THE SMOOTH COLONEL AND THE REVEREND FIND COMMON GROUND

NICHOLAS M. KIEFER AND JEFFREY S. RACINE

Abstract. A semiparametric regression estimator that exploits categorical (i.e. discrete-support) kernel functions is developed for a broad class of hierarchical models including the pooled regression estimator, the fixed-effects estimator familiar from panel data, and the varying coefficient estimator, among others. Separate shrinking is allowed for each coefficient. Regressors may be continuous or discrete. The estimator is motivated as an intuitive and appealing generalization of existing methods. It is then supported by demonstrating that it can be realized as a posterior mean in the Lindley & Smith (1972) framework. As a demonstration of the flexibility of the proposed approach, the model is extended to non-parametric hierarchical regression based on B-splines.

1. Introduction

Kernel smoothing of coefficients across groups of related regressions provides an attractive method of combining common information without forcing a choice between constrained and unconstrained regressions. Choosing the extent of smoothing is subjective, but cross validation (Stone 1974) provides a practical and appealing method for choosing smoothing parameters in a wide range of settings. Though these methods appear to perform well in examples and simulations, they lack a firm statistical foundation. In fact, kernel smoothed estimators are closely related to posterior means in a normal hierarchical model. We explore that relationship, in the process providing a sound foundation and interpretation for the kernel smoother. Further, we extend the class of models with a new smoother suggested by the Bayesian formulation. The new class includes constrained regressions (the pooled model), unconstrained regressions, the fixed-effect model familiar from panel data analysis

Date: March 8, 2016.
We thank Bent Jesper Christensen and participants at the AU Info-Metrics conference in Riverside CA for their thoughtful comments and suggestions. Kiefer acknowledges support from the Alfred P Sloan Foundation. Racine acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).
Nicholas M. Kiefer: Departments of Economics and Statistical Science, 490 Uris Hall, Cornell University, Ithaca, NY 14853, nicholas.kiefer@cornell.edu; CREATES, funded by the Danish Science Foundation, University of Aarhus, Denmark. Jeffrey S. Racine: Department of Economics, Kenneth Taylor Hall, McMaster University, Hamilton, ON Canada L8S 4M4, racinej@mcmaster.ca.

(cf Breusch et al. (1989)), some classical combined estimators (cf Theil & Goldberger (1961), Judge & Bock (1976)) and a model with common intercepts but potentially different slopes. These are in a sense endpoints of the class of models. All in-between models are obtained by choice of the bandwidth.

We fix ideas by considering briefly the simple regression model, where the calculations are instructive. This allows stressing the essential ideas without unduly complicating notation. In this setting we compare the kernel and Bayes estimators. We then turn to the general hierarchical model. We extend the kernel estimator to this case. The general setting suggests a generalized kernel estimator allowing differential smoothing across coefficients. The development is ad hoc, as it results from modifying the equations defining the kernel estimator. Turning to the Bayesian formulation, we find a sound foundation for the new estimator. To demonstrate the flexibility inherent to our approach, we extend the method to a formulation based on B-splines which delivers a convenient and flexible nonparametric estimator. This simple extension substantially increases the range of application for these methods. The methods are illustrated with an application to a wage regression which demonstrates advantages over common parametric models based on the same functional form, and also highlights potential benefits from pursuing the more flexible nonparametric B-spline extension.

Though there exists a literature on Bayesian nonparametric regression that predictably involves mixtures of densities and Dirichlet priors (see Griffin & Steel (2010) and Karabatsos & Walker (2012) by way of illustration), our aim in the current context is to provide a firm statistical foundation for frequentist kernel estimators and suggest new estimators having the same solid foundations by demonstrating that they can be realized as a posterior mean in the Lindley & Smith (1972) framework.

2. The Simple Linear Model

2.1. **Ordinary and Kernel Estimators.** We begin by considering a single-regressor parametric hierarchical model[1] of the form

$$y_{ji} = x_{ji}\beta_i + \epsilon_{ji}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, c,$$

where n_i is the number of observations drawn from group i, and where there exist c groups. For the ith group we write this as

$$y_i = x_i\beta_i + \epsilon_i, \quad i = 1, \ldots, c,$$

[1]We ignore the intercept for notational simplicity (perhaps the data are centered) but return to that case in Section 3. Here β_i is a scalar).

where x_i is the vector $x_i = (x_{1i}, x_{2i}, \ldots, x_{n_i i})'$, $y_i = (y_{1i}, y_{2i}, \ldots, y_{n_i i})'$ and $\epsilon_i = (\epsilon_{1i}, \ldots, \epsilon_{n_i i})'$. Assume $E(\epsilon_i | x_i) = E(\epsilon_i) = 0$ and $E(\epsilon_i \epsilon_j') = 0$ so we abstract from consideration of endogeneity and error covariance. For the full sample we write this using matrix notation as

$$\mathbf{y} = A\beta + \epsilon,$$

where $\mathbf{y}$ is the n-vector of observations ($n = \sum_{i=1}^{c} n_i$), A is the $(n \times c)$ design matrix, and $\beta = (\beta_1, \ldots, \beta_c)'$, the vector of group derivatives.

Let $\mathbf{1}(l = i)$ be the indicator function taking value 1 when $l = i$ and 0 otherwise. The frequency estimator of β_i, which we denote $\hat{\beta}_i$, is the solution to

$$\begin{aligned}
\hat{\beta}_i &= \arg\min_{\beta_i} \sum_{l=1}^{c} \sum_{j=1}^{n_l} (y_{jl} - x_{jl}\beta_i)^2 \mathbf{1}(l = i) \\
&= \sum_{l=1}^{c} \sum_{j=1}^{n_l} y_{jl} \frac{x_{jl}}{\sum_{l=1}^{c} \sum_{j=1}^{n_l} x_{jl}^2 \mathbf{1}(l = i)} \mathbf{1}(l = i) \\
&= \frac{\sum_{l=1}^{c} x_l' y_l \mathbf{1}(l = i)}{\sum_{l=1}^{c} x_l' x_l \mathbf{1}(l = i)} = \frac{x_i' y_i}{x_i' x_i}
\end{aligned}$$

We express the estimator in this form to facilitate comparison with the kernel estimator. The semiparametric kernel estimator[2] $\hat{\beta}_{i,\lambda}$ of β_i is the solution to

$$\hat{\beta}_{i,\lambda} = \arg\min_{\beta_i} \sum_{l=1}^{c} \sum_{j=1}^{n_l} (y_{jl} - x_{jl}\beta_i)^2 L(l, i, \lambda),$$

where we use the kernel function

$$L(l, i, \lambda) = \begin{cases} 1, & \text{when } l = i, \\ \lambda, & \text{otherwise,} \end{cases}$$

where the "bandwidth" $\lambda \in [0, 1]$ The case $\lambda = 0$ leads to an indicator function, and $\lambda = 1$ gives a uniform weight function. We can also express this kernel as $L(l, i, \lambda) = \lambda^{\mathbf{1}(l \neq i)}$, where $\mathbf{1}(cond)$ is the usual indicator function taking on value 1 when $(cond)$ is true, 0 otherwise. Looking ahead to generalization it is useful to consider the FOC:

$$\sum_{l=1}^{c} \sum_{j=1}^{n_l} L(l, i, \lambda) x_{jl} (y_{jl} - x_{jl}\beta_i) = 0.$$

[2]The approach is semiparametric since it uses kernel smoothing for categorical (discrete) covariates while the relationship between y and x is parametrically specified. See Li et al. (2013) for detailed analysis of this class of kernel estimators and a demonstration of the asymptotic optimality of cross-validation for selecting smoothing parameters.

