Kiefer, Nicholas M.; Racine, Jeffrey S.

Working Paper

The smooth colonel meets the reverend

CAE Working Paper, No. 08-01

Provided in Cooperation with:
Center for Analytical Economics (CAE), Cornell University

Suggested Citation: Kiefer, Nicholas M.; Racine, Jeffrey S. (2008) : The smooth colonel meets the reverend, CAE Working Paper, No. 08-01, Cornell University, Center for Analytical Economics (CAE), Ithaca, NY

This Version is available at:
https://hdl.handle.net/10419/70459

ISSN 1936-5098

CAE Working Paper #08-01

The Smooth Colonel meets the Reverend

by

Nicholas M. Kiefer
and
Jeffrey S. Racine

May 2008

THE SMOOTH COLONEL MEETS THE REVEREND

NICHOLAS M. KIEFER AND JEFFREY S. RACINE

Abstract. Kernel smoothing techniques have attracted much attention and some notoriety in recent years. The attention is well deserved as kernel methods free researchers from having to impose rigid parametric structure on their data. The notoriety arises from the fact that the amount of smoothing (i.e., local averaging) that is appropriate for the problem at hand is under the control of the researcher. In this paper we provide a deeper understanding of kernel smoothing methods for discrete data by leveraging the unexplored links between hierarchical Bayes models and kernel methods for discrete processes. A number of potentially useful results are thereby obtained, including bounds on when kernel smoothing can be expected to dominate non-smooth (e.g., parametric) approaches in mean squared error and suggestions for thinking about the appropriate amount of smoothing.

1. Introduction

We investigate the relationship between nonparametric discrete kernel methods and hierarchical Bayes models of the type considered by Lindley & Smith (1972). By exploiting certain similarities among the approaches, we not only gain a deeper understanding of the nature of kernel-based methods, but also leverage some theoretical apparatus developed for hierarchical Bayes models which is immediately relevant for kernel-based techniques.

This paper proceeds as follows. Section 2 provides some background material for the kernel smoothing of discrete probabilities and conditional means that is necessary for what follows. Section 3 presents a three-stage hierarchical Bayes framework and makes explicit the connection between the prior variance of a multivariate mean vector and the smoothing parameter in the kernel estimator. Section 4 considers some implications for applied discrete kernel regression, while Section 5 presents some summary remarks along with directions for future research.

Date: April 9, 2008.
Key words and phrases. Kernel estimation, Bayesian Methods, hierarchical models, nonparametrics, bandwidth selection.
We would like to thank but not implicate Esfandiar Maasoumi for his insightful comments and suggestions.

ISSN 1936-5098

CAE Working Paper #08-01

The Smooth Colonel meets the Reverend

by

Nicholas M. Kiefer
and
Jeffrey S. Racine

May 2008

2. Background

We first consider an unordered discrete variable having c outcomes, which is used strictly to indicate group membership. We let $X \in \mathcal{S} \equiv \{1, 2, \ldots, c\}$. For arbitrary $i \in \mathcal{S}$, let n_i denote the number of $X_{jk} = i$, in any given sample. The indices i and k denote the 'group' from which X is drawn $(i, k = 1, \ldots, c)$, while the index j denotes the jth draw from the group, $j = 1, \ldots, n_i$. The total number of observations will be $n = \sum_{i=1}^{c} n_i$, so that $n - n_i$ is the number of $X_{jk} \neq i$.

Our interest lies with conditional mean models of the type recently considered by Ouyang, Li & Racine (2008, in press). Given that such models are a function of the underlying probabilities, we take this as a starting point for developing some background and notation.

2.1. **Probability Function Estimation.**

We begin by assuming that interest lies in estimating $\Pr(X = i) = p(i)$ given a sample of realizations $\{X_{ji}\}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, c$. We consider two approaches, i) the traditional ('frequency' i.e., non-smooth) estimator and ii) a kernel (smooth) estimator.

Define the frequency estimator of $p(i)$ to be

$$p_i = \frac{1}{n} \sum_{k=1}^{c} \sum_{j=1}^{n_k} \mathbf{1}(X_{jk} = i) = \frac{n_i}{n},$$

where $\mathbf{1}(X_{jk} = i)$ is the usual indicator function. The kernel estimator of $p(i)$ is given by

$$p_{i,\lambda} = \frac{1}{n} \sum_{k=1}^{c} \sum_{j=1}^{n_k} L(X_{jk}, i, \lambda),$$

where

$$L(X_{jk}, i, \lambda) = \begin{cases} 1 - \lambda & \text{if } X_{jk} = i \\ \lambda/(c-1) & \text{otherwise,} \end{cases} \tag{1}$$

and where $\lambda \in [0, (c-1)/c]$ is a 'smoothing parameter' or 'bandwidth' (Aitchison & Aitken (1976)). The restriction that $\lambda \in [0, (c-1)/c]$ ensures that $p_{i,\lambda}$ is a proper probability estimator (i.e., $p_{i,\lambda} \in [0, 1]$).

