Nonparametric Estimation in Economics: Bayesian and Frequentist Approaches

Joshua Chan[*], Daniel J. Henderson[†], Christopher F. Parmeter [‡], Justin L. Tobias[§]

Abstract

We review Bayesian and classical approaches to nonparametric density and regression estimation and illustrate how these techniques can be used in economic applications. On the Bayesian side, density estimation is illustrated via finite Gaussian mixtures and a Dirichlet Process Mixture Model, while nonparametric regression is handled using priors that impose smoothness. From the frequentist perspective, kernel-based nonparametric regression techniques are presented for both density and regression problems. Both approaches are illustrated using a wage data set from the Current Population Survey.

[*] Department of Economics, University of Technology Sydney.

[†] Department of Economics, University of Alabama.

[‡] Department of Economics, University of Miami.

[§] Department of Economics, Purdue University.

INTRODUCTION

Significant improvements in computing power coupled with the development of powerful new statistical methods have served to push forward the frontier of what can be accomplished in serious empirical research. While early empirical investigations in economics were significantly limited by the power of computational machinery (and to a lesser extent the development of theory), this is no longer the case. Researchers are now equipped to fit models that seek to impose as few restrictions as possible, and to use the data to uncover relationships that may commonly be misrepresented as linear or Gaussian.

A goal of this paper is to review, from both Bayesian and frequentist (classical) perspectives, several nonparametric techniques that have been employed in the economics literature, to illustrate how these methods are applied, and to describe the value of their use. In the first part of our review we focus on density estimation. When discussing the issue of density estimation we begin by reviewing frequentist approaches to the problem, as commonly seen in economics, then illustrate those methods in an example. Once this has been completed, we repeat that same process - first reviewing methods and then focusing on their application - although this time we do so from a Bayesian perspective. We follow the same general pattern as we cover nonparametric estimation of regression functions. For both density and regression estimation, we pay particular attention to what are perceived as key implementation issues: the selection of the smoothing parameters and kernel functions in the frequentist case, and the treatment of smoothing parameters and the number of mixture components in the Bayesian paradigm.

DENSITY ESTIMATION

There are many reasons why economists seek to recover density estimates: as a summary tool for visualizing salient features of the data, as an input toward estimating and quantifying specific parameters of interest such as quantiles or tail probabilities (e.g., the probability of family income falling below the poverty line), or as a method for motivating other techniques, such as regression discontinuity approaches. Nonparametric density estimation techniques in particular have considerable appeal for economic applications, as researchers value methods that can adapt to the problem at hand, and can produce estimates of objects of interest that are not sensitive to specific (and potentially incorrect) parametric structures.

In this section, we review both classical and Bayesian methods for density estimation, and illustrate those methods in an economic problem by estimating hourly wage densities. We begin with a discussion of classical kernel-based approaches, apply those estimate densties of (log) hourly wages, and then move on to Bayesian techniques.

Classical Approach

We begin with the simple case of a continuous, univariate random variable X. Let $F(x)$ denote the cumulative distribution function of X. From the definition of the density, we know that

$$\begin{aligned} f(x) &= \frac{d}{dx}F(x) \\ &= \lim_{h\to 0}\frac{F(x+\frac{h}{2})-F(x-\frac{h}{2})}{h}, \end{aligned}$$

where h is the width of the interval. We plan to estimate $f(x)$ using a random sample of data $(x_1, x_2, \ldots, x_n)$. The simplest estimator would be to count the number of observations around the point x and divide that number by nh. The resulting estimator would be given as

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} 1\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right),$$

where $1(\cdot)$ takes the value 1 if the argument is true and 0 otherwise; this is the common histogram. We replace the indicator function with the more general notation of a kernel function $k(\cdot)$ and the estimator is now given as

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right).$$

Below we will discuss alternative choices for the kernel function as well as the choice of h which we refer to as the bandwidth. The use of the kernel allows the density to be smooth.

In practice it is likely that we will have mixed data - composed of both continuous and discrete variables. Define $\boldsymbol{x} = [\boldsymbol{x}^c, \boldsymbol{x}^D]$, where $\boldsymbol{x}^c$ contains the continuous variables and $\boldsymbol{x}^D = [\boldsymbol{x}^u, \boldsymbol{x}^o]$ contains the discrete data, further partitioned as unordered and ordered data. The total number of covariates can be decomposed as $q = q_c + q_D = q_c + (q_u + q_o)$. We will smooth the continuous data using bandwidth h and our discrete data with bandwidth $\lambda = [\lambda^u, \lambda^o]$.

To smooth mixed data, we deploy the generalized product kernel function[4], defined as

$$\begin{aligned} W_{ix} &\equiv K_h\left(\boldsymbol{x}_i^c, \boldsymbol{x}^c\right) L^u_{\lambda^u}\left(\boldsymbol{x}_i^u, \boldsymbol{x}^u\right) L^o_{\lambda^o}\left(\boldsymbol{x}_i^o, \boldsymbol{x}^o\right) \\ &= \prod_{d=1}^{q_c} k\left(\frac{x_{id}^c - x_d^c}{h_d}\right) \prod_{d=1}^{q_u} \ell^u(x_{id}^u, x_d^u, \lambda_d^u) \prod_{d=1}^{q_o} \ell^o(x_{id}^o, x_d^o, \lambda_d^o). \end{aligned}$$

This gives rise to the generalized product kernel density estimator

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{n\,|\mathbf{h}|} \sum_{i=1}^{n} W_{ix},$$

where $|\mathbf{h}|$ is the product of the bandwidths for only the continuous variables $(h_1 h_2 \cdots h_{q_c})$.

To implement the kernel density estimator, we need to select the kernels and the associated bandwidths. The MSE of the density estmator depends on the kernel functions used and the size of the bandwidths. MSE goes to zero as the sample size tends towards infinity and each bandwidth tends towards zero while at the same time the product of the continuous bandwidths and the sample size tend towards infinity[4]. In other words, as the sample size gets larger, each bandwidth shrinks to zero, but it shrinks slow enough so that $n|\mathbf{h}| \to \infty$. The intuition is that as the sample size gets larger, we do not need to smooth over individuals who are different from us as we will have a large number of observations which are identical (in terms of their $\boldsymbol{x}$ values) to us.