The kernel estimator of β_i is given by

$$\hat{\beta}_{i,\lambda} = \sum_{l=1}^{c}\sum_{j=1}^{n_l} y_{jl}\frac{x_{jl}}{\sum_{l=1}^{c}\sum_{j=1}^{n_l} x_{jl}^2 L(l,i,\lambda)}L(l,i,\lambda) = \frac{\sum_{l=1}^{c} x_l' y_l L(l,i,\lambda)}{\sum_{l=1}^{c} x_l' x_l L(l,i,\lambda)}.$$

We rewrite this for comparison with the Bayes estimator as

$$\hat{\beta}_{i,\lambda} = \frac{x_i' y_i + \lambda \sum_{l\neq i}^{c} x_l' y_l}{x_i' x_i + \lambda \sum_{l\neq i}^{c} x_l' x_l},$$

and define the pooled (overall) OLS estimator $\hat{\beta} = \sum_{l=1}^{c} x_l' y_l / \sum_{l=1}^{c} x_l' x_l$. Note that

$$\sum_{l\neq i}^{c} x_l' y_l = \sum_{l=1}^{c} x_l' x_l \hat{\beta} - x_i' x_i \hat{\beta}_i$$

since $x_i' y_i = x_i' x_i \hat{\beta}_i$. Therefore, the kernel estimator can be written as

$$\hat{\beta}_{i,\lambda} = \frac{(1-\lambda) x_i' x_i \hat{\beta}_i + \lambda \sum_{l=1}^{c} x_l' x_l \hat{\beta}}{(1-\lambda) x_i' x_i + \lambda \sum_{l=1}^{c} x_l' x_l},$$

which we could write as

$$\hat{\beta}_{i,\lambda} = \frac{x_i' x_i \hat{\beta}_i + \frac{\lambda}{1-\lambda}\delta\hat{\beta}}{x_i' x_i + \frac{\lambda}{1-\lambda}\delta}, \tag{1}$$

where $\delta = \sum_{l=1}^{c} x_l' x_l$ depends on data through the covariates but not on the group or the responses.

It is useful to gain intuition by considering the balanced case in which $s = x_i' x_i$ is the same for all i. In this case

$$\hat{\beta}_{i,\lambda} = \frac{\hat{\beta}_i + \frac{c\lambda}{1-\lambda}\hat{\beta}}{1 + \frac{c\lambda}{1-\lambda}}, \tag{2}$$

showing clearly that the kernel estimator in each group is a weighted average of the within group OLS estimator and the pooled OLS estimator.

2.2. **Bayes Estimators.** We consider a three-stage hierarchical Bayes model. The first stage is given by

$$\mathbf{y} \sim (A_1\beta, C_1).$$

As a function of β and C_1 for given y, this first stage specification can be regarded as the likelihood function for the normally distributed case, otherwise as a quasi likelihood based on two moments (Heyde 1997). We return to A_1 below.

The second stage,

$$\beta \sim (A_2\theta_2, C_2),$$

can be regarded as a prior distribution for β given $A_2\theta_2$ and C_2 in the normal case (where it is conjugate) or as an approximation to the prior if not normal, or from a frequency viewpoint as a second stage in the data generating process (DGP). The first stage "parameters" are themselves generated by a random process in this view. This interpretation focuses attention on the hyperparameters θ_2 (and C_2) rather than β which strictly speaking is not a parameter in the frequency sense.

The third stage,

$$\theta_2 \sim (A_3\theta_3, C_3),$$

can again be regarded as a prior on the second stage parameter θ_2, or as an additional stage in the DGP.

Our interest lies in estimating the $c \times 1$ vector of coefficients β. Following Lindley & Smith (1972) we are thinking of normal distributions at each stage. For our purposes we can also regard the stages as approximate distributions characterized by two moments noting the calculations are exact only for the normal. The point of the stages is that the dimension of the conditioning parameter is reduced at each step.

We are using the Bayesian hierarchical setup to obtain insight into the kernel estimator. The full Bayesian analysis will require additional specification in the form of a prior on C_1 and possibly C_2. Lindley & Smith (1972) suggest specifications proportional to identity matrices and inverted gamma densities for the factors of proportion (and related generalizations). They suggest using modal estimators in the expressions for the posterior means of interest. Using MCMC methods it is now possible to marginalize with respect to these variances, probably a better procedure; see Seltzer et al. (1996).

For the problem at hand, we try to stick with the notation of Lindley & Smith (1972) as closely as possible. The first stage is

$$A_1 = \{a_{ji}\} \text{ with } a_{ji} \in \{0, X_{ji}\}, \sum_{i=1}^{c} a_{li} = X_{ji}, \sum_{l=1}^{n} a_{li} = \iota_i' x_i,$$

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_c \end{bmatrix},$$

$$C_1 = \sigma^2 I_n,$$

A_1 is the $n \times c$ design matrix with $A_1'A_1$ the $c \times c$ diagonal matrix with $x_i'x_i$, the sum of squared regressors in the ith group, as the ith diagonal element, β is a $c \times 1$ vector of coefficients σ^2 is the within-group variance (i.e., $\text{var}(y_{ij})$), and I_n is the $n \times n$ identity matrix. The idea here is to get at the relation between kernel and Bayes estimators in a very simple

model where the effects of a single continuous covariate are different across c groups. The general case with l regressors, continuous or discrete, is discussed in the next section. Next, the second stage will become

$$\begin{aligned} A_2 &= \iota_c, \\ \theta_2 &= \beta_{\cdot}, \\ C_2 &= \tau^2 I_c, \end{aligned}$$

where $\beta_{\cdot}$ is the average effect (the average slope), and $\tau^2 = \mathrm{var}(\beta_i)$. Note that $A_2\theta_2 = \iota_c\beta_{\cdot}$ is simply a $c \times 1$ vector with elements being the mean effect $\beta_{\cdot}$ to which the Bayes (and kernel) estimators can shrink. Finally, we let the scalar

$$C_3^{-1} \to 0$$

so that the prior on $\beta_{\cdot}$ is improper. Note that the impropriety is confined to one dimension. The frequency analysis corresponds to an improper prior on the c-vector β, so that we expect inadmissibility of the frequency estimator through a Stein effect if $c > 2$. By adding a third stage, we reduce the improper prior in this single regressor setting to one dimension. The results are seen below.