Note that we can rewrite $p_{i,\lambda}$ as follows,

$$\begin{aligned} p_{i,\lambda} &= \frac{1}{n} \sum_{k=1}^{c} \sum_{j=1}^{n_k} L(X_{jk}, i, \lambda) \\ &= \frac{n_i(1-\lambda) + (n-n_i)\lambda/(c-1)}{n} \\ &= \frac{n_i \left(1 - \lambda c/(c-1)\right)}{n} + \frac{\lambda}{(c-1)} \\ &= p_i \left(1 - \frac{\lambda c}{(c-1)}\right) + \frac{\lambda}{(c-1)}. \end{aligned}$$

Note that when $\lambda = 0$, $p_{i,\lambda} = p_i = n_i/n$ (the frequency estimator), while when $\lambda = (c-1)/c$ (i.e., $(1 - \lambda c/(c-1)) = 0$), $p_{i,\lambda} = 1/c$, the discrete uniform (rectangular) distribution.

2.2. **Conditional Mean Estimation.** Now suppose we are interested in estimating $\mu_i = E(Y|X = i)$, the expectation of Y conditional upon $X = i$ based on a sample of realizations $\{X_{ji}, Y_{ji}\}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, c$. We again consider two approaches, a traditional frequency approach and a kernel-based approach. We first define some frequency estimators of certain population moments that shall be used to simplify the kernel-based estimator.

Let y_i be the frequency estimator of μ_i defined as

$$y_i = \frac{1}{n_i} \sum_{k=1}^{c} \sum_{j=1}^{n_k} Y_{jk} \mathbf{1}(X_{jk} = i), \tag{2}$$

i.e., the sample mean of Y when $X = i$ (a 'cell' mean). Let $y_{\bar{i}}$ be defined as

$$y_{\bar{i}} = \frac{1}{(n-n_i)} \sum_{k=1}^{c} \sum_{j=1}^{n_k} Y_{jk} \mathbf{1}(X_{jk} \neq i),$$

i.e., the sample mean of Y over all values of X other than $X = i$ ($\bar{i}$ is taken to be the complement of i), while the frequency estimator of $E(Y)$ (the 'overall' mean) is

$$y_{\cdot} = \frac{1}{n}\sum_{k=1}^{c}\sum_{j=1}^{n_k} Y_{jk} = \frac{n_i y_i + (n-n_i)y_{\bar{i}}}{n}.$$

The kernel estimator of μ_i is defined as

$$y_{i,\lambda} = \frac{n^{-1}\sum_{k=1}^{c}\sum_{j=1}^{n_k} Y_{jk}L(X_{jk},i,\lambda)}{p_{i,\lambda}}.$$

See Ouyang et al. (2008, in press) for the theoretical underpinnings of this estimator.

In order to facilitate a comparison of the Bayesian approach of Lindley & Smith (1972) and the kernel approach, we wish to express $y_{i,\lambda}$ as a weighted average of y_i and $y_{\cdot}$. The kernel estimator $y_{i,\lambda}$ can be rewritten as follows,

$$\begin{aligned}
y_{i,\lambda} &= \frac{n^{-1}\sum_{k=1}^{c}\sum_{j=1}^{n_k} Y_{jk}L(X_{jk},i,\lambda)}{p_{i,\lambda}} \\
&= \frac{n^{-1}\left(n_i y_i(1-\lambda) + (n-n_i)y_{\bar{i}}\lambda/(c-1)\right)}{n^{-1}\left(n_i(1-\lambda) + (n-n_i)\lambda/(c-1)\right)} \\
&= \frac{n_i y_i(1-\lambda) + \left(ny_{\cdot} - n_i y_i\right)\lambda/(c-1)}{n_i(1-\lambda) + (n-n_i)\lambda/(c-1)} \\
&= \left[\frac{n_i/n\left(1-\lambda c/(c-1)\right)}{n_i/n\left(1-\lambda c/(c-1)\right) + \lambda/(c-1)}\right] y_i + \left[\frac{\lambda/(c-1)}{n_i/n\left(1-\lambda c/(c-1)\right) + \lambda/(c-1)}\right] y_{\cdot} \\
&= (1-\Phi_i)y_i + \Phi_i y_{\cdot},
\end{aligned}$$