Kernel Choice

It is feasible to reduce the MSE of the estimator by appropriate choice of the kernel function.[7] was the first to study this issue and determined the optimal kernel which now bears his name. While the use of the Epanechnikov kernel results in the lowest MSE, this does not imply that it is the best kernel. The Epanechnikov kernel possesses only one continuous derivative. Economists typically employ the Gaussian kernel which has derivatives of all orders. Formally, the Gaussian kernel is given as

$$k\left(\frac{x_i - x}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - x}{h}\right)^2},$$

and we will employ this kernel in our empirical illustrations.

Many authors argue that kernel choice is less important than bandwidth choice. While we do not necessarily disagree, when the dimension of the data increases, this choice becomes more important. The efficiency of kernel functions relative to the Epanechnikov kernel worsen as the dimension increases. Further discussion can be found in Chapter 3 of[2].

For discrete variables, there are also several choices for kernel functions. The first and most popular unordered discrete kernel function is developed in[6], but it requires knowledge of the support of the data (not an issue with binary data). In our empirical illustrations we employ the unordered discrete kernel in[5]. Their kernel function is given as

$$l^u(x_i^u, x^u, \lambda) = \lambda^{1\left(x_i^u \neq x^u\right)}.$$

When $\lambda = 0$, we resort back to an indicator function. When $\lambda = 1$, the kernel function becomes a constant and we have the possibility of uniform smoothing. One issue with this kernel is that the kernel weights do not sum to one. This would imply that the kernel density estimator will not be a proper probability density, but this is easily remedied by normalizing the density estimator.

Bandwidth Selection

Perhaps the most important aspect of applied nonparametric estimation is selection of the bandwidths.[4] discuss data driven bandwidth selection in the mixed data case. The optimal smoothing parameters for the mixed data kernel density estimator can be obtained by minimizing the integrated squared difference between the estimated density and the true density as

$$CV^m(h,\lambda) = \min_{h_1,\ldots,h_{q_c}\lambda_1,\ldots,\lambda_{q_D}} \int \left[\widehat{f}(v) - f(v)\right]^2 dv.$$

Replacing population moment conditions with sample moments and using a leave-one-out estimator to avoid the bandwidth tending towards zero, it is possible to show the feasible cross-validation function

$$CV_1^m(h,\lambda) = \min_{h_1,\ldots,h_{q_c}\lambda_1,\ldots,\lambda_{q_D}} \left[\frac{1}{n^2\left|\mathbf{h}\right|^2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}^{(2)} - \frac{2}{n(n-1)\left|\mathbf{h}\right|^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} W_{ij}\right],$$

where $W_{ij}^{(2)} = K_{h,ij}^{(2)} L_{\lambda,ij}^{(2)}$ is the convolution kernel.

Empirical Illustration

Here we present an illustration of the methods discussed previously. We consider a relatively simple example, but one that still demonstrates how the methods are employed and what can be learned from their application.

We examine the distribution of hourly wages for college educated men and women. The data that we use come directly from the 2013 March Supplement of the Current Population Survey (CPS), compiled by the Bureau of Labor Statistics. Our cross-section consists of white, married (with spouse present) men and women, aged 18-64, who are engaged full-time in the labor market. In addition, we focus here only on those whose highest level of education is a bachelor's degree.

These specific restrictions serve two purposes. First, they produce a relatively homogeneous sample for which to compare wages between men and women. Second, after the restrictions are imposed, we obtain a reasonably large, but manageable, working data set, given the wealth of observations available in the CPS. Specifically, our sample has 8,112 observations of which 4564 are male. For now, we focus our attention only on differences across gender; after describing both Bayesian and frequentist approaches to nonparametric density estimation problems, we will also consider the role of age in explaining variation in conditional mean functions.

Figure 1 gives the densities of log hourly wages for each gender. We initially used cross-validation methods to obtain our bandwidths (both least-squares and likelihood cross-validation), but this led to bandwidths which were too small to distinguish any features of the data. Hence, we resorted to rule-of-thumb bandwidths. We can see that the mode of the male density is to the right of the female density. This result holds true for men and women who are otherwise relatively homogeneous. It is not possible to determine (simply with this figure) whether this difference is brought about by different levels of experience, discrimination and/or other factors. We consider a common proxy for experience in the next sub-section when we consider nonparametric approaches to regression problems.

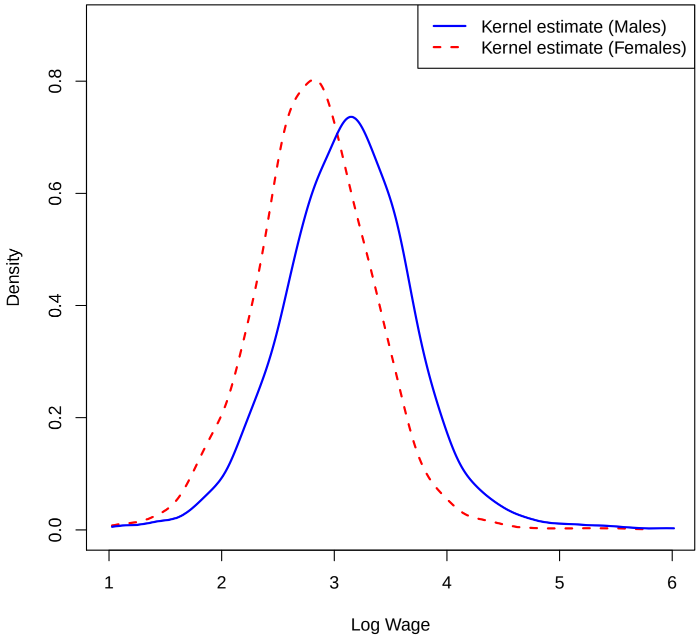

Figure 1: Kernel density estimates of log hourly wages by gender - using a Gaussian kernel and bandwidths equal to $1.06\sigma_x n^{-1/5}$

Bayesian Approach

We continue to consider the problem of density estimation, but now describe how the problem can be approached from a Bayesian point of view. Bayesians, of course, combine prior and data information to obtain posterior distributions of the model's parameters. Data information enters the process through specification of the likelihood function, as the researcher puts forward an assumed model for the data density.