The three stage Bayes estimate is (Lindley & Smith 1972, page 7, Equation (16))

$$\beta^* = D_0 d_0 \tag{3}$$

where

$$\begin{aligned} D_0^{-1} &= \left(A_1' C_1^{-1} A_1 + C_2^{-1} - C_2^{-1} A_2 \left(A_2' C_2^{-1} A_2 \right)^{-1} A_2' C_2^{-1} \right) \\ d_0 &= \left(A_1' C_1^{-1} \mathbf{y} \right). \end{aligned} \tag{4}$$

β^* is the posterior mean and is an optimal estimator under quadratic loss. Writing

$$\Lambda = A_1' C_1^{-1} A_1 = \frac{1}{\sigma^2} \begin{bmatrix} x_1'x_1 & 0 & 0 & \dots \\ 0 & x_2'x_2 & 0 & \dots \\ \vdots & & \ddots & \\ \vdots & 0 & 0 & x_n'x_n \end{bmatrix}$$

we see that

$$\begin{aligned} D_0^{-1} &= \left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota_c \iota_c'/c\right), \\ d_0 &= A_1' C_1^{-1} \mathbf{y} \\ &= \sigma^{-2} \begin{pmatrix} x_1' y_1 \\ \vdots \\ x_c' y_c \end{pmatrix}. \end{aligned}$$

Thus the vector of posterior means satisfies

$$\left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota_c' \iota_c/c\right) \beta^* = d_0$$

or, element-wise

$$(\sigma^{-2} x_j' x_j + \tau^{-2}) \beta_j^* - \tau^{-2} \beta_\cdot^* = \sigma^{-2} x_j' y_j,$$

where $\beta_\cdot^* = \sum_{j=1}^c \beta_j^*/c$. Thus

$$\begin{aligned} \beta_j^* &= \frac{\sigma^{-2} x_j' y_j + \tau^{-2} \beta_\cdot^*}{\sigma^{-2} x_j' x_j + \tau^{-2}} \\ &= \frac{\sigma^{-2} x_j' x_j \hat{\beta} + \tau^{-2} \beta_\cdot^*}{\sigma^{-2} x_j' x_j + \tau^{-2}} \end{aligned}$$

and the Bayes estimator for the jth mean is a weighted average of the group OLS estimator and the overall posterior mean.

We now re-express this estimator in terms of the OLS estimators alone for comparison with the kernel specification. First, we use a convenient partitioned inversion formula, namely the Woodbury identity:

$$Q = (A + BDB')^{-1} = A^{-1} - A^{-1} B \left(B' A^{-1} B + D^{-1}\right)^{-1} B' A^{-1}. \tag{5}$$

Letting

$$\begin{aligned} A &= \Lambda + \tau^{-2} I_c, \\ B &= \iota, \\ D &= -\tau^{-2}/c, \end{aligned}$$

we have

$$Q = \left(\Lambda + \tau^{-2} I_c\right)^{-1} - \left(\Lambda + \tau^{-2} I_c\right)^{-1} \iota \left(\iota' \left(\Lambda + \tau^{-2} I_c\right)^{-1} \iota - c\tau^2\right)^{-1} \iota' \left(\Lambda + \tau^{-2} I_c\right)^{-1}.$$

Let $w_i = x_i'x_i/\sigma^2$, $d_i = x_i'x_i/\sigma^2 + \tau^{-2} = w_i + \tau^{-2}$. Note that

$$\iota' \left(\Lambda + \tau^{-2} I_c\right)^{-1} \iota - c\tau^2 = \sum_{i=1}^{c} \frac{1}{d_i} - c\tau^2 = -\tau^2 \sum_{i=1}^{c} \frac{w_i}{d_i} = -\tau^2 \eta.$$

Next, the Bayes estimator of the ith component of β is given by

$$\beta_i^* = d_i^{-1}\sigma^{-2}x_i'y_i + \tau^{-2}\eta^{-1}d_i^{-1}\sigma^{-2}\sum_{j=1}^{c}\frac{x_j'y_j}{d_j},$$

which can be written in terms of the OLS estimators for each group $\hat{\beta}_i$

$$\beta_i^* = d_i^{-1}\sigma^{-2}x_i'x_i\hat{\beta}_i + \tau^{-2}\eta^{-1}d_i^{-1}\sigma^{-2}\sum_{j=1}^{c}\frac{x_j'x_j\hat{\beta}_j}{d_j}.$$

The second term is $\tau^{-2}d_i^{-1}$ times a weighted average of the $\hat{\beta}_i$, which can be seen by verifying that $\eta\sigma^2$ is in fact the summed weights $\sum_{j=1}^{c}\frac{x_j'x_j}{d_j}$. The OLS estimator within each group is drawn toward an average of the OLS estimators over all the groups. A little more insight can be obtained in the balanced case ($w_i = w_j = s\sigma^{-2}$), in which the second term is an unweighted average of the group-specific OLS estimators, which with balance is equal to the overall OLS estimator $\hat{\beta}$, i.e.

$$\beta_i^* = \frac{\hat{\beta}_i + s^{-1}\sigma^2\tau^{-2}\hat{\beta}}{1 + s^{-1}\sigma^2\tau^{-2}}. \tag{6}$$

2.3. **Comparison of Estimators.** We now compare (2) with (6). Recall that $\tau^2 = \mathrm{var}(\beta_i)$, hence $\tau^{-2} \in [0, \infty]$, and that the kernel smoothing parameter $\lambda \in [0, 1]$, hence $\lambda/(1-\lambda) \in [0, \infty]$. The role of $s^{-1}\sigma^2\tau^{-2}$ for the Bayesian estimator defined in Equation (6) is that played by $c\lambda/(1-\lambda)$ for the kernel estimator defined in Equation (2). The role played by c in the balanced kernel case is that of the relative precision of $\hat{\beta}$ to $\hat{\beta}_j$. This is a little harder to break out in the Bayesian formulation. It is captured in $s^{-1}\sigma^2\tau^{-2}$, but τ also captures the influence of λ. Comparison of (2) and (6) also gives some intuition for the choice of the smoothing parameter λ if one chooses not to adopt the Bayesian approach explicitly. λ should be larger as the groups are thought to be more homogeneous (smaller τ^2) and smaller as the groups are thought to be less similar. Recalling that $s = x_i x_i'$, higher variance in x should lead to lower λ. Higher error variance should indicate higher λ. The signal to noise ratio s/σ^2 when higher leads to lower λ. Of course, if one is to do this thinking, it is natural to use the Bayesian specification directly, noting that the logic applies equally in the unbalanced case. In a special case (the c-means problem) Kiefer & Racine (2009) obtained conditions giving bounds on λ under which a MSE improvement was assured. Simulations showed that cross-validation produced λ-values satisfying these conditions.