where the third equality follows from (2) by noting that

$$ny_{\cdot} - n_i y_i = (n-n_i)y_{\bar{i}},$$

where

$$1-\Phi_i = \left[\frac{n_i/n\left(1-\lambda c/(c-1)\right)}{n_i/n\left(1-\lambda c/(c-1)\right) + \lambda/(c-1)}\right] \quad \text{and} \quad \Phi_i = \left[\frac{\lambda/(c-1)}{n_i/n\left(1-\lambda c/(c-1)\right) + \lambda/(c-1)}\right],$$

and where $\lambda \in [0,(c-1)/c]$ implies that $\Phi_i \in [0,1]$.

When $\lambda = 0$ (i.e., $\Phi_i = 0 \forall i$), $y_{i,\lambda} = y_i$ (the frequency estimator), while when $\lambda = (c-1)/c$ (i.e., $(1 - \lambda c/(c-1)) = 0$ or $\Phi_i = 1 \forall i$), $y_{i,\lambda} = y_{\cdot}, i = 1, \ldots, c$ (the global mean).

3. Bayes Estimates for the Linear Model

We consider hierarchical models of the form

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

where n_i is the number of observations drawn from group i, and where there exist c groups.

For the ith group,

$$\begin{pmatrix} y_{1i} \\ \vdots \\ y_{n_i i} \end{pmatrix} = \iota_{n_i} \mu_i + \epsilon_i, \quad i = 1, \ldots, c,$$

where ι_{n_i} is a vector of ones of length n_i, $\epsilon_i = (\epsilon_{1i}, \ldots, \epsilon_{n_i i})'$, and, for the sample, $\mathbf{y} = A\mu + \epsilon$ where $\mathbf{y}$ is the n-vector of observations, A is the $(n \times c)$ design matrix, and $\mu = (\mu_1, \ldots, \mu_c)'$, the vector of group means.

Our aim is to understand the connection between hierarchical Bayes models and kernel estimators of multivariate means. Just like an important special case of the Bayes estimates we consider below, the kernel estimator $y_{i,\lambda}$ is a weighted average of the group mean for group i, i.e., y_i, and the overall mean for all groups, i.e., $y_{\cdot}$. The weights themselves are a function of the total number of observations, n, the number of observations in group i, n_i, and the smoothing parameter, λ, which is typically of order $O(n^{-1/2})$.

3.1. **Comparing Kernel and Bayes Estimates.**

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

$$\mathbf{y} \sim (A_1 \theta_1, C_1).$$

As a function of θ_1 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,

$$\theta_1 \sim (A_2\theta_2, C_2),$$

can be regarded as a prior distribution for θ_1 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 θ_1 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 means θ_1. 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, Wong & Bryk (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

$$A_1 = \{a_{ji}\} \text{ with } a_{ji} \in \{0,1\}, \sum_{i=1}^{c} a_{ki} = 1, \sum_{k=1}^{n} a_{ki} = n_i,$$

$$\theta_1 = \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_c \end{bmatrix},$$

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

A_1 is the $n \times c$ design matrix with $A_1'A_1$ the $c \times c$ diagonal matrix with n_i, the number of observations in the ith group, as the ith diagonal element, μ is a $c \times 1$ vector of (population) group means, σ^2 is the within-group variance (i.e., $\text{var}(y_{ij})$), and I_n is the $n \times n$ identity matrix. Next, the second stage will become

$$A_2 = \iota_c,$$

$$\theta_2 = \mu_{\cdot},$$

$$C_2 = \tau^2 I_c,$$

where $\mu_{\cdot}$ is the (population) 'overall mean', and $\tau^2 = \text{var}(\mu_i)$. Note that $A_2\theta_2 = \iota\mu_{\cdot}$ is simply a $c \times 1$ vector with elements being the overall mean $\mu_{\cdot}$ to which the Bayes (and kernel) estimators can shrink. Finally, we let the scalar

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

so that the prior on $\mu_{\cdot}$ is improper. Note that the impropriety is confined to one dimension. The frequency analysis corresponds to an improper prior on the c-vector θ_1, 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 to one dimension. The results are seen below.