Finite Mixture Models

For the case of density estimation considered here, we might choose to assume that the true density function - whatever it may be - can be adequately approximated by a finite mixture of Gaussian (normal) distributions (focusing on the univariate case for simplicity, although multivariate extensions are straightforward):

$$y_i|\mu, \sigma, \pi \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2), \quad i = 1, 2, \ldots, n. \tag{1}$$

In the above, we have represented the density of the scalar random variable y as a mixture of underlying Gaussian distributions, with $\mathcal{N}(\mu, \sigma^2)$ denoting a normal distribution with mean μ and variance σ^2. Note that this specification does not impose that the underlying true data generating process is normal; by mixing together several different Gaussian distributions, departures from normality are permitted. In practice the number of mixing components K is chosen to be reasonably large so that the model exhibits sufficient flexibility to capture skew, multimodality, fat tails, and other salient features of the data. For most density estimation exercises in economic applications, the approximation in (1) for small-to-moderate K is likely to be quite accurate.

The parameters π serve to weight the individual mixture components, with $\sum_{k=1}^{K} \pi_k = 1$. The number of components K, for now, is taken as given. Estimation can be conducted in a number of number of ways, including maximum likelihood, moments-based approaches and the expectation-maximization (EM) algorithm. Below we discuss another fully Bayesian alternative: a simulation-based estimation algorithm via Markov Chain Monte Carlo (MCMC) methods, namely the Gibbs sampler.

To this end, it is useful to introduce an equivalent representation of (1) which incorporates a latent variable vector $\mathbf{z}_i$. Specifically, let $\mathbf{z}_i = [z_{i1}\ z_{i2}\ \cdots\ z_{iK}]$ denote a component label vector. One and only one of the entries of this vector has a unit value (with the others all being zero), and a one in the j^{th} position denotes that y_i is drawn from the j^{th} component of the mixture. The specification in (1) can then be reproduced by writing:

$$y_i|\mu, \sigma, \mathbf{z} \sim \prod_{k=1}^{K} \left[\mathcal{N}(\mu_k, \sigma_k^2)\right]^{z_{ik}},$$

with a multinomial prior placed over the component label vector:

$$\mathbf{z}_i|\boldsymbol{\pi} \sim Mult(1, \boldsymbol{\pi}) \Rightarrow p(\mathbf{z}_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}},$$

with $\boldsymbol{\pi} = [\pi_1\ \pi_2\ \cdots\ \pi_K]$. Given this structure, a model equivalent to (1) is produced; when integrating the conditional (on $\mathbf{z}_i$) sampling distribution of the data over the multinomial prior for $\mathbf{z}_i$, the unconditional likelihood in (1) is obtained.

A Bayesian analysis of this model is completed upon specifying priors for the component specific parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma^2}$ and $\boldsymbol{\pi}$. Below we make the following choices:[1]

$$\begin{aligned} \mu_k &\overset{iid}{\sim} \mathcal{N}(\mu_0, V_\mu), \quad k = 1, 2, \ldots, K \\ \sigma_k^2 &\overset{iid}{\sim} IG(a, b), \quad k = 1, 2, \ldots, K \\ \boldsymbol{\pi} &\sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_K). \end{aligned}$$

All of the hyperparameters μ_0, V_μ, a, b and $\{\alpha_k\}_{k=1}^{K}$ are assumed fixed and selected by the researcher.

An MCMC-based strategy via the Gibbs sampler involves cycling through draws from the complete posterior conditionals of the model parameters. This involves four steps, one for each of the sets of parameters $\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\pi}, \mathbf{z}$. With a little patience (and a little algebra),

[1] Here, IG denotes an inverse (or inverted) gamma distribution and is parameterized as: $x \sim IG(a, b) \Rightarrow p(x) \propto x^{-(a+1)} \exp(-[bx]^{-1})$. In practice, component-specific hyperparmeters of the priors can be employed; here we focus on the case of common priors only for simplicity.

one can derive the following forms for the conditional posterior distributions:

$$\mu_k|\theta_{-\mu_g}, \text{Data} \overset{ind}{\sim} \mathcal{N}(D_{\mu_k} d_{\mu_k}, D_{\mu_k}), \quad k = 1, 2, \cdots K \tag{2}$$

where θ_{-x} denotes all quantities in our posterior other than x and

$$D_{\mu_k} = \left[n_k/\sigma_k^2 + V_\mu^{-1}\right]^{-1}, \quad d_{\mu_k} = \left[\sum_{i=1}^{n} (z_{ik} y_i)\right] / \sigma_k^2 + V_\mu^{-1} \mu_0,$$

where $n_k \equiv \sum_{i=1}^{n} z_{ik}$ denotes the number of observations "in" the g^{th} component of the mixture. The term $\sum_i z_{ik} y_i$ simply selects and sums the subset of y observations currently assigned to the k^{th} mixture component. As for the remaining posterior conditionals, we obtain:

$$\sigma_k^2|\theta_{-\sigma_k^2}, \text{Data} \overset{ind}{\sim} IG\left((n_k/2) + a, \left[b^{-1} + (1/2) \sum_i z_{ik}(y_i - \mu_k)^2\right]^{-1}\right) \quad k = 1, 2, \cdots K, \tag{3}$$

$$z_i|\theta_{-z_i}, \text{Data} \overset{ind}{\sim} \text{Mult}\left(1, \left[\frac{\pi_1 \phi(y_i; \mu_1, \sigma_1^2)}{\sum_{k=1}^{K} \pi_k \phi(y_i; \mu_k, \sigma_k^2)} \; \frac{\pi_2 \phi(y_i; \mu_2, \sigma_2^2)}{\sum_{k=1}^{K} \pi_k \phi(y_i; \mu_k, \sigma_k^2)} \cdots \frac{\pi_K \phi(y_i; \mu_K, \sigma_K^2)}{\sum_{k=1}^{K} \pi_k \phi(y_i; \mu_k, \sigma_k^2)}\right]\right) \tag{4}$$

and

$$\boldsymbol{\pi} \mid \theta_{-\boldsymbol{\pi}}, \text{Data} \sim \text{Dirichlet}(n_1 + \alpha_1, n_2 + \alpha_2, \cdots n_G + \alpha_G). \tag{5}$$

A Gibbs algorithm to this problem involves cycling through the distributions in (2) - (5). An initial set of simulations, or a "burn-in" period is discarded, and the final set of simulations are retained for estimation purposes. An estimate of the mixture density can be calculated as follows:

$$\widehat{p(y)} = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \pi_k^{(m)} \phi(y; \mu_k^{(m)}, \sigma_k^{2(m)}),$$

with $\theta^{(m)}$ denoting the m^{th} post-convergence simulation of the parameter θ, M denoting the total number of posterior simulations, and $\phi(x; \mu, \sigma^2)$ denoting a normal density function for the random variable x with mean μ and variance σ^2.