Turning to implications for consistency, we note that standard arguments give consistency of the Bayes estimator. Looking at (6) we see that the effect of the prior vanishes as $n \to \infty$. In the unbalanced case we require $n_i/n \sim O(1)$. Looking at (1) we see that the effect of the kernel smoother λ does not vanish. Hence, for consistency we require $\lambda \sim O(n^{-1})$. Use less smoothing in larger samples.

3. The General Hierarchical Linear Model

There exist a number of variations on the hierarchical model according to the hierarchy structure, number of levels, and so forth. Below we consider a framework that is useful for not only fostering a direct comparison between kernel and Bayes estimators, but suggesting novel estimators that, to the best of our knowledge, have not been explored in the hierarchical setting.

The general hierarchical model allows for multiple covariates as well as multiple groups. Write

$$y_i = X_i\beta_i + \epsilon_i, \quad i = 1, \ldots, c, \tag{7}$$

as above, but now allow X_i to be the $n_i \times k$ matrix of n_i observations on k covariates in group c. The covariates can be continuous or discrete. By choice of regressor interactions and choice of the grouping into c groups the model accommodates a number of popular specifications. The OLS estimator can be characterized as above as

$$\hat{\beta}_i = \arg\min_{\beta_i} \sum_{j=1}^{c} (y_j - X_j\beta_i)'(y_j - X_j\beta_i)\mathbf{1}(j = i)$$

implying $\hat{\beta}_i = (X_i'X_i)^{-1}X_i'y_i$.

3.1. **A Categorical Kernel Approach.**

Li et al. (2013) propose a semiparametric kernel-based approach to the estimation of a smooth kernel model where the coefficients are grouped into c groups. For the general model the corresponding kernel estimator is

$$\hat{\beta}_{i,\lambda} = \arg\min_{\beta_i} \sum_{j=1}^{c} (y_j - X_j\beta_i)'(y_j - X_j\beta_i)L(i, j, \lambda) \tag{8}$$

with associated FOC

$$\sum_{l=1}^{c} L(l, i, \lambda)X_l'(y_l - X_l\beta_i) = 0 \tag{9}$$

leading to

$$\hat{\beta}_{i,\lambda} = \left[\sum_{j=1}^{c} L(i,j,\lambda) X_j' X_j\right]^{-1} \sum_{j=1}^{c} L(i,j,\lambda) X_j' y_j.$$

With $L(l,i,\lambda) = \lambda^{\mathbf{1}(l\neq i)}$ this simplifies to

$$\hat{\beta}_{i,\lambda} = \left[(1-\lambda) X_i' X_i + \lambda \sum_{j=1}^{c} X_j' X_j\right]^{-1} \left[(1-\lambda) X_i' y_i + \lambda \sum_{j=1}^{c} X_j' y_j\right], \tag{10}$$

which can be rewritten as a matrix-weighted average of the within-group OLS estimator and the pooled OLS estimator

$$\hat{\beta}_{i,\lambda} = \left[X_i' X_i + \frac{\lambda}{(1-\lambda)} \sum_{j=1}^{c} X_j' X_j\right]^{-1} \left[X_i' X_i \hat{\beta}_i + \frac{\lambda}{(1-\lambda)} \sum_{j=1}^{c} X_j' X_j \hat{\beta}_j\right]. \tag{11}$$

Equation (11) is useful for interpretation but (10) is more general as it accommodates the important case in which some or all of the $X_j' X_j$ are singular. Of course, the sum $\sum_{j=1}^{c} X_j' X_j$ must be nonsingular.

3.2. **A Novel Kernel Estimator.** A natural and practically useful generalization of this estimator is obtained by re-representing the kernel as $\mathbf{L}(l,i,\lambda)$, a $k \times k$ diagonal matrix with diagonal elements $\lambda_l^{\mathbf{1}(l\neq i)}, l = 1, \ldots, k$. Thus λ is now a vector. Substituting into the FOC in Equation (9) gives

$$\sum_{l=1}^{c} \mathbf{L}(l,i,\lambda) X_l' (y_l - X_l \beta_i) = 0. \tag{12}$$

Note that the equation system on the LHS of (12) is not the first derivative of a scalar function of β. Nevertheless, the resulting estimator is intuitively appealing and a sound foundation is given by the Bayesian analysis to follow. Solving (12) gives

$$\hat{\beta}_{i,\lambda} = \left[\sum_{j=1}^{c} \mathbf{L}(i,j,\lambda) X_j' X_j\right]^{-1} \sum_{j=1}^{c} \mathbf{L}(i,j,\lambda) X_j' y_j.$$

Letting $\boldsymbol{\lambda}$ be the $k \times k$ diagonal matrix with diagonal element λ_l we can write

$$\hat{\beta}_{i,\lambda} = \left[(I - \boldsymbol{\lambda}) X_i' X_i + \boldsymbol{\lambda} \sum_{j=1}^{c} X_j' X_j\right]^{-1} \left[(I - \boldsymbol{\lambda}) X_i' y_i + \boldsymbol{\lambda} \sum_{j=1}^{c} X_j' y_j\right]. \tag{13}$$

This can be represented in terms of the least-squares estimators as

$$\hat{\beta}_{i,\lambda} = \left[(I-\boldsymbol{\lambda})X_i'X_i + \boldsymbol{\lambda}\sum_{j=1}^{c} X_j'X_j\right]^{-1}\left[(I-\boldsymbol{\lambda})X_i'X_i\hat{\beta}_i + \boldsymbol{\lambda}\sum_{j=1}^{c} X_j'X_j\hat{\beta}_j\right].$$

This representation requires that each $X_i'X_i$ be invertible. In the special case in which each λ_l is less than one (so each element of β is "shrunk") the matrix $(I-\boldsymbol{\lambda})$ is invertible and we have an alternative representation useful for comparison with the Bayes estimator to come,

$$\hat{\beta}_{i,\lambda} = \left[X_i'X_i + (I-\boldsymbol{\lambda})^{-1}\boldsymbol{\lambda}\sum_{j=1}^{c} X_j'X_j\right]^{-1}\left[X_i'X_i\hat{\beta}_i + (I-\boldsymbol{\lambda})^{-1}\boldsymbol{\lambda}\sum_{j=1}^{c} X_j'X_j\hat{\beta}_j\right].$$

Equation (13), a generalization of Equation (11), does not appear to have been explored in a hierarchical setting. Some progress toward putting it on a sound foundation is provided by the Bayesian analysis. Additional interpretation can be developed by considering the "balanced" case $X_j'X_j = X'X\ \forall j$:

$$\hat{\beta}_{i,\lambda} = [(I-\boldsymbol{\lambda})X'X + \boldsymbol{\lambda}cX'X]^{-1}\left[(I-\boldsymbol{\lambda})X'X\hat{\beta}_i + \boldsymbol{\lambda}cX'X\hat{\beta}\right],$$

where $\hat{\beta}$ is the overall OLS estimator. Thus $\hat{\beta}_{i,\lambda}$ is a matrix-weighted average of the within estimator and the pooled estimator.

