol{\theta}$ that serves to smooth the regression curve. As a starting point, consider a Gaussian prior on $\mathbf{G}\boldsymbol{\theta}$ given as

$$(\mathbf{G}\boldsymbol{\theta} \,|\, \mathbf{a}, \tau^2) \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_{\mathbf{a}}, \tau^2 \mathbf{I}_k) \Longleftrightarrow \mathbf{G}\boldsymbol{\theta} = \widetilde{\boldsymbol{\mu}}_{\mathbf{a}} + \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}_k), \tag{3}$$

where $\mathbf{a} = (a_1, a_2)'$ and $\widetilde{\boldsymbol{\mu}}_{\mathbf{a}} = (a_1, a_2, 0, \ldots, 0)'$ with a_1 and a_2 being unknown parameters. The prior in (3) thus centers the model over a linear specification, with a_1 and a_2 governing the slope and intercept of the line; the parameter τ^2 controls how tightly the model is centered over a linearity.

As described below, priors are also employed over the hyperparameters a_1, a_2 and τ^2. Of course, in practice, the prior employed on the parameter τ^2 will have a large impact on the smoothness of the function f, and is akin to the familiar bandwidth parameter used in classical kernel-based regression. In the limiting case where $\tau^2 = 0$, the model is linear, and previous applications of this method have interpreted "small" τ as evidence in favor of linearity. Large values of τ, on the other hand, can produce curves that are erratic and essentially connect the scatterplot of data points.

Our method departs from previous applications of the traditional smoothness prior approach in two ways: First, we separately consider the case of a linear specification, following an idea described in Frühwirth-Schnatter and Wagner (2010) in the context of

state-space models. In particular, we employ a variant of the prior in (3) and instead specify:

$$\mathbf{G}\boldsymbol{\theta} = \widetilde{\boldsymbol{\mu}}_{\mathbf{a}} + d\tau\mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k), \tag{4}$$

where $d \in \{0.1\}$ is a binary variable, to be estimated in the model. When $d = 0$, note that $\mathbf{G}\boldsymbol{\theta} = \widetilde{\boldsymbol{\mu}}_{\mathbf{a}}$, thus reproducing the linear model exactly. The prior in (4) can be expressed equivalently as:

$$(\boldsymbol{\theta} \,|\, \mathbf{a}, d, \tau) \sim \begin{cases} \boldsymbol{\delta}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_{\mathbf{a}}) & \text{if } d = 0 \\ \mathcal{N}\left[\boldsymbol{\mu}_{\mathbf{a}}, \tau^2(\mathbf{G}'\mathbf{G})^{-1}\right] & \text{if } d = 1 \end{cases} \iff \boldsymbol{\theta} = \boldsymbol{\mu}_{\mathbf{a}} + d\tau\boldsymbol{\gamma}, \tag{5}$$

where $\delta_x(z)$ is a delta function, equal to one when $x = z$ and is otherwise zero, $\boldsymbol{\mu}_{\mathbf{a}} = \mathbf{G}^{-1}\widetilde{\boldsymbol{\mu}}_{\mathbf{a}}$ and $\boldsymbol{\gamma} = \mathbf{G}^{-1}\mathbf{v} \sim \mathcal{N}(\mathbf{0}, (\mathbf{G}'\mathbf{G})^{-1})$.

Our second departure from the traditional approach relates to the prior employed on τ: we employ a truncated Gaussian prior instead of the more typical inverse-gamma specification. Frühwirth-Schnatter and Wagner (2010) argue that the conventional inverse-gamma prior for the smoothing parameter τ^2 is often too informative and distorts information from the likelihood. Instead, they adopt a normal prior for τ, which implies a gamma prior $\mathcal{G}(0.5, 0.5/V_\tau)$ on the variance τ^2. The sign of τ, however, is not identified in their scheme.

Bounding the prior away from zero via a truncated normal prior in our setting allows for an approximate separation of the linear and nonlinear alternatives; if either an inverse gamma or truncated normal prior is employed, and $\tau \approx 0$, the model is essentially linear, thus creating a redundancy where a linear model is effectively reproduced when either $d = 0$ or $d = 1, \tau \approx 0$. By specifying our prior in this way, the $d = 1$ regime places most of its prior mass over values of τ that imply departures from linearity.

The model is completed upon specifying priors for the remaining parameters. These are described below:

$$\begin{aligned} \mathbf{a} = (a_1, a_2)' &\sim \mathcal{N}(\mathbf{a}_0, \mathbf{V}_{\mathbf{a}}), && (6) \\ \sigma^2 &\sim \mathcal{IG}(\nu_{\sigma^2}, S_{\sigma^2}), && (7) \\ d &\sim Bern(\underline{p}) && (8) \\ \tau &\sim \mathcal{TN}_{(\underline{\tau},\infty)}(\mu_\tau, V_\tau), && (9) \end{aligned}$$

with $\mathbf{a}_0 = \mathbf{0}, \mathbf{V}_{\mathbf{a}} = 100\mathbf{I}_2, \nu_{\sigma^2} = 5, S_{\sigma^2} = 4, b_0 = c_0 = 0.5$, $\mu_\tau = 0$ and $V_\tau = 0.1$. In the above, $\mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 that

is truncated to the interval (a, b), $Bern(p)$ denotes a Bernoulli distribution with success probability p and $\mathcal{B}(a, b)$ denotes a Beta distribution with parameters a and b.