Density Estimates via Dirichlet Process Priors

A limitation of the preceding approach lies in the determination of the number of mixture components K. If K is selected to be too small, then the model may not be rich enough to capture key features of the data. If, on the other hand, K is chosen to be too large, some of the mixture components may be redundant or, as the Gibbs algorithm is run, some mixture components may be assigned few or no observations, resulting in overfitting and a loss of efficiency.

An alternate approach that seeks to surmount these deficiencies is to, instead, allow K to be endogenized within the model. One possible avenue here is to employ reversible jump MCMC methods[15] which allows a sampler to navigate across models of varying dimensions. More recently, approaches within economics have instead employed the Dirichlet process prior, essentially allowing a fully nonparametric approach to the density estimation problem. We describe this approach below.

The specific model we employ is termed a Dirichlet process mixture model (DPMM) and is specified as follows:

$$\begin{aligned} y_i|\mu_i, \sigma_i^2 &\overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, 2, \ldots, n && (6) \\ \boldsymbol{\theta}_i &\equiv [\mu_i \ \sigma_i^2] \mid G \overset{iid}{\sim} G && (7) \\ G &\sim DP(G_0, \alpha). && (8) \end{aligned}$$

In the above, the parameters $\boldsymbol{\theta}_i$ are assumed to be generated from an unknown distribution G, and a prior over that distribution - the Dirichlet Process prior - is employed in (8). One can think about G_0 as the center of this prior, or the "base measure" in the sense that for any measurable set A, we have $E(G[A]) = G_0(A)$. The "concentration parameter" α controls how tightly G is distributed over this mean distribution G_0, as suggested by the result $\text{Var}(G[A]) = G_0(A)[1 - G_0(A)]/(\alpha + 1)$. This we can think about this specification as one that permits a general distribution over the coefficients θ_i, and employs a prior over that distributional space with G_0 denoting the center of that prior, and α controlling how tightly the prior is specified around G_0.

As shown in Sethuraman[19], we can represent the DPMM as an infinite mixture of Gaussian distributions, with a "stick-breaking" process for the generation of the component weights. Specifically, we can write:

$$\begin{aligned} y_i|\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma} &\sim \sum_{k=1}^{\infty} \omega_k \mathcal{N}(\mu_k, \sigma_k^2) \\ \omega_k &= v_k \prod_{l<k}(1 - v_l) \\ v_l &\overset{iid}{\sim} Beta(1, \alpha), \quad l = 1, 2, \ldots \end{aligned}$$

In this form, we can see that the DP model affords an infinite mixture of normals representation for the sampling distribution, and offers a prescription for how the component-specific weights are generated. The advantage of this model over the previous finite mixture representation is that the algorithm allows us to "test down" and determine the number of components endogenously rather than fixing the number of components *a priori*.

There are a variety of algorithms that exist for the estimation of these models - algorithms based on the Pólya-Urn scheme,[11] the so-called Chinese restaurant process, and others that employ auxiliary variables and slice sampling[16,17]. Approaches to sampling based on a truncated representation of the infinite summation have also been described,[12] and articles that review alternate computational approaches also exist and are quite useful for practitioners.[18]

In what follows, we apply both the finite mixture and DPMM methods to estimate the log hourly wage distribution for men and women, as previously done using kernel methods in Figure 1. Our results are provided in Figure 2.

The figure plots two sets of results: first, results from the finite mixture model are presented, setting $K = 5$. For this model, we set $\mu_0 = 0$, $V_\mu = 100$, $\alpha_k = 2\ \forall k$ and choose a and b of the inverse gamma priors so that the prior mean and prior variance of σ_k^2 are both .5. The sampler is run separately on the male and female data subsamples, and an estimate of the log wage density for each gender is plotted in the figure, using the final 5,000 of 6,000 Gibbs simulations to perform the calculations.

For comparison purposes, we also plot density estimates from the DP model alongside

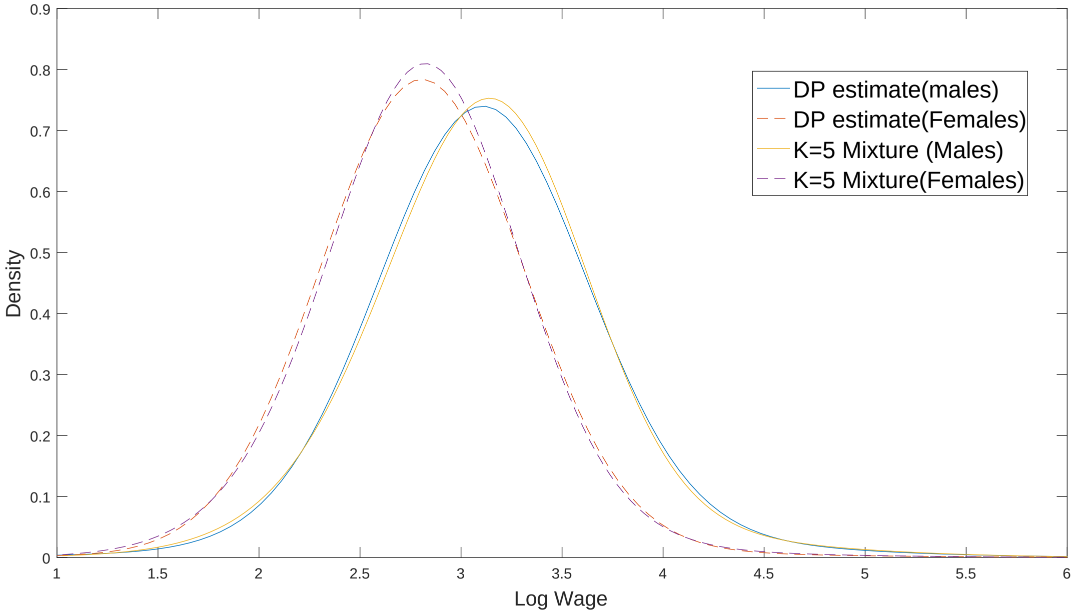

Figure 2: Density estimates of log hourly wages by gender - finite mixture and DPMM results

the finite mixture plots, and those are found to be quite similar to the 5-component mixture model results.[2] Looking more deeply at our posterior simulations, the DP model suggests that 5 components may be more than is needed to model this data, as for the females sample, $\Pr(K = 2|\text{Data}) = .65$, $\Pr(K = 3|\text{Data}) = .27$ and $\Pr(K \geq 4|\text{Data}) = .08$. A similar pattern is found for the males sample. Thus, the model clearly supports a movement away from the standard one-component Gaussian model, but also suggests that the full flexibility afforded by the $K = 5$ case may be unnecessary. Furthermore, results obtained here are quite similar to those obtained using kernel methods in the previous section.