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

$$\theta_1^* = D_0 d_0$$

where

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

θ_1^* is the posterior mean and is an optimal estimator under quadratic loss. Writing

$$\Lambda = A_1' C_1^{-1} A_1 = \frac{1}{\sigma^2} \begin{bmatrix} n_1 & 0 & 0 & \dots \\ 0 & n_2 & 0 & \dots \\ \vdots & & \ddots & \\ \vdots & 0 & 0 & n_c \end{bmatrix}$$

we see that

$$\begin{aligned} D_0^{-1} &= \left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota (\iota'^{-2} \iota)^{-1} \iota'^{-2} \right) \\ &= \left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota \iota' / c \right), \\ d_0 &= A_1' C_1^{-1} \mathbf{y} \\ &= \begin{pmatrix} \frac{y_1 n_1}{\sigma^2} \\ \vdots \\ \frac{y_c n_c}{\sigma^2} \end{pmatrix}. \end{aligned}$$

Recall that y_i is the mean for group i. Thus the vector of posterior means satisfies

$$\left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota (\iota' \iota)^{-1} \iota' \right) \theta_1^* = d_0$$

or, element-wise

$$(\sigma^{-2}n_j + \tau^{-2})\theta^*_{1j} - \tau^{-2}\theta^*_{1.} = \sigma^{-2}n_j y_j,$$

where $\theta^*_{1.} = \sum_{j=1}^{c} \theta^*_{1j}/c$. Thus

$$\theta^*_{1j} = (\sigma^{-2}n_j y_j + \tau^{-2}\theta^*_{1.})/(\sigma^{-2}n_j + \tau^{-2})$$

and the Bayes estimator for the jth mean is a weighted average of the group mean and the overall posterior mean. This cannot in general be expressed as a weighted average of the group mean and the overall mean. We explore the implications of this fact below.

We adopt a partitioned inverse, namely

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

Letting

$$A = \Lambda + \tau^{-2}I_c,$$

$$B = \iota,$$

$$D = -\tau^{-2}/c,$$

we have

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

We let $w_i = n_i/\sigma^2$, $d_i = n_i/\sigma^2 + \tau^{-2} = w_i + \tau^{-2}$. Note that

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

Note also that

$$d_0 = \Lambda y = \begin{pmatrix} w_1 y_i \\ \vdots \\ w_c y_c \end{pmatrix}.$$

Next, the Bayes estimator of the ith component of μ (i.e., the ith component of θ_1 in Lindley & Smith's (1972) notation) is given by

$$\mu_i^* = d_i^{-1} w_i y_i - \gamma d_i^{-1} \sum_{j=1}^{c} \frac{w_j y_j}{d_j}.$$

It is useful to recast this expression in terms of the between-to-within variance ratio $\kappa = \tau^2/\sigma^2$. Let $v_i = n_i/(n_i + \kappa^{-1})$. Then

$$\mu_i^* = v_i y_i + (1 - v_i) \sum_{j=1}^{c} v_j y_j / \sum_{j=1}^{c} v_j.$$

Except in a special case, this cannot be expressed as a weighted average of the group and overall mean. The reason is that the different group mean estimators have different precisions, since the prior variance is the same for each group mean but the data contribution depends on the group sample size. Naturally, the overall mean that should be used weights the different group means according to their precisions, and these differ nonlinearly in group sample sizes since the precision depends on the sum of the data and prior precisions. However, some insight can be gained by considering the important special case of a balanced design.

3.2. **The Balanced Case (n_i equal for all i).** Let $n_i = n^*$ for all i. The kernel estimator of the ith component of μ can be written as

$$\begin{aligned} y_{i,\lambda} &= \left[\frac{n^* \left(1 - \lambda c/(c-1)\right)}{n^* \left(1 - \lambda c/(c-1)\right) + n\lambda/(c-1)} \right] y_i + \left[\frac{n^* \lambda/(c-1)}{n^* \left(1 - \lambda c/(c-1)\right) + n^* \lambda/(c-1)} \right] y_{\cdot} \\ &= \left[\frac{n^*}{n^* + n^*/((c-1)/\lambda - c)} \right] y_i + \left[\frac{n^*/((c-1)/\lambda - c)}{n^* + n^*/((c-1)/\lambda - c)} \right] y_{\cdot}, \end{aligned} \tag{3}$$

where λ is a smoothing parameter to be set by the researcher.