2.1 Posterior Simulation

The joint posterior distribution associated with this model follows from the likelihood implied by (2) and priors in (5) and (6) - (9). Specifically,

$$p(\boldsymbol{\theta}, d, \tau, \mathbf{a}, \sigma^2 \,|\, \mathbf{y}) \propto \phi\left(\mathbf{y}; \mathbf{D}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n\right) f(\boldsymbol{\theta} \,|\, \mathbf{a}, d, \tau) f(d \,|\, \underline{p}) f(\mathbf{a}) f(\sigma^2) f(\tau). \tag{10}$$

Posterior simulation is accomplished via the Gibbs sampler, and involves sequentially sampling from:

1. $f(\boldsymbol{\theta} \,|\, \mathbf{y}, d, \tau, \sigma^2, \mathbf{a}) = f(\boldsymbol{\theta} \,|\, \mathbf{y}, d, \tau, \sigma^2, \mathbf{a})$;
2. $f(d, \tau \,|\, \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \mathbf{a}) = f(d \,|\, \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \mathbf{a}) f(\tau \,|\, \mathbf{y}, \boldsymbol{\theta}, d, \sigma^2, \mathbf{a})$;
3. $f(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\theta}, d, \tau, \mathbf{a}) = f(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\theta})$;
4. $f(\mathbf{a} \,|\, \mathbf{y}, \boldsymbol{\theta}, d, \tau, \sigma^2) = f(\mathbf{a} \,|\, \mathbf{y}, \boldsymbol{\theta}, d, \tau, \sigma^2)$.

We describe each of these steps, in order, below.

Step 1. To implement Step 1, first note (5) implies $\boldsymbol{\theta}$ becomes degenerate when $d = 0$, precluding the adoption of what might be considered a "standard" posterior simulator. To sidestep this problem, we follow the idea discussed in Frühwirth-Schnatter and Wagner (2010) and will, instead, sample $\boldsymbol{\gamma}$ conditional on the data and other parameters, and use the sampled value of $\boldsymbol{\gamma}$ to calculate $\boldsymbol{\theta}$.

To that end, we substitute (5) into (2) to get:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}} + d\tau \mathbf{D}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \qquad (\boldsymbol{\epsilon} \,|\, \mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where $\boldsymbol{\mu}_{\mathbf{a}} = \mathbf{G}^{-1}\widetilde{\boldsymbol{\mu}}_{\mathbf{a}}$ with $\widetilde{\boldsymbol{\mu}}_{\mathbf{a}} = (a_1, a_2, 0, \ldots, 0)'$, and $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, (\mathbf{G}'\mathbf{G})^{-1})$. Using standard results from linear regression, we have,

$$(\boldsymbol{\gamma} \,|\, \mathbf{y}, d, \tau, \sigma^2, \mathbf{a}) \sim \mathcal{N}(\widehat{\boldsymbol{\gamma}}, \mathbf{K}_{\boldsymbol{\gamma}}^{-1}), \tag{11}$$

where

$$\mathbf{K}_{\boldsymbol{\gamma}} = \mathbf{G}'\mathbf{G} + \frac{d\tau^2}{\sigma^2}\mathbf{D}'\mathbf{D}, \qquad \widehat{\boldsymbol{\gamma}} = \frac{d\tau}{\sigma^2}\mathbf{K}_{\boldsymbol{\gamma}}^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}}).$$

One difficulty of drawing from the above Gaussian distribution is that $\boldsymbol{\gamma}$ is typically high-dimensional. Consequently, the conventional sampling approach that requires the Cholesky factor of the covariance matrix $\mathbf{K}_{\boldsymbol{\gamma}}^{-1}$ is time-consuming - perhaps prohibitively so - particularly when $k \approx n$. However, since both $\mathbf{G}$ and $\mathbf{D}$ are band matrices, so is the precision matrix $\mathbf{K}_{\boldsymbol{\gamma}}$. Therefore, one can efficiently sample from $\mathcal{N}(\widehat{\boldsymbol{\gamma}}, \mathbf{K}_{\boldsymbol{\gamma}}^{-1})$ using band matrix routines as proposed in Chib, Greenberg, and Jeliazkov (2009), and we employ those here.

Finally, given the draw $\boldsymbol{\gamma}$, we can then obtain a draw of $\boldsymbol{\theta}$ using (5), given current values of $\boldsymbol{\mu}_{\mathbf{a}}$, d and τ:

$$\boldsymbol{\theta} = \boldsymbol{\mu}_{\mathbf{a}} + d\tau\boldsymbol{\gamma}. \tag{12}$$

Step 2. Since d and τ enter the likelihood multiplicatively, we sample them jointly to improve efficiency. To that end, we first sample the indicator $d \in \{0, 1\}$ marginally of τ (yet conditioned on the remaining model parameters), followed by drawing τ from its complete posterior conditional distribution. The latter of these two steps is easy: again using standard linear regression results, we have,

$$(\tau \,|\, \mathbf{y}, \boldsymbol{\theta}, d, \sigma^2, \mathbf{a}) \sim \mathcal{TN}_{(\underline{\tau},\infty)}(\widehat{\tau}, K_\tau^{-1}), \tag{13}$$

where

$$K_\tau = V_\tau^{-1} + \frac{d}{\sigma^2}\boldsymbol{\gamma}'\mathbf{D}'\mathbf{D}\boldsymbol{\gamma}, \qquad \widehat{\tau} = K_\tau^{-1}\left[\frac{d}{\sigma^2}\boldsymbol{\gamma}'\mathbf{D}'(\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}}) + V_\tau^{-1}\mu_\tau\right]. \tag{14}$$