REGRESSION ESTIMATION

While density estimation is a useful tool, regression is the backbone of applied econometric research. The vast majority of economic research still assumes, without any theoretical justification, that regressors enter the conditional mean linearly and that each regressor is separable. Here we discuss how to estimate regression functions where we are unsure of the

[2] For the DP analysis, we make use of Matlab code provided by Song[20].

underlying functional form.

Classical Approach

We consider a nonparametric regression function where we allow for some of the regressors to be discrete in nature. Our nonparametric regression model, as given in[22], is

$$y_i = m\left(\boldsymbol{x}_i\right) + u_i, \quad i = 1, 2, \ldots, n, \tag{9}$$

where $m(\cdot)$ is the unknown smooth conditional mean with regression vector $\boldsymbol{x}_i$ defined earlier, and u_i is a mean zero additive error term which we assume is uncorrelated with $\boldsymbol{x}_i$.

Using the mixed data generalized product kernel, regression estimators can be obtained by minimizing the kernel weighted sum of squared errors

$$\sum_{i=1}^{n} u_i^2 W_{ix} = \sum_{i=1}^{n} \left[y_i - m\left(\boldsymbol{x}_i\right)\right]^2 W_{ix}.$$

The so-called local-constant least-squares (LCLS) estimator is the solution to this objective function:

$$\widehat{m}\left(\boldsymbol{x}\right) = \left(\sum_{i=1}^{n} y_i W_{ix}\right) / \left(\sum_{i=1}^{n} W_{ix}\right). \tag{10}$$

The intuition behind this estimator follows from a simple example. If we were estimating the expected log hourly wage for an individual, we would place more weight on male observations if the point $\boldsymbol{x}$ were for a male than we would for female observations. Similarly, we would place more weight on individuals with higher levels of education if the point $\boldsymbol{x}$ were for an individual with a college degree than we would for observations who dropped out of high school (noting that we only need a single categorical variable for level of education and not multiple dummies as in a parametric model).

The asymptotic properties of the LCLS estimator in the presence of mixed data can be found in[22]. As is the case for density estimation with mixed data, we require the conditions that each bandwidth $h \to 0$ and $\lambda \to 0$ as $n \to \infty$ and that $nh_1 h_2 \cdots h_{q_c} \to \infty$. This is almost a free lunch as additional discrete regressors do not slow down the rate of convergence and hence do not add to the curse of dimensionality (one cost is that we must calculate additional bandwidths).

Estimating the regression model in (9) using a constant ($m(\cdot)$) is not the only way to locally approximate the unknown regression surface. As an alternative, a local-polynomial approximation can be obtained for a given point $\boldsymbol{x}$. The most popular version, the local-linear estimator, is obtained by taking a first-order Taylor expansion of $m(\cdot)$ to assist with construction of the estimator.

The choice of how many expansions to take is important. More expansions will lead to a reduction in the bias, but at a cost of an increase in variability. This is caused by the increase in the number of local parameters which must be estimated.[23] have an in-depth discussion of this issue, but we will limit ours to the following insight. It is often argued that if we are interested in the pth gradient, then we should use the $(p+1)$th-order expansion. For example, if we are interested in the conditional mean, the local-linear estimator is preferable.

Bandwidth Selection

The goal here is to produce the set of bandwidths which minimize the cross-validation function

$$CV\left(h, \lambda^u, \lambda^o\right) = \sum_{i=1}^{n} \left[y_i - \widehat{m}_{-i}\left(\boldsymbol{x}_i\right)\right]^2,$$

where $\widehat{m}_{-i}\left(\boldsymbol{x}_i\right)$ is the leave-one-out estimator $m(\cdot)$.

Note that the typical approach looks at minimizing the cross-validation function with respect to the conditional mean. It turns out that gradient estimates obtained from $\widehat{m}\left(\boldsymbol{x}\right)$, using a bandwidth determined through least-squares cross-validation is (asymptotically) too small for estimating $\partial\widehat{m}(\boldsymbol{x})/\partial\boldsymbol{x}$ and a rate adjustment is necessary. As an alternative,[24] develop a cross-validation function where minimization is based on the gradient of the unknown function.

Upper and Lower Bounds for Bandwidths

Historically, large-sample theory assumes that the bandwidths gravitate towards zero at a rate slow enough so that it does not dominate the fact that the sample size is growing toward infinity. What this implies (in large samples) is that we should see bandwidths that are close to zero. In a finite sample, it is impossible to know how 'close' to zero we are. In the continuous case, we can get a good sense of a large bandwidth by comparing it to the standard deviation of the regressor. If the bandwidth of a particular variable is say, three times its standard deviation, then we can be relatively confident that this is a large bandwidth.

The intuition is that for a really large bandwidth, the term within the kernel is small and so we can treat it as 0. Thus, the term does not depend on the observation (i) and hence it cancels from both the numerator and the denominator. In the LCLS case, this deems the variable irrelevant in terms of smoothing the function. For the local-linear estimator, we see that when the bandwidth for a continuous regressor gets large, the estimator treats the variable as if it enters linearly.

Empirical Results

Here our goal is to study the age-earnings profiles of college-educated married white men and women. We seek to uncover these relationships by applying estimators that make few assumptions regarding the shapes of these profiles, and to use these methods to describe differences in patterns across men and women. For the frequentist case, results are found to be relatively similar across estimation procedures. As a result we only show estimates for the local-linear least-squares estimator, with bandwidths selected via least-squares cross-validation.

The conditional mean estimates obtained via regressing log hourly earnings on age and gender are given in the left panel of Figure 3. We are able to plot these in two dimensions given that gender is binary. Each curve is consistent with past results in the theoretical and applied literatures. Log hourly wages increase quickly at younger ages, then begin to plateau and eventually fall. For men, the decline begins at roughly at 52 years of age, while the expected earnings decline of females appears to occur earlier (around 45 years of age).