Further, the Bayes estimator of the ith component of μ is given by (in the balanced case)

$$\begin{aligned}\mu_i^* &= \left[\frac{n^*}{n^* + \kappa^{-1}}\right] y_i + \left[\frac{\kappa^{-1}}{n^* + \kappa^{-1}}\right] y_{\cdot} \\ &= v y_i + (1 - v) y_{\cdot}\end{aligned} \tag{4}$$

where $v = n^*/(n^* + \kappa^{-1})$ is the common value of the v_i term from above. The correspondence between the two methods is given by

$$n^*/((c-1)/\lambda - c) = \kappa^{-1},$$

hence

$$\kappa = \frac{1}{n^*}((c-1)/\lambda - c).$$

Alternatively, λ can be expressed as

$$\lambda = (c-1)/(c + n^*\kappa). \tag{5}$$

This 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 κ or τ^2) and smaller as the groups are thought to be less similar. 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.

From a decision-theoretic point of view we can consider the admissibility of the frequency estimator (2), the kernel estimator (3), and the equivalent Bayes estimator (4). Consider the normal case with squared-error loss and note that the estimators are linear; $\mu^* = By$. Using Cohen (1966, Theorem 2.1) we see that μ^* is admissible if and only if the eigenvalues of B, b_i, satisfy $0 \leq b_i \leq 1$ with equality at unity at most twice. Here B has diagonal elements

$$\{B\}_{ii} = v + \frac{1 - v}{c}$$

and off-diagonal elements

$$\{B\}_{ij} = \frac{1-v}{c}$$

and the eigenvalues are v with multiplicity $c-1$ and unity with multiplicity one. In the unsmoothed case ($\tau^{-2} = 0$ from the Bayesian viewpoint, $\lambda = 0$ from the frequentist), all of the eigenvalues are unity and the estimators are inadmissible for $c > 2$.

Next, we turn to another frequency property, that of MSE. This is of limited interest from the Bayesian point of view (samples that did not arise are irrelevant for a particular application) but is useful in assessing properties of techniques used repeatedly in identical applications. We know that the MSE of the Bayes/kernel estimator (identical in the balanced case) improves over that of the frequency estimator y_i if and only if (Lindley & Smith (1972, page 3, Equation (2)))

$$\hat{\tau}^2 \leq 2\tau^2 + \sigma^2,$$

where

$$\hat{\tau}^2 = \sum_i \frac{(y_i - y_\cdot)^2}{c-1}. \tag{6}$$

This allows us to obtain an upper bound for λ that will ensure (in probability) that $\text{MSE}(y_{i,\lambda}) \leq \text{MSE}(y_i)$. Substituting, we have

$$\hat{\tau}^2 \leq 2\frac{\sigma^2}{n}((c-1)/\lambda - c) + \sigma^2,$$

which is equivalent to

$$\frac{n(\hat{\tau}^2 - \sigma^2)}{2\sigma^2} + c \leq \frac{c-1}{\lambda},$$

which implies that

$$\lambda \leq \frac{2\sigma^2(c-1)}{n(\hat{\tau}^2 - \sigma^2) + 2c\sigma^2}. \tag{7}$$

The only unknown in this formula is σ^2 which can be estimated directly from the data via

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n^*}(y_{ij} - y_i)^2}{n - c}. \tag{8}$$

It is widely known that the smoothing parameter must obey $\lambda \to 0$ as $n \to \infty$ for consistent estimation while, as noted earlier, λ is restricted to lie in $[0, (c-1)/c]$. Note that (7) tells us that an *oversmoothed* kernel estimator can be consistent but can be beaten by the frequency estimator on MSE grounds (i.e., when λ is overly large).

4. Implications for Kernel Estimation

The results obtained in Section 3 above yield a number of implications for applied kernel estimation with discrete data. The first is that they provide bounds for bandwidth selection that are previously unknown in the literature. The second is that they deliver a simple plug-in method of bandwidth selection with an empirical Bayes flavor (Efron & Morris (1973)) that possesses appealing finite-sample properties and, in addition, is computationally trivial.

4.1. **Bounds for λ.**

Recall that $[0, (c-1)/c]$ is the range of λ when using the kernel function defined in (1). We now incorporate the result summarized in (7) to obtain tighter bounds on λ.