To sample d marginal of τ, first note that

$$p(d \,|\, \mathbf{a}, \boldsymbol{\theta}, \sigma^2\mathbf{y}) \propto \left[\int_{\underline{\tau}}^{\infty} f(\mathbf{y} \,|\, d, \tau, \mathbf{a}, \boldsymbol{\theta}, \sigma^2) f(\tau)\mathrm{d}\tau\right] p(d \,|\, \underline{p}).$$

In the appendix we show that

$$\begin{aligned}\int_{\underline{\tau}}^{\infty} f(\mathbf{y} \,|\, d, \tau, \mathbf{a}, \boldsymbol{\theta}, \sigma^2) f(\tau)\mathrm{d}\tau =& (2\pi\sigma^2)^{-\frac{n}{2}} \frac{1 - \Phi((\underline{\tau} - \widehat{\tau})K_\tau)}{1 - \Phi((\underline{\tau} - \mu_\tau)V_\tau^{-1})} V_\tau^{-\frac{1}{2}} K_\tau^{-\frac{1}{2}} \\ &\times \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}})'(\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}}) - \frac{1}{2}V_\tau^{-1}\mu_\tau^2 + \frac{1}{2}K_\tau\widehat{\tau}^2},\end{aligned} \tag{15}$$

where $\widehat{\tau}$ and K_τ are defined in (14). When $d = 0$, $K_\tau = V_\tau^{-1}$ and $\widehat{\tau} = \mu_\tau$, and it follows that

$$\Pr(d = 0 \,|\, \mathbf{a}, \boldsymbol{\theta}, \sigma^2\mathbf{y}) \propto (1 - \underline{p})(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}})'(\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_{\mathbf{a}})\right). \tag{16}$$

Similarly,

$$\Pr(d=1\,|\,\mathbf{a},\boldsymbol{\theta},\sigma^2\mathbf{y}) \propto \underline{p}(2\pi\sigma^2)^{-\frac{n}{2}}V_\tau^{-\frac{1}{2}}K_\tau(1)^{-\frac{1}{2}}\frac{1-\Phi((\underline{\tau}-\widehat{\tau}(1))K_\tau(1))}{1-\Phi((\underline{\tau}-\mu_\tau)V_\tau^{-1})} \times \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_\mathbf{a})'(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_\mathbf{a})-\frac{1}{2}V_\tau^{-1}\mu_\tau^2+\frac{1}{2}K_\tau(1)\widehat{\tau}(1)^2}, \tag{17}$$

where $\widehat{\tau}(1)$ and $K_\tau(1)$ denote, respectively, $\widehat{\tau}$ and K_τ evaluated at $d=1$. A draw can easily be obtained from this conditional posterior, by first normalizing the probabilities in (16) and (17) and then sampling from the resulting two-point distribution.

Step 3. Sampling σ^2 from its full conditional distribution is standard. Specifically, let $\boldsymbol{\epsilon}=\mathbf{y}-\mathbf{D}\boldsymbol{\theta}$. Then,

$$(\sigma^2\,|\,\mathbf{y},\boldsymbol{\theta}) \sim \mathcal{IG}\left(\nu_{\sigma^2}+\frac{n}{2}, S_{\sigma^2}+\frac{1}{2}\sum_{i=1}^{n}\epsilon_i^2\right), \tag{18}$$

where ϵ_i is the ith element of $\boldsymbol{\epsilon}$.

Step 4. To derive the conditional distribution of $\mathbf{a}$, first rewrite (5) as $\boldsymbol{\theta}=\mathbf{X}_\mathbf{a}\mathbf{a}+d\tau\boldsymbol{\gamma}$ and substitute it into (2) to get:

$$\mathbf{y}=\mathbf{D}\mathbf{X}_a\mathbf{a}+d\tau\mathbf{D}\boldsymbol{\gamma}+\boldsymbol{\epsilon}, \qquad (\boldsymbol{\epsilon}\,|\,\mathbf{X}) \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_n),$$

where $\mathbf{X}_\mathbf{a}=\mathbf{G}^{-1}\begin{pmatrix}\mathbf{I}_2\\ \mathbf{0}\end{pmatrix}$. Then, using standard linear regression results, we have

$$(\mathbf{a}\,|\,\mathbf{y},\boldsymbol{\theta},d,\tau,\sigma^2) \sim \mathcal{N}(\widehat{\mathbf{a}},\mathbf{K}_\mathbf{a}^{-1}), \tag{19}$$

where

$$\mathbf{K}_\mathbf{a}=\mathbf{V}_\mathbf{a}^{-1}+\frac{1}{\sigma^2}\mathbf{X}_\mathbf{a}'\mathbf{D}'\mathbf{D}\mathbf{X}_\mathbf{a}, \qquad \widehat{\mathbf{a}}=\mathbf{K}_\mathbf{a}^{-1}\left(\mathbf{V}_\mathbf{a}^{-1}\mathbf{a}_0+\frac{1}{\sigma^2}\mathbf{X}_\mathbf{a}'\mathbf{D}'(\mathbf{y}-d\tau\mathbf{D}\boldsymbol{\gamma})\right).$$

Posterior simulation proceeds by sampling from (11), drawing from the binary distribution obtained by normalizing (16) and (17), drawing the smoothing parameter from (13) and then sampling from (18) and (19).