While it is interesting to see that the figure is consistent with previous findings in the literature, the more compelling result is the difference between the two curves. Both have the same general shape, yet expected log hourly wages of women are always below those of

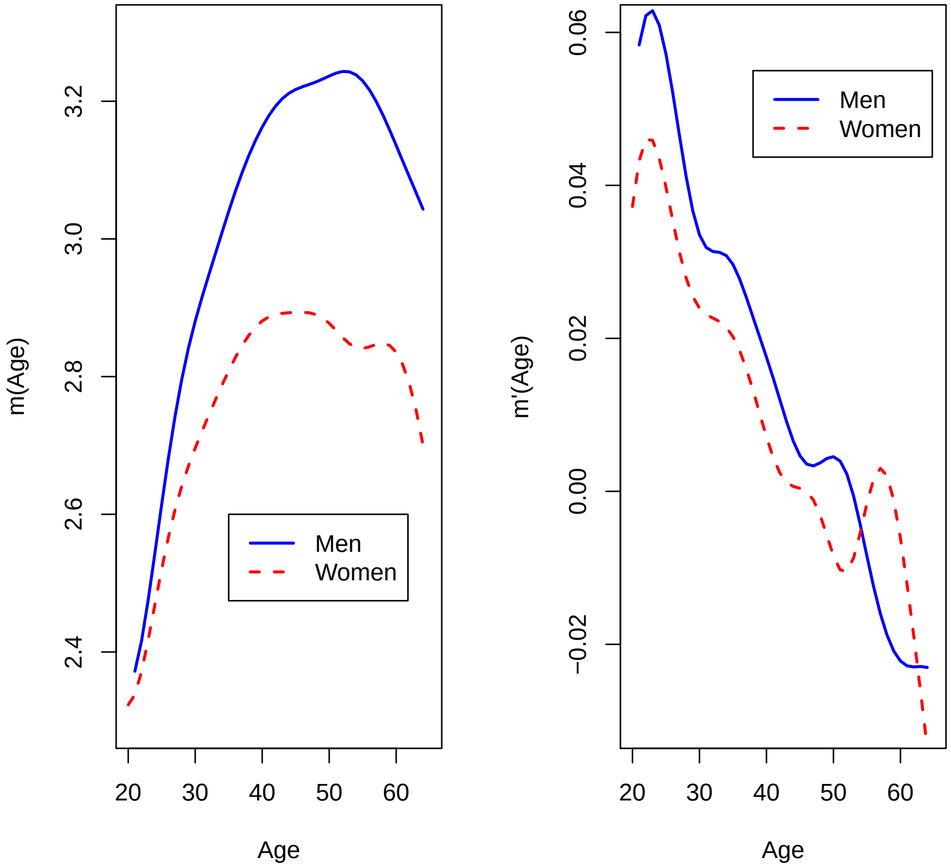

Figure 3: Kernel Estimated Conditional Mean Function Relating Age to Log Hourly Earnings (Left Panel); Kernel Estimated Marginal Effect (Gradient) of Age on Earnings (Right Panel)

males (albeit very close initially), and this difference increases with age. Many explanations have been given for this wage disparity (e.g., discrimination, lower levels of experience given child rearing, etc.) and it is likely that many of these explanations can help to explain the gender gap.

We plot the gradient of the conditional mean for each regressor versus age in the right panel of Figure 3. We see that the slope decreases with age and eventually becomes negative (around 45 for females and 52 for males). It is interesting to note that the rate of decay is actually quite similar between the two groups. This gives some credibility for the experience argument put forth in the literature (e.g.,[25]).

Bayesian Approach

As in the previous section, we consider a standard nonparametric regression problem, yet add to it an assumption of normally distributed disturbances. We consider a univariate case for simplicity, although generalizations exists for higher-dimension problems. Specifically, and with an eye toward estimating age-earnings profiles as considered previously, we review Bayesian techniques for estimating the following model:

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad \text{with} \quad \boldsymbol{\epsilon}|\mathbf{X} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n).$$

Following an approach described in the literature,[13,14] our method addresses this problem by treating each point on the regression curve as a parameter to be estimated, by employing a prior that shrinks neighboring parameters together, and by using well-known and computationally convenient results for Bayesian linear regression with conditionally conjugate priors.

To this end, suppose that there are $K \leq n$ distinct x_i values in the sample and denote these as $\{x_k^*\}_{k=1}^K$ with $x_1^* < x_2^* < \cdots < x_K^*$. Furthermore, let $\mathbf{D}$ denote an $n \times K$ assignment matrix, where the i^{th} row of $\mathbf{D}$ simply maps that observation's x_i value to the corresponding element in $\mathbf{x}^*$. Specifically, the k^{th} element of the i^{th} row of $\mathbf{D}$, or D_{ik}, is calculated as $D_{ik} = I(x_i = x_k^*)$, with $I(\cdot)$ denoting the standard indicator function. Thus, each row of $\mathbf{D}$ contains one an only one unit value, and all other row entries are zero.

With this in hand, we can then write our model in traditional linear regression fashion as

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

Note that the elements of the $K \times 1$ vector $\boldsymbol{\theta}$ denote (sorted) elements of the regression curve, i.e., $\boldsymbol{\theta} = [m(x_1^*) \; m(x_2^*) \; \cdots \; m(x_K^*)]'$. To incorporate the idea that the regression function should be "smooth," we employ a prior that expresses the idea that adjacent values of $\boldsymbol{\theta}$ should be "similar."

To this end, we first define $\Delta_k = x_k^* - x_{k-1}^*$, $k = 2, 3, \cdots, K$, and construct a $K \times K$

matrix $\mathbf{H}$ as follows:

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

We observe that $\mathbf{H}\boldsymbol{\theta}$ serves to transform the coefficient vector $\boldsymbol{\theta}$ into a vector of "initial conditions" (i.e., the first two points on the regression curve) and first-differences in point-wise slopes of the curve. That is,

$$\boldsymbol{\gamma} \equiv \mathbf{H}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \frac{\theta_3-\theta_2}{\Delta_3} - \frac{\theta_2-\theta_1}{\Delta_2} \\ \frac{\theta_4-\theta_3}{\Delta_4} - \frac{\theta_3-\theta_2}{\Delta_3} \\ \vdots \\ \frac{\theta_K-\theta_{K-1}}{\Delta_K} - \frac{\theta_{K-1}-\theta_{K-2}}{\Delta_{K-1}} \end{bmatrix} \approx \begin{bmatrix} \theta_1 \\ \theta_2 \\ m'(x_2) - m'(x_1) \\ m'(x_3) - m'(x_2) \\ \vdots \\ m'(x_{K-1}) - m'(x_{K-2}) \end{bmatrix}.$$

Beliefs about smoothness of the regression function can be incorporated through a prior distribution over the elements of $\boldsymbol{\gamma}$. If the prior for the last $K-2$ elements of $\boldsymbol{\gamma}$ is very tightly centered over zero, for example, we would effectively restrict the differences between consecutive point-wise slopes to be zero, thereby imposing global linearity of the regression curve; values of the initial conditions θ_1 and θ_2 would then determine the intercept and global slope of that curve. On the other hand, "loose" priors on the elements of $\boldsymbol{\gamma}$ may lead to an under-smoothed curve that essentially connects the data points.