Note that when $\hat{\tau}^2 = \sigma^2$, (7) equals $(c-1)/c$, the upper bound possible for λ, hence the bound is non-binding in this case. It is also non-binding when $\hat{\tau}^2 \leq \sigma^2$. However, when $\hat{\tau}^2 > \sigma^2$, then in order to outperform the frequency estimator on MSE grounds, the kernel estimator must obey $\lambda < (c-1)/c$ with the upper bound now given by (7). On MSE grounds, the range of λ is no longer $[0, (c-1)/c]$, rather it is

$$\left[0, \min\left\{\frac{c-1}{c}, \frac{2\sigma^2(c-1)}{n(\hat{\tau}^2 - \sigma^2) + 2c\sigma^2}\right\}\right]. \tag{9}$$

In other words, (7) tells us that when the idiosyncratic variation (i.e., $\sigma^2 = \mathrm{var}(y_{ij})$) is greater than the intergroup variation (i.e., $\hat{\tau}^2 = \mathrm{var}(y_i)$), there exists a λ in the feasible

range (i.e. $[0,(c-1)/c]$) that will outperform the frequency estimator on MSE grounds (e.g., that given by (5)). On the other hand, when the idiosyncratic variation is less than the intergroup variation, imposing this (reduced) bound on λ (rather than $(c-1)/c$) avoids situations where the frequency estimator may outperform the smoothed estimator. Note that (5) always satisfies the bound.

The reader may well be asking what effect this may have in applied settings. By way of example, we consider two illustrative cases and present the results in the form of two graphs given in Figure 1. In Figure 1 below we plot the upper bound on λ given by the above rule as a function of $\hat{\kappa} = \hat{\tau}^2/\sigma^2$ for $c = \{2, 10\}$ and $n = 25$ i.e., we plot the upper bound in (9) λ versus the relative variation in the group means ($\hat{\tau}^2$).

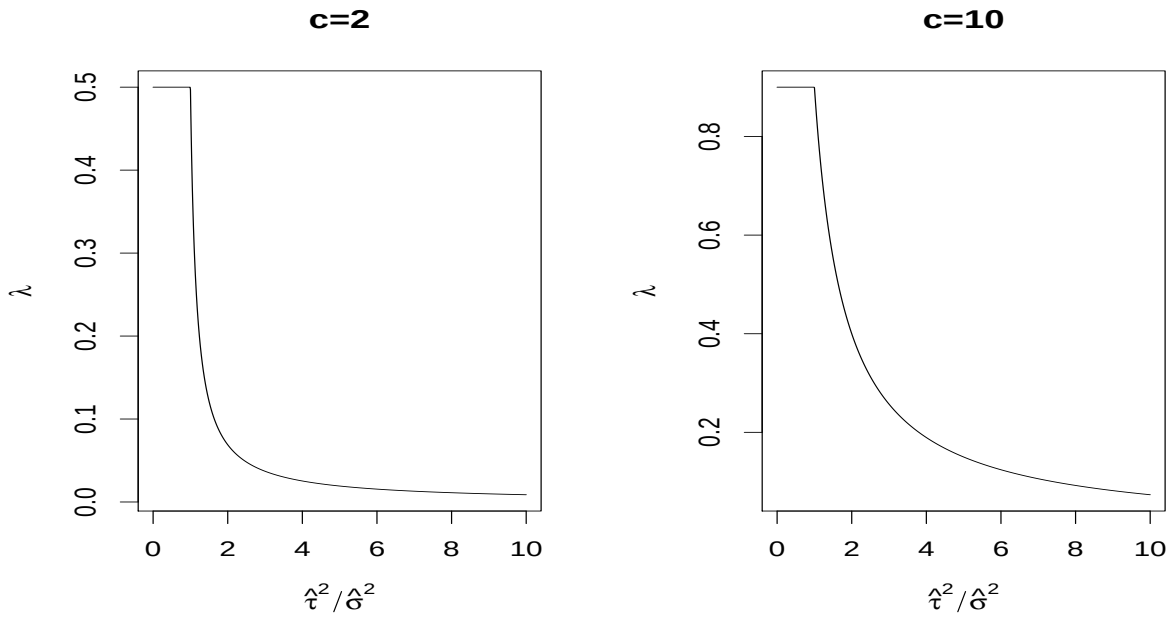

FIGURE 1. Upper bounds on λ given by Equation (9) when $\sigma^2 = 1$, $n = 25$, $c = \{2, 10\}$.

Figure 1 reveals that there are situations in which choosing λ in the permissible range ($[0,(c-1)/c]$) can result in smoothed estimates that are worse than the frequency (i.e.,

non-smooth) estimate on MSE grounds. In these situations (i.e., when $\hat{\tau}^2 > \sigma^2$) a restricted choice of λ can avoid this possibility.