3 Generated Data Experiments

In this section we perform a few generated data experiments. These experiments are conducted with the goals of: (a) providing evidence that the model and associated posterior simulator for model fitting perform well at recovering parameters of the true data

generating process when those are known, (b) showing that the methods generally accommodate both linear and nonlinear relationships when either is present in the data, and (c) arguing for the value of an automatically-produced posterior probability of linearity, as measured by $\Pr(d = 0 \,|\, \mathbf{y})$, and noting how some degree of caution and care should be given in that interpretation.

In the first set of generated data experiments, we simulate $n = 200$ observations separately from both linear and nonlinear specifications. Specifically, we consider two models:

$$y_i = .15x_i + .3\exp[-4(x_i+1)^2] + .7\exp[-16(x_i-1)^2] + \epsilon_i, \quad x_i \sim \mathcal{U}[-2,2], \quad \epsilon_i \sim \mathcal{N}(0,.01)$$

and

$$y_i = 2 + x_i + \epsilon_i, \quad x_i \sim \mathcal{U}[0,20], \quad \epsilon_i \sim \mathcal{N}(0,20),$$

with the nonlinear specification taken from Fan and Gijbels (1996) and DiNardo and Tobias (2001).

The sampler in both cases is run for 31,000 iterations, with the first 1,000 of those discarded as the burn-in period. We do not discuss convergence diagnostics in detail here, other than to note that the sampler, perhaps not surprisingly given its similarity in structure to a linear regression, converges very quickly to explore the joint posterior distribution, with the joint sampling of $(\tau, d| \cdot \mathbf{y})$ offering a considerable improvement in mixing performance. Posterior simulation in both cases takes under 11 seconds on a reasonably standard PC, noting that $\mathbf{D}$ in this case is a 200×200 matrix.

In both experiments we choose hyperparameters, as described below (9) and just before section 2.1. Prior hyperparameter choices that have the largest influence on posterior results are those made regarding τ in (9). In this capacity, we select $\mu_\tau = 0$, $V_\tau = .1$ and $\underline{\tau} = .05$, and employ this prior for both the linear and nonlinear experiment. Estimation results are provided in Figure 1. As one can see from the figure, the point estimate (posterior mean) of the function f adapts under this prior to accommodate both the nonlinear (upper graph) and linear (lower graph) specifications. Finally, for each experiment, we also calculate the posterior probability of linearity, as measured by

$$\widehat{Pr}(d = 0 \,|\, \mathbf{y}) = \frac{1}{30,000} \sum_{i=1}^{30,000} d_i,$$

where d_i represents the i^{th} post-converge draw produced from our simulator. In the nonlinear experiment, each of the posterior simulations yield $d_i = 1$, so that the estimated

probability of linearity is exactly zero. In the linear experiment, 72 of the 30,000 posterior simulations were associated with $d_i = 1$, yielding a posterior probability of linearity equal to .998.

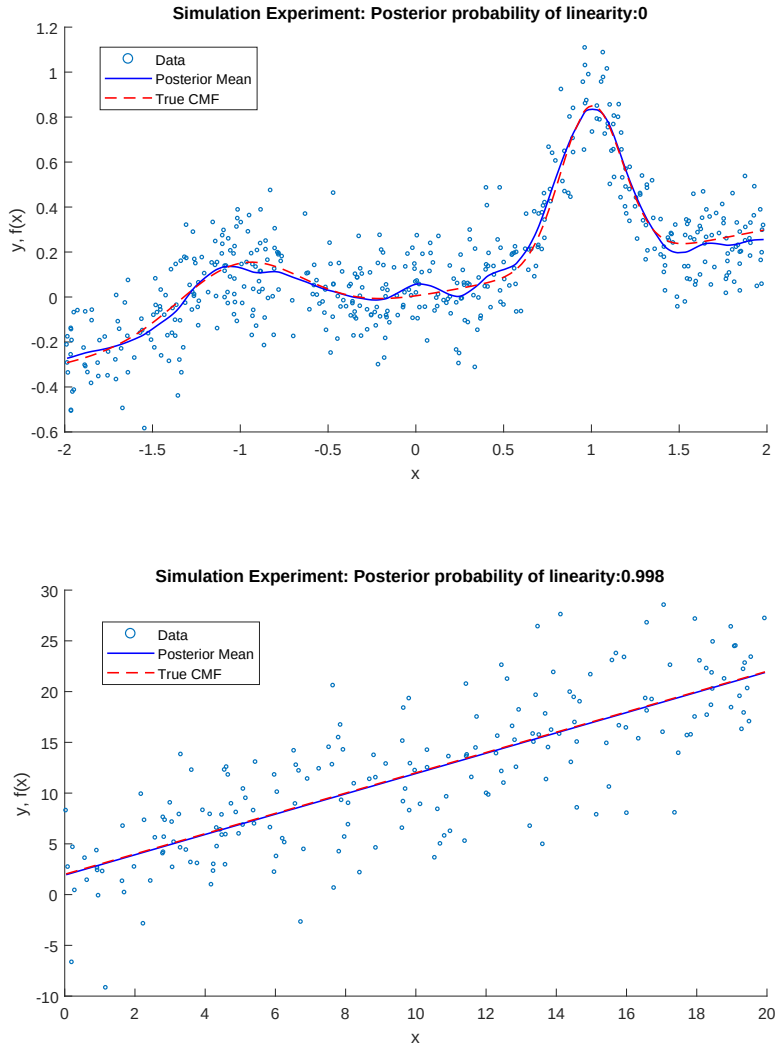

Figure 1: Results of Linear and Nonlinear Generated Data Experiments.