The relative ease and intuitive appeal with which smoothness considerations can be imposed on $\boldsymbol{\gamma}$ suggests reparameterizing the model in terms of $\boldsymbol{\gamma}$ and backing out information on the regression curve itself through $\boldsymbol{\theta} = \mathbf{H}^{-1}\boldsymbol{\gamma}$. Specifically, we can write

$$\begin{aligned} \mathbf{y} &= \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ &= \mathbf{D}\mathbf{H}^{-1}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \end{aligned}$$

where $\mathbf{X} \equiv \mathbf{D}\mathbf{H}^{-1}$. The model above, combined with a set of priors of the form

$$\begin{aligned} \boldsymbol{\gamma}|\eta &\sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \mathbf{0}_{K-2} \end{bmatrix}, \begin{bmatrix} \mathbf{V}_{\theta_1,\theta_2} & \mathbf{0}_{2,K-2} \\ \mathbf{0}_{K-2,2} & \eta \boldsymbol{I}_{K-2} \end{bmatrix}\right) \equiv \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \mathbf{V}_{\boldsymbol{\gamma}}(\eta)\right) \\ \eta &\sim IG(a_\eta, b_\eta) \\ \sigma^2 &\sim IG(a_{\sigma^2}, b_{\sigma^2}) \end{aligned}$$

completes the specification. Note the prior specification above allows for the possibility of a fairly diffuse stance surrounding the initial conditions θ_1 and θ_2 though a suitable choice of $\mathbf{V}_{\theta_1,\theta_2}$. In addition, the parameter η acts as a smoothing parameter, quite similar in spirit to the bandwidth parameter discussed previously for classical nonparametric regression. We add an inverse gamma prior over η and note that the degree of smoothing will be automatically updated by the data. Despite this learning, however, choice of smoothing parameter matters (like the classical case), and for the Bayesian this can manifest itself in sensitivity of the posterior results to the choice of hyperparameters a_η and b_η.

Fitting the model via the Gibbs sampler is a straightforward exercise. Specifically, the following complete posterior conditional distributions are obtained:

$$\boldsymbol{\gamma}|\eta, \sigma^2, \text{Data} \sim \mathcal{N}\left(\mathbf{D}_{\boldsymbol{\gamma}}\mathbf{d}_{\boldsymbol{\gamma}}, \mathbf{D}_{\boldsymbol{\gamma}}\right), \tag{11}$$

where

$$\mathbf{D}_{\boldsymbol{\gamma}} = \left(\mathbf{X}'\mathbf{X}/\sigma^2 + \mathbf{V}_{\boldsymbol{\gamma}}^{-1}(\eta)\right)^{-1}, \quad \mathbf{d}_{\boldsymbol{\gamma}} = \mathbf{X}'\mathbf{y}/\sigma^2 + \mathbf{V}_{\boldsymbol{\gamma}}^{-1}(\eta)\boldsymbol{\mu}_{\boldsymbol{\gamma}},$$

$$\sigma^2|\boldsymbol{\gamma}, \eta, \text{Data} \sim IG\left(\frac{n}{2} + a_{\sigma^2}, \left[b_{\sigma^2}^{-1} + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\gamma})'(\mathbf{y} - \mathbf{X}\boldsymbol{\gamma})\right]^{-1}\right) \tag{12}$$

and finally

$$\eta|\sigma^2, \boldsymbol{\gamma}, \text{Data} \sim IG\left(\frac{K-2}{2} + a_\eta, \left[b_\eta^{-1} + \frac{1}{2}\boldsymbol{\gamma}_{3:K}'\boldsymbol{\gamma}_{3:K}\right]^{-1}\right), \tag{13}$$

where $\boldsymbol{\gamma}_{3:K}'$ denotes the last $K-2$ elements of $\boldsymbol{\gamma}$, which form first-differences in point-wise slopes. A Gibbs algorithm involves successively sampling from these conditionals; at each iteration, estimates of the regression function can be obtained by calculating $\boldsymbol{\theta} = \mathbf{H}^{-1}\boldsymbol{\gamma}$.

Empirical Results

We make use of the previous results to fit our (log) wage-age model, as described in the previous section. Results are presented in Figure 4. The leftmost graph plots posterior means of the regression function $m(age)$ for both men and women. The graph on the right plots the first derivative of this regression function, $m'(age)$. These quantities (and their posterior standard deviations, although these are not reported in the figures for the sake of clarity) are easily calculated given posterior simulations of $\boldsymbol{\gamma}$. The sampler is run for 10,000 iterations, and the first 1,000 of these are discarded as the burn-in period. For our priors, we choose a_η and b_η so that the prior mean and prior standard deviation of η are both equal to 10^{-4} and choose the prior hyperparameters a_{σ^2} and b_{σ^2} so that the prior mean and standard deviation of σ^2 are both equal to 1. Finally, we choose a diffuse prior over the initial conditions by setting $\boldsymbol{\mu}_1 = 0$ and $\mathbf{V}_{\theta_1,\theta_2} = 100\mathbf{I}_2$.

Our Bayesian posterior estimates are very close to the point estimates reported in Figure 3. We see that the conditional mean function for men rises with age until a peak in the early 50's and then decreases. For women we see an increase in expected earnings until approximately 40 years of age, and then the function flattens, with an overall pattern on decline until 65. Plots of the marginal effects in the right panel echo this pattern; the derivative is found to equal zero at approximately 40 years of age for women and 50 for men. Furthermore, the fact that the marginal effect point estimate for men lies above that same effect for women (prior to age 50) suggests that, on average, men's wages tend to grow at a faster rate than women's wages year-over-year.