3.3. **A Bayesian Approach.** For the Bayes estimator we again consider the 3-stage hierarchical model $\mathbf{y} \sim (A_1\beta, C_1)$, $\beta \sim (A_2\theta_2, C_2)$, and $\theta_2 \sim (A_3\theta_3, C_3)$ with now

$$A_1 = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & X_c \end{bmatrix},$$

$\beta = (\beta_1', \beta_2', \ldots, \beta_c')'$ and $C_1 = \sigma^{-2}I_{ck}$. Here A_1 is $n \times ck$ and β $ck \times 1$. In the second stage we have $A_2 = (I_k, \ldots, I_k)'$, a stack of c $k \times k$ identity matrices, and $\theta_2 = \beta_0$, a k-dimensional common prior mean for the β_j, and

$$C_2 = \begin{bmatrix} T & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & T \end{bmatrix}$$

a block-diagonal matrix with c $k \times k$ blocks of the prior variance matrix T. In the third stage we again let $C_3^{-1} \to 0$.

Using (3) and (4) we see following the development above that the posterior means satisfy

$$(\sigma^{-2}X_j'X_j + T^{-1})\beta_j^* - T^{-1}\beta_\cdot^* = \sigma^{-2}X_j'y_j$$

where

$$\beta^*_{\cdot} = c^{-1} \sum_j \beta^*_j$$

so

$$\begin{aligned}\beta^*_i &= (\sigma^{-2} X_i' X_i + T^{-1})^{-1} (\sigma^{-2} X_j' y_j + T^{-1} \beta^*_{\cdot}) \\ &= (\sigma^{-2} X_i' X_i + T^{-1})^{-1} (\sigma^{-2} X_i' X_i \hat{\beta}_i + T^{-1} \beta^*_{\cdot}),\end{aligned}$$

a matrix-weighted average of the OLS estimator and the average of the posterior means. To get this in the form of a weighted average of the OLS and average OLS estimator we again use the partitioned inversion formula (5) and define $W_j = \sigma^{-2} X_j' X_j$ and $D_j = W_j + T^{-1}$. The matrix inverted in the RHS of (5) is $\sum_j D_j^{-1} - cT$. Simplify by noting that $D_j^{-1} - T = -TW_j D_j^{-1}$. Write $\Xi = \sum_j W_j D_j^{-1}$. Then we can write

$$\begin{aligned}\beta^*_j &= (\sigma^{-2} X_j' X_j + T^{-1})^{-1} (\sigma^{-2} X_j' y_j + \sigma^{-2} T^{-1} \Xi^{-1} \sum_i D_i^{-1} X_i y_i) \\ &= (\sigma^{-2} X_j' X_j + T^{-1})^{-1} (\sigma^{-2} X_j' X_j \hat{\beta}_j + \sigma^{-2} T^{-1} \Xi^{-1} \sum_i D_i^{-1} X_i X_i \hat{\beta}_i),\end{aligned}$$

giving the posterior mean as a matrix weighted average of the within-group and the pooled OLS estimators. Again, the first representation may be most useful since it does not require inversion of each $X_j' X_j$, while the second may be more useful for interpretation and comparison with the kernel estimator.

Again, we gain insight by examining the balanced case with $X_j' X_j = X'X$ for all j and hence equal W_j and D_j. Here

$$\beta^*_j = (\sigma^{-2} X'X + T^{-1})^{-1} (\sigma^{-2} X'X \hat{\beta}_j + T^{-1} \hat{\beta}),$$

where $\hat{\beta}$ is the pooled OLS estimator. The balanced case is unlikely to arise in practice except by design. Nevertheless it offers clear insight into the relationship between the estimators and hence some guidance for bandwidth selection. The connection is similar but more complicated in the unbalanced case, as seen from our general expressions.

4. What is this Class of Models?

The hierarchical model captures a wide class of useful models as special cases as well as a range of estimators that may offer MSE advantages and clearly allow incorporation of prior confidence in the specifications. To fix ideas consider the simple parametric model

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij},$$

where x_{ij} is a scalar. Using our kernel method the bandwidths $(\lambda_1, \lambda_2) \in [0,1]^2$ must be specified. The endpoint $\lambda = [0,0]$ is the case of separate regressions for each of the groups indexed by j. The endpoint $\lambda = [1,1]$ is the pooled regression estimator, with the (α, β) constrained to be the same across each group. The endpoint $\lambda = [0,1]$ is the "fixed effect" model familiar from panel data analysis. In this model the effect of the regressors are the same across groups, but differences in intercepts (group locations) due perhaps to fixed but unobserved variables are incorporated. Finally, the endpoint $\lambda = [1,0]$ fixes the intercepts across groups but allows different slopes. An example might be a system of demand equations, which require zero quantities at zero expenditure but which allow for different price responses.

Our general class allows these models and all models in between as defined by varying λ. As long as $\lambda \to 0$ as $n \to \infty$ the models are all consistent estimators for the most general specification. Previous work has shown MSE improvement in examples and we expect this is available generally.

5. Nonparametric Hierarchical Regression with B-Splines

The categorical kernel-based models outlined above are semiparametric since they require specification of the functional relationship among the non-categorical regressors and response along the lines of the semiparametric estimator proposed in Li et al. (2013). However, in many applications a fully nonparametric specification may be required. The framework we consider can immediately be generalized to a fully nonparametric specification by replacing the regressors with an appropriate spline basis. We consider splines simply because they constitute a powerful generalization that is immediately accessible to those familiar with polynomial regression via least squares fitting. Below we generalize the categorical kernel-based approach outlined in Section 3.1 to a semiparametric additive B-spline regression model (i.e. allow for nonlinearity with respect to each continuous regressor but retain the additive structure) or a fully nonparametric B-spline regression model using the approaches of Ma & Racine (2013) and Ma et al. (2012), respectively. This approach may appeal to practitioners comfortable with weighted least-squares estimation who otherwise might resist semiparametric and nonparametric methods. Some background is provided for the interested reader who may not be familiar with B-splines, while others can skip to the proceeding section.