We move on to present a few additional generated data experiments. In these experiments, we generate data when $f(x)$ is quadratic and look to determine how well our

model performs at detecting the presence of departures from linearity, when different degrees of nonlinearity are present in the data. The experiments that follow are sampling experiments: we generate 200 observations from a quadratic model 100 different times. For each 200-observation data set, we record our estimate (posterior mean) of $E(d \,|\, \mathbf{y})$, $Std(d \,|\, \mathbf{y})$ and whether or not a classical test rejects, at the 5% level, the null hypothesis that the quadratic coefficient equals zero. Data are generated from the following specification:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \epsilon_i, \quad x_i \overset{iid}{\sim} \mathcal{U}[0, 20], \quad (\boldsymbol{\epsilon} \,|\, \mathbf{X}) \sim \mathcal{N}(\mathbf{0}, 20\mathbf{I}_{200}).$$

In the experiments, $\alpha_0 = 2, \alpha_1 = 1$ and $\alpha_2 \in \{0, .01, .03, .05, .10\}$. Results from these sampling experiments are presented in the table below:

Table 1: Results of Sampling Experiments Across Different Degrees of Nonlinearity

	$\alpha_2 = 0$	$\alpha_2 = .01$	$\alpha_2 = .03$	$\alpha_2 = .05$	$\alpha_2 = .10$
$E_{y \,\mid\, \alpha}[\widehat{\Pr(d = 1 \,\mid\, \mathbf{y})}]$	.005	.018	.172	.512	.999
$E_{y \,\mid\, \alpha}[\widehat{\text{Std}(d = 1 \,\mid\, \mathbf{y})}]$	.058	.078	.257	.309	.022
Reject $H_{0,\alpha=.05} : \alpha_2 = 0$	.03	.15	.60	1	1

As suggested by results in the table, our algorithm performs well at identifying a linear specification when the model is, in fact, linear: the (sampling) average posterior probability that $d = 0$ when $\alpha_2 = 0$ is .005. This, in fact, outperforms the classical test which, by construction, will reject the null when true 5 percent of the time. Similarly, the average posterior standard deviation of .058 indicates that the marginal posterior of d tends to be reasonably tightly concentrated about zero in these cases. As α_2 increases, and commensurately the degree of nonlinearity increases, the model begins to detect these departures: when α_2 is as large as .10, the marginal posterior of d becomes nearly concentrated about $d = 1$.

Further inspection of the table shows that in the intermediate cases the model has difficulty, as expected, in detecting small degrees of nonlinearity. When $\alpha_2 = .05$, for example, the (average) posterior probability of linearity is .512; these results essentially leave the researcher positioned uncomfortably in the middle of supporting a linear or nonlinear specification. In contrast, the classical results reject each null hypothesis that $\alpha_2 = 0$ in this experiment, suggesting a greater degree of power in the classical test than the Bayesian counterpart. While one could argue that the classical test both assumes correct specification in the unrestricted model and the comparison between two parametric

alternatives in the classical case and a linear versus a nonparametric one in the Bayesian case are answering different questions, the degree of discrepancy between the approaches may remain unsettling to some.

In this regard we refer, again, the important role of the prior in these calculations. Our specification seeks to interpret $d = 0$ as the linear case; in forming the alternative model, large values of τ may produce a prior that places a lot of mass over specifications that exhibit high degrees of nonlinearity. It may be the case that the data will support the $d = 0$ (linear) regime when nonlinearities are, in fact, present, if the alternative when $d = 1$ steers the analysis toward something that is excessively nonlinear. To this end, we repeat the sampling experiment with $\alpha_2 = .05$, this time under a prior where $\underline{\tau} = .001$ and $V_\tau = .05$. In this case, the average sampling probability that $d = 1$ jumps to .96, and the average posterior standard deviation of d is .12. These results move significantly closer to the classical procedure, where linearity was rejected in each case.

The movement of the prior toward "smaller" values of τ is not, however, without cost. The linear model can be exactly reproduced within the $d = 0$ regime or approximately reproduced when $(d = 1, \tau \approx 0)$. This, in turn, creates a practical identification problem and may call into question the use of $\Pr(d = 1 \,|\, \mathbf{y})$ as the posterior probability of linearity. Indeed, we find that this is the case in an additional sampling experiment: Keeping the same prior with $\underline{\tau} = .001$ and $V_\tau = .05$ but now setting $\alpha_2 = 0$ (i.e., linearity), we obtain $E_{\mathbf{y}\,|\,\mathbf{r}}[\widehat{\Pr(d = 1 \,|\, \mathbf{y})}] = .20$, and $E_{\mathbf{y}\,|\,\mathbf{r}}[\widehat{\mathrm{Std}(d = 1 \,|\, \mathbf{y})}] = .37$. Although something as simple as a plot of posterior results may still lead the researcher toward a linear conclusion, simple use of d as evidence in favor of linearity / nonlinearity may lead the researcher astray. As a result, we suggest adoption of a prior that works to differentiate the linear and nonlinear alternatives. The downside to this approach may be that probabilities of nonlinearity are conservative if the prior puts too much distance between the linear specification and nonlinear alternatives. The methods are not fully automatic, and care must be taken with the choice of prior in these "intermediate" cases.