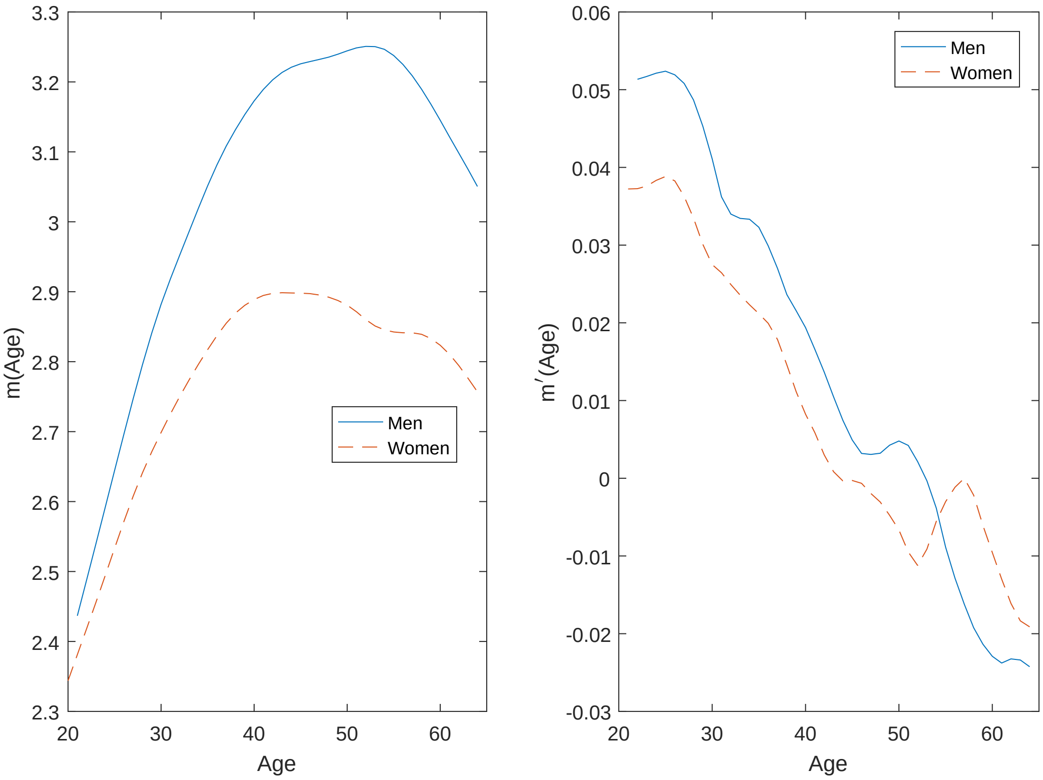

Figure 4: Posterior Mean of the Regression Function Relating Age to Log Hourly Earnings (Left Panel); Posterior Mean of the Marginal Effect (Gradient) of Age on Earnings (Right Panel)

CONCLUSION

We have provided a general overview of modern nonparametric methods that are commonly used in economic applications. We have reviewed both Bayesian and frequentist approaches and illustrated how both can be used in problems of hourly wage density estimation and flexible estimation of age-earnings profiles.

The allure of nonparametric methods, regardless of the perspective in which they are applied, is the ability to reduce potentially stringent parametric assumptions on the problem at hand. Within economics a precise functional form is rarely provided by theory, and so applied researchers often operate with little guidance regarding the specifications that should be taken to the data. Nonparametric methods can offer significant advantages here, as they require less in terms of inputs and assumptions made by the practitioner.

References

1. Rossi, P.E. Bayesian Non- and Semi-parametric Methods and Applications. **2014**, Princeton University Press, Princeton.
2. Henderson, D.J., Parmeter, C.F. Applied Nonparametric Econometrics. **2015**, Cambridge University Press, New York.
3. DiNardo, J., Tobias, J.L. Nonparametric Density and Regression Estimation. *Journal of Economic Perspectives* **2001**, 15(4), 11-28.
4. Li, Q., Racine, J. Nonparametric Estimation of Distributions with Categorical and Continuous Data. *Journal of Multivariate Analysis* **2003**, 86, 266-292.
5. Ouyang, D., Li, Q., Racine, J. Cross-validation and the Estimation of Probability Distributions with Categorical Data. *Journal of Nonparametric Statistics* **2006**, 18(1), 69-100.
6. Aitchison, J., Aitken, C.G.G. Multivariate Binary Discrimination by the Kernel Method. *Biometrika* **1976**, 63(3), 413-420.
7. Epanechnikov, V.A. Nonparametric Estimation of a Multidimensional Probability Density. *Teoriya Veroyatnostei i ee Primeneniya* **1969**, 14, 156-161.
8. Blackwell, D., MacQueen, J.B., Ferguson distributions via Polya Urn Schemes. *Annals of Statistics*, **1973**, 1(2), 353-355.
9. Ferguson, T.S., A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1973**, 1(2), 209-230.
10. Escobar, M.D., Estimating normal means with a Dirichlet Process Prior. *Journal of the American Statistical Association*, **1994**, 89, 268-277.
11. Escobar, M.D. West, M., Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **1995**, 90(43), 577-588.

12. Ishwaran, H., James L.F. Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. *Journal of Computational and Graphical Statistics* **2002**, 11(3), 508-532.

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

14. Koop, G., Poirier, D.J., Tobias, J.L. Semiparametric Bayesian inference in multiple equations models. *Journal of Applied Econometrics* **2004**, 19(7), 827-849.

15. Green, P.J., Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **1995**, 82(4), 711-732.

16. Kalli, M., J.E. Griffin, S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, **2011**, 21, 93-105.

17. Walker, S.G. Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics - Simulation and Computation*, **2007**, 36, 45-54.

18. Neal, R.M., Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **2000**, 9(2), 249-265.

19. Seutheraman, J., A constructive definition of Dirichlet priors. *Statistica Sinica* **1994**, 4, 639-650.

20. Song, Y., DPM in Applications. Notes provided for a short course at the Melbourne Bayesian Econometrics Workshop, November 2016.

21. Silverman, B.W., Density Estimation for Statistics and Data Analysis. **1986**, Chapman and Hall, London.

22. Racine, J.S., Q. Li. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **2004**, 119, 99-130.

23. Fan, J., I. Gijbels. Local polynomial modelling and its applications. **1996**, Chapman and Hall, London.

24. Henderson, D.J., Q. Li, C.F. Parmeter, S. Yao. Gradient-based smoothing parameter selection for nonparametric regression estimation. *Journal of Econometrics* **2016**, 184, 233-241.

25. Polachek, S.W. Differences in expected post-school investment and determinant of market wage differentials. *International Economic Review* **1975**, 16, 451-470.