5.1. **A Brief Overview.**

Spline regression is a nonparametric technique that involves nothing more than replacing a model's regressors (and perhaps 'raw polynomials' thereof) with

their B-spline bases (which are themselves polynomials). Spline methods can deliver consistent estimates of a broad range of DGPs, hence their appeal. Naturally we must determine the optimal 'order', number of interior 'knots', and bandwidths for the model (see below), however, the reader may immediately recognize that, when extended to admit categorical regressors, this involves little more than weighted least squares estimation. A 'spline' is a function that is constructed piece-wise from polynomial functions, and we focus attention on a class of splines called 'B-splines' ('basis-splines'). We consider 'regression spline' methodology which differs in a number of ways from 'smoothing splines', both of which are popular in applied settings. The fundamental difference between the two approaches is that smoothing splines use the data points themselves as potential knots whereas regression splines place knots at equidistant/equiquantile points. Also, smoothing splines explicitly penalize 'roughness' where curvature (i.e. second derivative) is a proxy for roughness. We direct the interested reader to Wahba (1990) for a treatment of smoothing splines.

The B-spline is a generalization of the Bézier curve and is popular due to a fundamental theorem (cf de Boor (2001)) stating that every spline of a given degree and smoothness can be represented as a linear combination of B-splines (the B-spline function is the maximally differentiable interpolative basis function, while a B-spline with no 'interior knots' is a Bézier curve). B-splines are defined by their 'order' m and number of interior 'knots' N (there are two 'endpoints' which are themselves knots so the total number of knots will be $N+2$ which we denote by $t_0, \ldots, t_{N+1}$). The degree d of the B-spline polynomial is the spline order minus one (i.e. $d = m - 1$).

A B-spline of degree d is a parametric curve composed of a linear combination of basis B-splines $B_{i,d}(x)$ of degree d given by

$$B(x) = \sum_{i=0}^{N+n} \beta_i B_{i,d}(x), \quad x \in [t_0, t_{N+1}].$$

The β_i are called 'control points' or 'de Boor points', the t_j the knots. For an order m B-spline having N interior knots there are $K = N + m = N + d + 1$ control points (one when $j = 0$). The B-spline order m must be at least 2 (hence at least linear, i.e. degree d is at least 1) and the number of interior knots must be non-negative ($N \geq 0$).

Figure 1 presents an illustration where we consider order $m = 4$ (i.e. degree = 3) basis B-splines $B_{0,3}(x), \ldots, B_{6,3}(x)$ (left) with 4 sub-intervals (segments) using uniform knots ($N = 3$ interior knots, 5 knots in total (2 endpoint knots)) and the 1st-order derivative basis B-splines $B'_{0,3}(x), \ldots, B'_{6,3}(x)$ which is needed for computation of marginal effects (right). The dimension of $B(x)$ is $K = N + m = 7$.

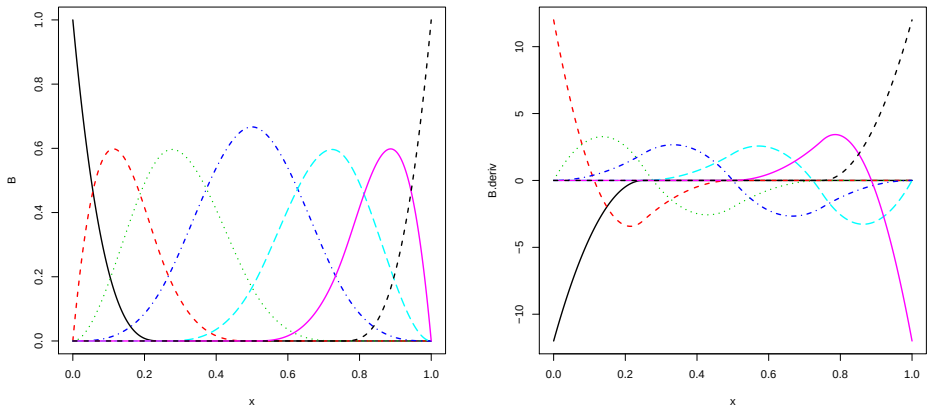

FIGURE 1. Fourth-order B-spline basis functions with three interior knots and the associated 1st derivative B-spline basis functions.

In general we will have k (continuous) regressors, $\mathbf{X} = (X_1, \ldots, X_k)'$, each having its own basis. There are two types of multivariate B-spline basis systems used, namely the 'tensor-product' and 'additive' bases. Letting $\otimes$ denote tensor product, then $\mathcal{B}(\mathbf{x}) = B_1(x_1) \otimes \cdots \otimes B_k(x_k)$ is a tensor basis system where the m_j and N_j represent the spline order and number of interior knots for the jth regressor, $j = 1, \ldots, k$. This multivariate tensor-product B-spline is quite powerful, but may exhaust degrees of freedom fairly rapidly as k increases. Similarly, we can define the multivariate additive B-spline which is naturally simpler as they simply involve concatenation of the univariate spline bases and consume fewer degrees of freedom than their tensor-based counterpart (i.e. $\mathcal{B}(\mathbf{x}) = B_1(x_1) + \cdots + B_k(x_k)$). In high-dimensional settings additive splines may be preferred/necessary, though as an anonymous referee pointed out, we emphasize that the additive B-spline model is applicable only to an additive regression model.

For the general model (7) the corresponding B-spline-based nonparametric estimator is

$$\hat{\beta}_{i,\lambda} = \arg\min_{\beta_i} \sum_{j=1}^{c} (y_j - \mathcal{B}(X_j)\beta_i)'(y_j - \mathcal{B}(X_j)\beta_i) L(i, j, \lambda) \tag{14}$$

with associated FOC

$$\sum_{l=1}^{c} L(l, i, \lambda)\mathcal{B}(X_l)'(y_l - \mathcal{B}(X_l)\beta_i) = 0 \tag{15}$$

leading to

$$\hat{\beta}_{i,\lambda} = \left[\sum_{j=1}^{c} L(i,j,\lambda)\mathcal{B}(X_j)'\mathcal{B}(X_j)\right]^{-1} \sum_{j=1}^{c} L(i,j,\lambda)\mathcal{B}(X_j)'y_j. \tag{16}$$

Note again that we have simply replaced X_j by $\mathcal{B}(X_j)$, all else is unchanged. All results obtained for the model considered in Section 3.1 hold for this estimator via simple substitution of $\mathcal{B}(X_j)$ for X_j. Next we discuss data-driven selection of spline degree(s), knot(s), and bandwith(s).