4 Illustrative Example: The Effect of BMI on Earnings

In this section we use our methods to briefly investigate the impact of BMI, or body mass index, on labor market earnings. The data that we employ is the same as that analyzed by Kline and Tobias (2008), which is taken from the 1970 British Cohort Study.

This data set tracks outcomes for all people born in Great Britain between April 5 and April 11, 1970; we use labor market outcomes for those individuals present during the 1999-2000 interview wave, when the respondents were between 29 and 30 years of age. The data set consists of 4,343 observations in total (2,561 observations for men and 1,782 for women).

The basic model relates log wages (in pounds) to a variety of demographic variables. These include discrete education controls that consist of a completion of high school degree indicator ($HighSchool$), a separate $Alevel$ indicator (denoting that the respondent passed at least one Alevel exam) and $Degree$ (indicating that the respondent completed a college degree program). In addition, we include a marriage indicator, a quadratic in months of tenure at the current job and a quadratic in potential labor market experience.

Our primary variable of interest is BMI, or body mass index, defined as weight (in kilograms) divided by the square of height (in meters). In the sample we employ, we investigate how BMI relates to log earnings over a support of [18,36]; "normal weight" is typically defined by BMI values in the interval $[18.5, 25]$, while values in excess of 30 are considered obese. We allow the the relationships between BMI and earnings to differ across gender, and consider a specification of the following form:[2]

$$\log w_i = f_1(BMI_i) + m_i f_2(BMI_i) + \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i, \quad (\boldsymbol{\epsilon} \,|\, \mathbf{X}, BMI) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n).$$

In the above m_i is a binary variable, denoting if the respondent is male: the conditional mean function for men is therefore given by $f_1 + f_2$, while the conditional mean function for women is f_1. The variables in $\mathbf{X}$ include those mentioned above: categorical education variables, experience and tenure variables and a marriage indicator.

The data set yields 1,268 unique BMI values, thus creating a reasonably high-dimensional matrix $\mathbf{D}$. Typically this introduces some computational challenges, given the need to calculate a high-dimensional inverse at each iteration, but these typical impediments are be mitigated here using band matrix calculations, as described previously. Since the model in this application introduces two regression functions, we introduce two priors to smooth the respective curves. Priors of the form in (5) are employed to smooth each regression function, and we denote the associated binary variables as d_1, d_2 and τ_1, τ_2, corresponding to the functions f_1 and f_2 and associated function vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

[2]Kline and Tobias (2008) consider the potential endogeneity of BMI and use parental BMI as instruments. We do not consider the endogeneity problem here simply to fix ideas on implementation of our approach to nonparametric regression. Kline and Tobias (2008) also run separate analyses for men and women; we pool the results in our specification but allow for different regression functions across gender.

Estimation results are presented in Figure 1 and Table 2. These results are obtained after running the posterior simulator for 31,000 iterations and discarding the first 1,000 of those draws. For these results, we employ a prior that sets $\mu_\tau = 0$, $\underline{\tau} = .0001$ and $V_\tau = .1$. Computations are produced, again, on a standard PC, and are completed in a little under 2.5 minutes.

Table 2 reveals evidence of a quadratic relationship in job tenure and experience, a monotonic relationship in educational attainment and shows that married individuals earn about 4.6 percent more than those who are not married. Posterior statistics associated with d_1 and d_2 provide strong evidence of a linear conditional mean for women (as the posterior probability of linearity is approximately .992 and is robust to a variety of prior choices), and modest evidence of nonlinearity for males, given that $\Pr(d_2 = 1 \,|\, \mathbf{y}) \approx .78$.

Table 2: Parameter posterior means, standard deviations and probabilities begin positive for the BMI example.

Wage Equation			
Variable	$\mathrm{E}(\cdot \,|\, \mathbf{y})$	$\sqrt{\mathrm{Var}(\cdot \,|\, \mathbf{y})}$	$\Pr(\cdot > 0 \,|\, \mathbf{y})$
JobTenure	0.025	0.0053	1.00
JobTenure^2	-0.001	0.0004	0.001
Experience	0.029	0.010	0.997
Experience^2	-0.001	0.0006	0.051
FamilyIncome	0.001	0.0001	1.00
HighSchool	0.068	0.015	1.00
ALevel	0.310	0.034	1.00
Degree	0.408	0.031	1.00
Married	0.046	0.012	1.00
Other Parameters			
Variable	$\mathrm{E}(\cdot \,|\, \mathbf{y})$	$\sqrt{\mathrm{Var}(\cdot \,|\, \mathbf{y})}$	$\Pr(\cdot > 0 \,|\, \mathbf{y})$
τ_f	0.179	0.137	1.00
τ_m	0.041	0.097	1.00
d_f	0.008	0.089	0.008
d_m	0.783	0.412	0.783
σ^2	.149	.003	1.00

Figure 2 plots the estimated regression functions for both men and women across the BMI support. In the figure, we also present 95% posterior probability intervals, but do so only for males simply for the sake of clarity and to minimize clutter within the graph. Interestingly, the shape of the estimated log earnings - BMI relationship is negative and

linear for women: women appear to be penalized for increments to BMI throughout the support of the BMI distribution. For men on the other hand, expected log earnings actually increase with BMI over the left-tail of the ability distribution, but then decline, with the location of the downturn approximately near the upper limit of the "normal weight" category. This pattern is consistent with some previous findings in the literature: Cawley (2004), for example, finds a negative impact of BMI on the wages of white women, while McLean and Moon (1980) find evidence of a BMI wage premium for men, which they term, and the literature has since adopted, the "portly banker effect". Such a positive effect for men is also documented more recently by Majumder (2013).