4.2. **A Plug-In Bandwidth Selector.** Equation (5) suggests a computationally trivial formula for a plug-in bandwidth selector for the kernel estimator of a multivariate mean. By way of example, we compare the MSE performance of the frequency estimator ($\lambda = 0$), least-squares cross-validated bandwidth selection (Ouyang et al. (2008, in press)), and that based upon (5) evaluated using the estimators (6) and (8) of τ^2 and σ^2. We vary τ and σ, set $c = 2$, $n_i = 25$, and draw $M = 10,000$ Monte Carlo replications where the setup is that described in Section 3. For each replication we compute the MSE. Results are summarized in figures 2 and 3 via box-and-whisker plots. The median MSE over the M replications is given below each figure.

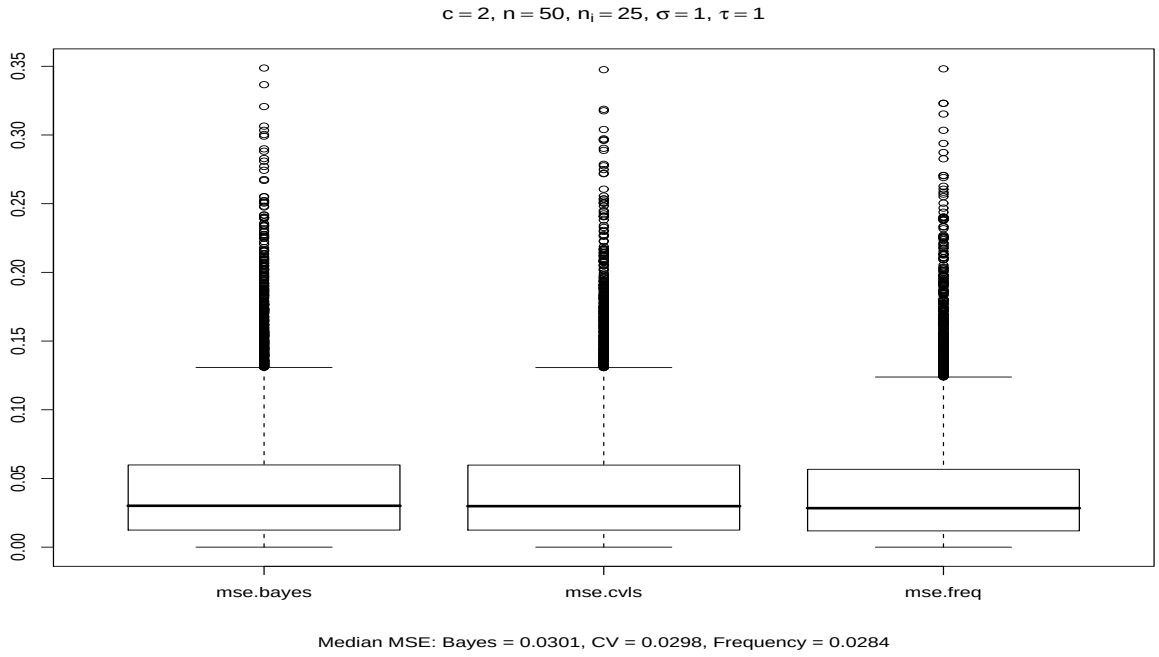

FIGURE 2. Boxplots for the MSE of the Bayes-plug-in, cross-validated, and frequency estimators $\sigma = 1$, $\tau = 1$. Note that results for $\tau > \sigma$ are qualitatively identical to those for $\tau = \sigma$ hence are omitted for space considerations.