Cross-validation has a rich pedigree in the regression spline arena and has been used for decades to choose the appropriate number of interior knots. Following in this tradition we can choose both the degree and number of interior knots (i.e. the vectors m and N) and kernel smoothing parameters (i.e. the bandwidth vector λ) by minimizing the cross-validation function defined by

$$CV(m,N,\lambda) = n^{-1}\sum_{i=1}^{n}(Y_i - \mathcal{B}(X_i)'\hat{\beta}_{-i,\lambda})^2,$$

where $\hat{\beta}_{-i,\lambda}$ denotes the leave-one-out estimate of β. Cross-validation has a number of appealing theoretical and practical properties, including the ability to automatically remove irrelevant regressors with probability approaching one asymptotically without the need for pretesting. For further details we refer the reader to Ma & Racine (2013) and Ma et al. (2012).

6. An Illustrative Example

For the following illustration we consider Wooldridge's `wage1` dataset and consider three models for an earnings equation, namely i) a parametric model, ii) a semiparametric kernel-based model, and iii) a nonparametric kernel-based model.[3] The aim of this section is simply to demonstrate that the estimators considered above may appeal to practitioners.

We first estimate a parametric regression model where the response is `lwage` and the regressors are a constructed variable Z having 8 outcomes being the unique combinations of the categorical variables `female`, `nonwhite` and `married`, along with the variables `educ`, `exper`, and `tenure` which are treated as continuous, and we also allow `exper` to enter as a quadratic via `I(exper**2)` (this model delivers the OLS estimators for each group, $\hat{\beta}_i$, $i = 0, \ldots, c-1$, $c = 2 \times 2 \times 2 = 8$, hence the model contains 40 parameters in total). This

[3]For what follows we consider an implementation in the `R` language for statistical computing and graphics. See `?wage1` in the `R` (R Core Team (2013)) package `np` (Hayfield & Racine (2008)).

corresponds directly to running separate regressions for each of the $c = 8$ groups, and is sometimes called the 'frequency estimator'.

We then fit a semiparametric categorical kernel-based model where the coefficients can change with the categorical covariates (i.e. $\hat{\beta}_{i,\lambda}$ in Equation (8)). When $\lambda = 0$ this model is the parametric model described above, i.e. the frequency estimator. Cross-validation is used to determine the appropriate value of λ. When $\lambda > 0$ the coefficients shrink towards the overall OLS coefficients, while when $\lambda = 1$ this is equivalent to pooled OLS. Both the parametric and semiparametric models, as noted above, specify a linear additive relationship between the regressors and response.

Finally, we fit a nonparametric categorical kernel-based B-spline model (i.e. $\hat{\beta}_{i,\lambda}$ in Equation (16) where here we drop the regressor `I(exper**2)` since we use B-splines to model potential nonlinearity). The kernel and kernel B-spline models use cross-validation for selecting smoothing parameters (bandwidths, spline degree, number of knots) while the kernel B-spline model in addition uses cross-validation for determining whether to use the additive or tensor basis. When the additive basis is selected the model is a semiparametric additive kernel model which is more flexible than the linear-in-parameters semiparametric kernel model outlined in the previous paragraph since the relationship between the response and *each* regressor is modelled nonlinearly using B-splines. See Ma & Racine (2013) and Ma et al. (2012) along with the R (R Core Team 2013) packages 'crs' (Racine & Nie 2014) and 'np' (Hayfield & Racine 2008) for implementation and further details.

Semiparametric and nonparametric methods are sometimes criticized for 'overfitting' the data at hand. The parametric (i.e. frequency) model itself could be overfit since all parameters are allowed to vary with respect to all realizations of the categorical covariates. Readers are no doubt properly skeptical of model comparison based upon in-sample measures of fit such as R^2 and their ilk. In this illustration in-sample R^2 are 0.4976 for the parametric model, 0.4666 for the kernel model, and 0.5116 for the kernel B-spline model.

Readers would also likely concur that the model that performs best on *independent* data taken from the same DGP is closest to the true unknown DGP. In order to assess which of the above models performs the best in terms of squared prediction error on *unseen* data taken from the same DGP, we follow Racine & Parmeter (2014) and assess each model's out-of-sample performance by splitting the data set $S = 10,000$ times into two independent samples of size $n_1 = 520$ and $n_2 = 6$. Predicted square error (PSE) is computed for the n_2 hold-out observations via $n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{Y}_i)^2$ where the predictions $\hat{Y}_i$ are those obtained from the regressors in the hold-out sample (for comparison purposes we compute the same measure for the in-sample measures based on the full sample predictions). We then assess

expected performance on the hold-out data via the median and means taken over all S replications.

Results are summarized in Figure 2 and tables 1-3. Median out-of-sample PSE is 0.1513 for the parametric model, 0.1338 for the kernel model, and 0.1276 for the kernel B-spline model (mean out-of-sample PSE is 2.3946 for the parametric model, 0.1663 for the kernel model, and 0.1576 for the kernel B-spline model). Applying the test for revealed performance of Racine & Parmeter (2014) indicates that the kernel B-spline model performs significantly better than its peers (p-value $< 2.2e-16$) These results indicate that the kernel B-spline model performs best in terms of its ability to predict unseen data taken from the same DGP. We note that in-sample PSEs (i.e. residual variances) are 0.1417 for the parametric model, 0.1504 for the kernel, and 0.1374 for the kernel B-spline model which mirrors R^2 and is of limited utility as a basis for model selection.

Based on these results, it would appear that the kernel-based approaches may hold much appeal to practitioners. And the fact that these methods have been placed on a sound footing by drawing the connection between them and Bayes models delivers additional insight into their performance and behaviour.

7. Concluding Remarks

We have established a relation between kernel and Bayesian hierarchical models. This relationship provides a sound statistical foundation for kernel methods that have proven themselves practically useful. Exploring this relationship led to a new class of kernel estimators for grouped data including pooled regression, separate regressions, fixed-effect models, and models with common intercepts but different slopes as special cases. All models "in between" these are covered. In the Bayesian case the model is determined by prior information. In the kernel case the model can be determined by cross validation. Extension to a fully nonparametric approach via B-splines is straightforward. An application shows that the approach can deliver specifications with good (prediction) properties.