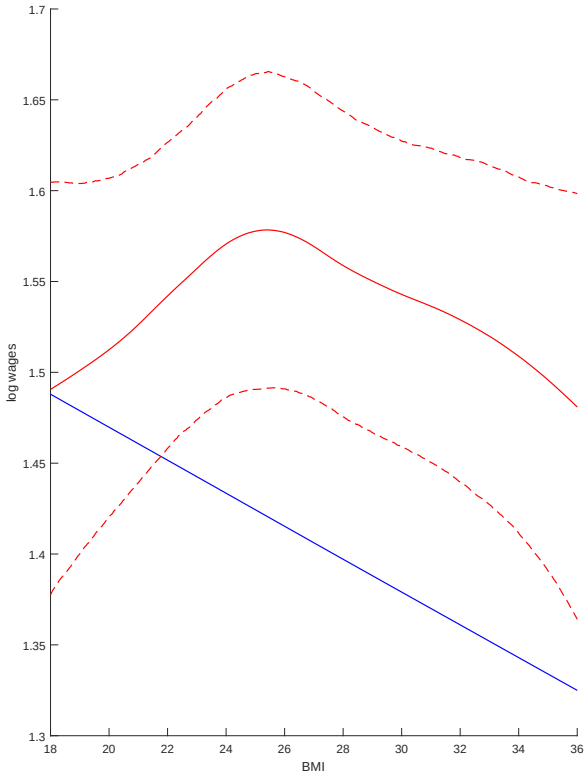

Figure 2: Estimated Regression Functions for Men and Women.

With a little further investigation, we can put this positive marginal effect for males into context. To this end, let us focus our attention on estimation of the average derivative, defined as $E_x[f'(x)] \approx \frac{1}{n}\sum_i f'(x_i)$. Each posterior simulation of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be used

to obtain a simulated value of this average derivative, as follows:

$$\widehat{AvgDeriv}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{m-1} n_i \left[\frac{\theta_{i+1}-\theta_i}{BMI_{i+1}-BMI_i}\right]. \tag{20}$$

In the above formula, recall that m denotes the number of distinct BMI values in the sample. The variable n_i recognizes that some BMI values occur with multiplicity in the data, since some individuals have exactly the same height and weight measurements. Finally, $n = \sum_{i=1}^{n_i}$ denotes the total number of observations in the data, excluding those possibly clustered at the smallest observed value. It is understood that the average derivative in (20) is calculated in this way for both females and males, with the $\boldsymbol{\theta}_1$ simulations used to calculate the former and $(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)$ used to calculate the latter. The average derivative in (20) is a function of $\boldsymbol{\theta}$, and thus a posterior distribution associated with this average derivative can be obtained via simulation: this quantity is calculated for each post-convergence draw, from which a variety of statistics can be obtained.

Interestingly, we note that the average derivative for women has a posterior mean of -.009, (i.e., every point increase in BMI for women leads to an expected earnings decrease of slightly less than .1 percent), and its calculated value is negative for every post-convergence simulation, i.e., $\Pr(\widehat{AvgDeriv(\boldsymbol{\theta}_1)} > 0 \,|\, \mathbf{y}) = 0$. For men, the posterior mean of the average derivative is .0018 - consistent with the so-called "portly banker effect" - and approximately 70 percent of the post-convergence simulations of $AvgDeriv(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)$ are positive. Our results, thus refine this conventional wisdom and suggest that this overall positive pattern is driven by earnings increases associated with BMI increments throughout the "normal weight" range only; once a male has a BMI in excess of 26, he actually experiences a penalty in the labor market for further increases to BMI. Said differently: our banker would not be happy at all should he find himself portly, although he is certainly happy to push the upper limit of the normal-weight range. Finally, we note that $\Pr(AvgDeriv(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) > AvgDeriv(\boldsymbol{\theta}_1) \,|\, \mathbf{y}) \approx .99$, providing strong evidence that the functions relating BMI to expected log earnings differ across men and women.

5 Conclusion

We presented a new approach for Bayesian nonparametric estimation of regression functions. The novelty of our approach is to permit, within the context of a standard posterior simulator, separate consideration of a linear specification and to deliver the posterior prob-

ability of linearity as a by-product of a standard Gibbs algorithm. The priors we employ allow linearity to be treated as a separate case within the model itself; a separate prior is then placed over an additional component that places most of its mass over nonlinear alternatives. Results, of course, are sensitive to the prior employed (just as the degree of smoothing is sensitive to bandwidth choice in kernel-based regression), and care must be taken to both differentiate the model components - if $\Pr(d = 0 \,|\, \mathbf{y})$ is to be interpreted as a posterior probability of linearity - yet not differentiate them excessively so that highly nonlinear alternatives will seldom find support for the data.

We applied our methods in generated data experiments. Those results suggested that our algorithm performs well, and generally adapts to reflect the curvature of the true regression function. In addition, we applied our methods to investigate the impact of BMI on (log) wages. In so doing, we find that the BMI - log wage profile for women is linear and negative. For men, however, we find the opposite result: the overall slope (i.e., average derivative) is weakly positive, consistent with the so-called "portly banker effect" in this literature. However, this positive slope is completely driven by an increasing relationship in the left tail of the BMI distribution only. For men, marginal increments to BMI, when BMI is greater than 26, are negative.