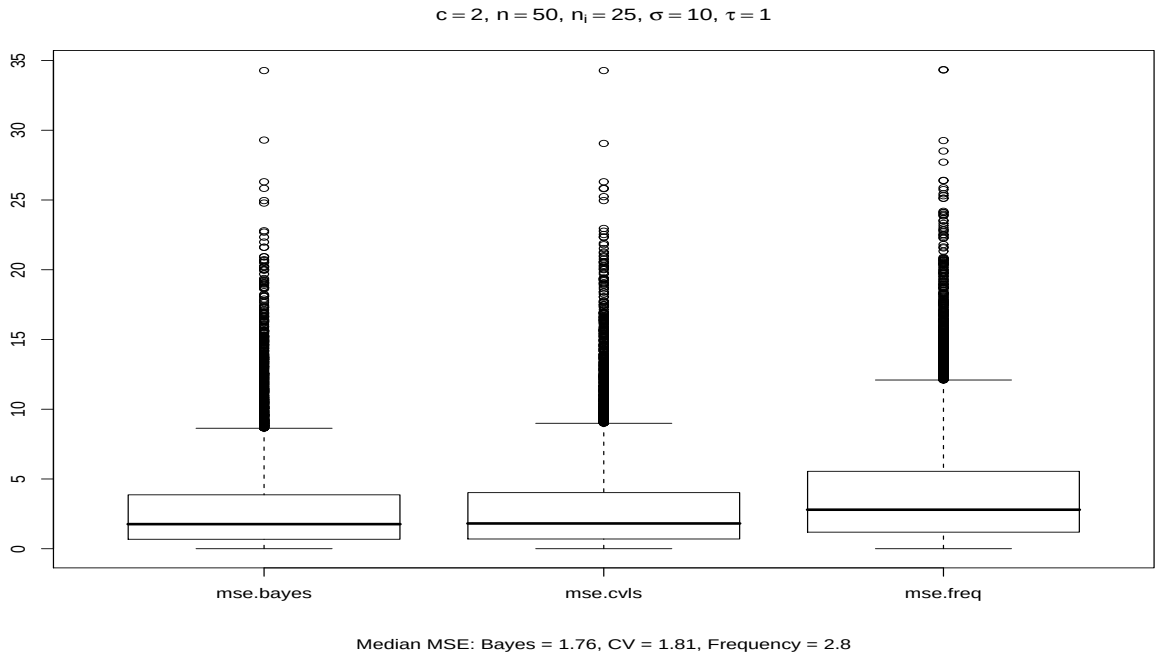

FIGURE 3. Boxplots for the MSE of the Bayes-plug-in, cross-validated, and frequency estimators $\sigma = 10$, $\tau = 1$.

It can be seen that the relative performances of the frequency estimator ($\lambda = 0$), the least-squares cross-validated estimator, and that based on the Bayes-plug-in rule (5) are equivalent for $\tau \geq \sigma$. However, for $\tau < \sigma$, the Bayes-plug-in rule remains competitive with the least-squares rule and outperforms the frequency estimator. Given that the Bayes-plug-in bandwidth is trivial to compute, it may to be of interest to practitioners.

5. Conclusion

In this paper we investigate the relationship between the kernel smoothing of a multivariate mean and the Bayes estimate thereof. We show that the smoothing parameter adopted for the kernel estimator is related to the prior variance in a three stage hierarchical Bayes model, which then provides an upper bound on the degree of smoothing that can be applied in order for the kernel method to improve upon the frequency (i.e., non-smooth) estimator. To the best of our knowledge these bounds are previously unknown in the literature. We also

propose a Bayes-plug-in bandwidth for kernel estimation that is computationally trivial and possesses appealing finite-sample properties.

Many remain uncomfortable with the kernel smoothing of discrete data and, in particular, with the kernel smoothing of datasets consisting of both discrete and continuous data. For instance, it is common to encounter separate kernel estimates of earnings equations for different industries where industry grouping is determined by, say, Standard Industrial Classification (SIC) codes, which is clearly a frequency approach (i.e., separate kernel estimates are generated for each SIC code). Methods for the kernel estimation of unconditional distributions, conditional distributions, and conditional means that smooth the discrete covariate in the manner described above in the presence of both discrete and continuous data have recently been developed; see Li & Racine (2003), Hall, Li & Racine (2004), Racine & Li (2004), and also Li & Racine (2007). In finite-sample settings, the estimators that smooth the discrete covariates often outperform their frequency-based counterparts on MSE grounds. There are, however, no finite-sample results that indicate when this will be the case, and it would be helpful to have some guidance on this matter. We expect that the approach considered herein can be extended to this setting providing enhanced understanding of kernel smoothing in these settings along with bounds on bandwidths for discrete covariates thereby ensuring that the kernel estimator that smooths the discrete covariates dominates the frequency-based kernel estimator that does not.

References

Aitchison, J. & Aitken, C. G. G. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**(3), 413–420.

Cohen, A. (1966), 'All admissible linear estimates of the mean vector', *The Annals of Mathematical Statistics* **37**, 458–463.

Efron, B. & Morris, C. (1973), 'Stein's estimation rule and its competitors - an empirical Bayes approach', *Journal of the American Statistical Association* **68**(341), 117–130.

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

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

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. & Racine, J. S. (2003), 'Nonparametric estimation of distributions with categorical and continuous data', *Journal of Multivariate Analysis* **86**, 266–292.

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

Ouyang, D., Li, Q. & Racine, J. S. (2008, in press), 'Nonparametric estimation of regression functions with discrete regressors', *Econometric Theory* .

Racine, J. S. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99–130.

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.

Nicholas M. Kiefer: Department 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.