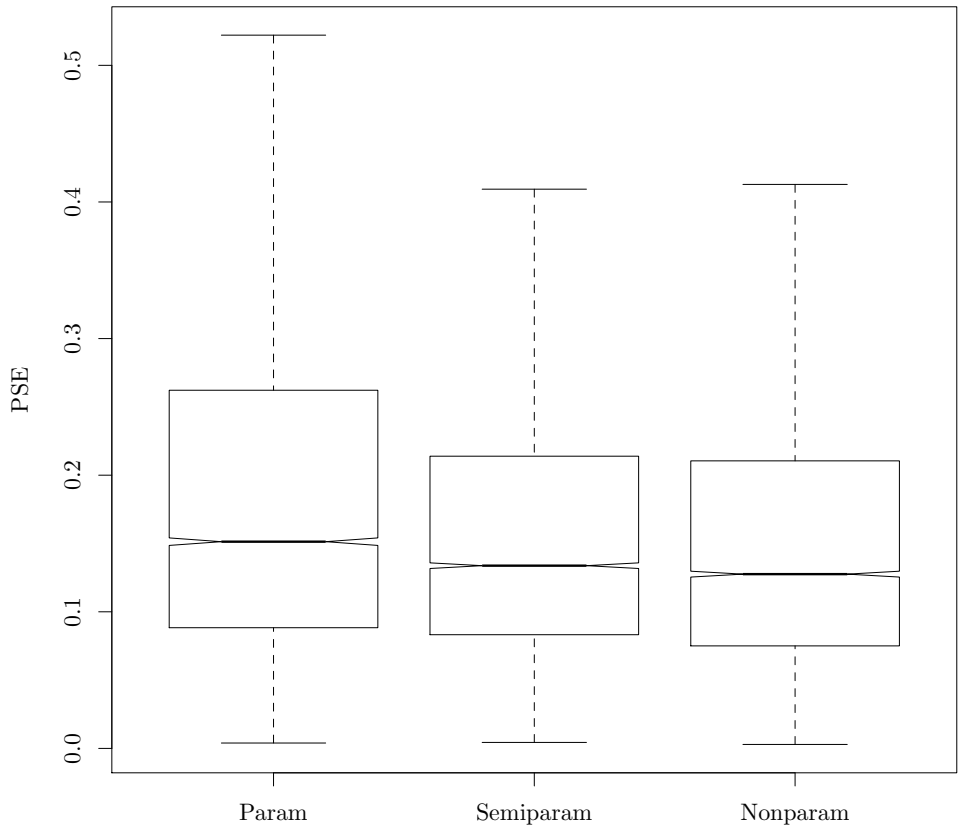

FIGURE 2. Comparison of out-of-sample predictive performance when the data is split into 10,000 training and validation samples of size $n_1 = 520$ and $n_2 = 6$. Median out-of-sample PSE is 0.1513 for the parametric model ('Param'), 0.1338 for the kernel model ('Semiparam'), and 0.1276 for the kernel B-spline model ('Nonparam').

References

Breusch, T. S., Mizon, G. E. & Schmidt, P. (1989), 'Efficient estimation using panel data', *Econometrica* **57**(3), 695–700.

de Boor, C. (2001), *A practical guide to splines*, Springer.

Griffin, J. E. & Steel, M. F. J. (2010), 'Bayesian nonparametric modelling with the dirichlet process regression smoother', *Statistica Sinica* **20**, 1507–1527.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
URL: *http://www.jstatsoft.org/v27/i05/*

Heyde, C. (1997), *Quasi-likelihood and Its Application*, Springer-Verlag.

Judge, G. G. & Bock, M. E. (1976), 'A comparison of traditional and stein-rule estimators under weighted squared error loss', *International Economic Review* **17**(1), pp. 234–240.
URL: *http://www.jstor.org/stable/2526079*

Karabatsos, G. & Walker, S. G. (2012), 'Adaptive-modal bayesian nonparametric regression', *Electronic Journal of Statistics* **6**, 1935–7524.

Kiefer, N. M. & Racine, J. S. (2009), 'The smooth colonel meets the reverend', *Journal of Nonparametric Statistics* **21**, 521–533.

Li, Q., Ouyang, D. & Racine, J. (2013), 'Categorical semiparametric varying coefficient models', *Journal of Applied Econometrics* **28**, 551–579.

Lindley, D. V. & Smith, A. F. M. (1972), 'Bayes estimates for the linear model', *Journal of the Royal Statistical Society* **34**, 1–41.

Ma, S. & Racine, J. S. (2013), 'Additive regression splines with irrelevant categorical and continuous regressors', *Statistica Sinica* **23**, 515–541.

Ma, S., Racine, J. S. & Yang, L. (2012), 'Spline regression in the presence of categorical predictors', *Journal of Multivariate Analysis* . Revised and Resubmitted.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Racine, J. & Parmeter, C. (2014), Data-driven model evaluation: A test for revealed performance, *in* A. Ullah, J. Racine & L. Su, eds, 'Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics', Oxford University Press, pp. 308–345.

Racine, J. S. & Nie, Z. (2014), *crs: Categorical Regression Splines*. R package version 0.15-22.
URL: *https://github.com/JeffreyRacine/R-Package-crs/*

Seltzer, M. H., Wong, W. H. & Bryk, A. S. (1996), 'Bayesian analysis in applications of hierarchical models: Issues and methods', *Journal of Educational and Behavioral Statistics* **21**, 131–167.

Stone, C. J. (1974), 'Cross-validatory choice and assessment of statistical predictions (with discussion)', *Journal of the Royal Statistical Society* **36**, 111–147.

Theil, H. & Goldberger, A. S. (1961), 'On pure and mixed statistical estimation in economics', *International Economic Review* **2**(1), pp. 65–78.
URL: *http://www.jstor.org/stable/2525589*

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM.

Appendix: Model Summaries

Table 1. Parametric Model Summary

```
Linear Regression Model
Regression data: 526 training points, in 1 variable(s)

              categ8
Bandwidth(s):      0

Bandwidth Type: Fixed

Residual standard error: 0.3764
R-squared: 0.4976

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 1
```

Table 2. Semiparametric Kernel-Based Model Summary

```
Smooth Coefficient Model
Regression data: 526 training points, in 1 variable(s)

              categ8
Bandwidth(s): 0.3491

Bandwidth Type: Fixed

Residual standard error: 0.3879
R-squared: 0.4666

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 1
```

TABLE 3. Nonparametric Kernel-Based B-Spline Model Summary

```
Kernel Weighting/B-spline Bases Regression Spline

There are 3 continuous predictors
There is 1 categorical predictor
Spline degree/number of segments for educ: 1/2
Spline degree/number of segments for exper: 3/3
Spline degree/number of segments for tenure: 1/3
Bandwidth for categ8: 0.1466
Model complexity proxy: degree-knots
Knot type: quantiles
Basis type: additive
Training observations: 526
Rank of model frame: 11
Trace of smoother matrix: 32

Residual standard error: 0.3751 on 515 degrees of freedom
Multiple R-squared: 0.5116,   Adjusted R-squared: 0.5022
F-statistic: 16.68 on 31 and 494 DF, p-value: 2.186e-58

Cross-validation score: 0.15646184
Number of multistarts: 10
Estimation time: 314.5 seconds
```