Technical Appendix

In this appendix we show the derivation of the analytical expression given in (15).

$$\begin{aligned}
&\int_{\underline{\tau}}^{\infty} f(\mathbf{y} \mid d, \tau, \mathbf{a}, \boldsymbol{\theta}, \sigma^2) f(\tau) \mathrm{d}\tau \\
&\quad = \int_{\underline{\tau}}^{\infty} (2\pi\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_{\mathbf{a}}-d\tau\mathbf{D}\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_{\mathbf{a}}-d\tau\mathbf{D}\boldsymbol{\gamma})} \times \frac{(2\pi V_\tau)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2V_\tau}(\tau-\mu_\tau)^2}}{1-\Phi((\underline{\tau}-\mu_\tau)V_\tau^{-1})} \mathrm{d}\tau \\
&\quad = \frac{(2\pi\sigma^2)^{-\frac{n}{2}}(2\pi V_\tau)^{-\frac{1}{2}}}{1-\Phi((\underline{\tau}-\mu_\tau)V_\tau^{-1})} \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_{\mathbf{a}})'(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_{\mathbf{a}})-\frac{1}{2V_\tau}\mu_\tau^2} \times \int_{\underline{\tau}}^{\infty} \mathrm{e}^{-\frac{1}{2}\left(\tau^2 K_\tau - 2\tau K_\tau \widehat{\tau}\right)} \mathrm{d}\tau, \quad (21)
\end{aligned}$$

where K_τ and $\widehat{\tau}$ are given in (14). Next, we complete the square to get

$$\tau^2 K_\tau - 2\tau K_\tau \widehat{\tau} = (\tau - \widehat{\tau})^2 K_\tau - K_\tau \widehat{\tau}^2,$$

and it follows that

$$\int_{\underline{\tau}}^{\infty} \mathrm{e}^{-\frac{1}{2}\left(\tau^2 K_\tau - 2\tau K_\tau \widehat{\tau}\right)} \mathrm{d}\tau = (2\pi K_\tau^{-1})^{\frac{1}{2}} \mathrm{e}^{\frac{1}{2} K_\tau \widehat{\tau}^2} (1 - \Phi((\underline{\tau} - \widehat{\tau}) K_\tau)).$$

Substituting this expression back to (21) we get

$$\begin{aligned}
\int_{\underline{\tau}}^{\infty} f(\mathbf{y} \mid d, \tau, \mathbf{a}, \boldsymbol{\theta}, \sigma^2) f(\tau) \mathrm{d}\tau = &(2\pi\sigma^2)^{-\frac{n}{2}} \frac{1-\Phi((\underline{\tau}-\widehat{\tau})K_\tau)}{1-\Phi((\underline{\tau}-\mu_\tau)V_\tau^{-1})} V_\tau^{-\frac{1}{2}} K_\tau^{-\frac{1}{2}} \\
&\times \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_{\mathbf{a}})'(\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_{\mathbf{a}}) - \frac{1}{2}V_\tau^{-1}\mu_\tau^2 + \frac{1}{2}K_\tau\widehat{\tau}^2\right)
\end{aligned}$$

as claimed.

References

J. Cawley. The impact of obesity on wages. *Journal of Human Resources*, 39(2):451–474, 2004.

S. Chib and E. Greenberg. Semiparametric modeling and estimation of instrumental variable models. *Journal of Graphical and Computational Statistics*, 16:86–114, 2007.

S. Chib and E. Greenberg. Additive cubic spline regression with dirichlet process mixture errors. *Journal of Econometrics*, 156:322–336, 2010.

S. Chib and E. Greenberg. On conditional variance estimation in nonparametric regression. *Statistics and Computing*, (23):261–270, 2013.

S. Chib, E. Greenberg, and I. Jeliazkov. Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, 18:321–348, 2009.

J. DiNardo and J.L. Tobias. Nonparametric density and regression estimation. *Journal of Economic Perspectives*, 14(4):11–28, 2001.

J. Fan and I. Gijbels. *Local Polynomial Modeling and its Applications*. Chapman & Hall, 1996.

S. Frühwirth-Schnatter and H. Wagner. Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154:85–100, 2010.

B. Kline and J. L. Tobias. The wages of BMI: Bayesian analysis of a skewed treatment-response model with nonparametric endogeneity. *Journal of Applied Econometrics*, 23 (6):767–793, 2008.

R. Kohn, M. Smith, and D. Chan. Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4):313–322, 2001.

G. Koop and D. J. Poirier. Bayesian variants of some classical semiparametric regression techniques. *Journal of Econometrics*, 123(2):259–282, 2004.

G. Koop and J.L. Tobias. Semiparametric bayesian inference in smooth coefficient models. *Journal of Econometrics*, (134):283–315, 2006.

G. Koop, D. J. Poirier, and J. Tobias. Semiparametric Bayesian inference in multiple equation models. *Journal of Applied Econometrics*, 20(6):723–747, 2005.

Md. A. Majumder. Does obesity matter for wages? evidence from the united states. *Economic Papers*, pages 200–217, 2013.

R.A. McLean and M. Moon. Health, obesity and earnings. *American Journal of Public Health*, 70:1006–1009, 1980.

M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75:317–343, 1996.

M. Smith and R. Kohn. Nonparametric seemingly unrelated regression. *Journal of Econometrics*, 98(2):257–281, 2000.