Empirical Bayes Regression With Many Regressors

Thomas Knox

Graduate School of Business, University of Chicago

James H. Stock

Department of Economics, Harvard University

and

Mark W. Watson*

Department of Economics and Woodrow Wilson School, Princeton University

This revision: January 2004

*The authors thank Gary Chamberlain, Sid Chib, Ron Gallant, Carl Morris, and Whitney Newey for useful comments, and Josh Angrist, Alan Krueger, and Doug Staiger for supplying the data used in this paper. This research was supported in part by National Science Foundation grants SBR-9730489 and SBR-0214131.

ABSTRACT

We consider frequentist and empirical Bayes estimation of linear regression coefficients with T observations and K orthonormal regressors. The frequentist formulation considers estimators that are equivariant under permutations of the regressors. The empirical Bayes formulation (both parametric and nonparametric) treats the coefficients as i.i.d. and estimates their prior. Asymptotically, when $K = \rho T^{\delta}$ for $0 < \rho < 1$ and $0 < \delta \leq 1$, the empirical Bayes estimator is shown to be: (i) optimal in Robbins' (1955, 1964) sense; (ii) the minimum risk equivariant estimator; and (iii) minimax in both the frequentist and Bayesian problems over a wide class of error distributions. Also, the asymptotic frequentist risk of the minimum risk equivariant estimator is shown to equal the Bayes risk of the (infeasible subjectivist) Bayes estimator in the Gaussian model with a "prior" that is the weak limit of the empirical c.d.f. of the true parameter values.

Key Words: Large model regression, equivariant estimation, minimax estimation, shrinkage estimation

1. Introduction

This paper considers the estimation of the coefficients of a linear regression model with dependent variable y and a large number (K) of orthonormal regressors X under a quadratic loss function. When K is large, this and the related K-mean problem have received much attention ever since Stein (1955) showed that the ordinary least squares (OLS) estimator is inadmissible for $K \geq 3$. Many approaches have been proposed, including model selection, Bayesian and otherwise [e.g. George (1999)], Bayesian model averaging [e.g. Hoeting, Madigan, Raftery and Volinsky (1999)], shrinkage estimation, ridge regression, and dimension-reduction schemes such as factor models [e.g. Stock and Watson (2002)]. However, outside of a subjectivist Bayesian framework, where the optimal estimator is the posterior mean, estimators with attractive optimality properties have been elusive.

We consider this problem using both frequentist and Bayesian risk concepts. Our frequentist approach exploits a natural permutation equivariance in this problem. Suppose for the moment that the regression errors are i.i.d. normally distributed, that they are independent of the regressors, and that the regressor and error distributions do not depend on the regression parameters; this shall henceforth be referred to as the "Gaussian model." In the Gaussian model, the likelihood does not depend on the ordering of the regressors, that is, the likelihood is invariant to simultaneous permutations of the indices of the regressors and their coefficients. Moreover, in this model with known error variance, the OLS estimator is sufficient for the regression coefficients. These two observations lead us to consider the class of "Gaussian equivariant estimators," that is, estimators that are equivariant functions of the OLS estimator under permutations of the regressor indices. This class is large and contains common estimators in this problem, including OLS, OLS with information criterion selection, ridge regression, the James-Stein (1960) estimator, and common shrinkage estimators. If it exists, the estimator that minimizes expected quadratic loss among all equivariant estimators is the minimum risk equivariant estimator. If this estimator achieves the minimum risk uniformly for all values of the regression coefficients in an arbitrary closed ball around the origin, the estimator is uniformly minimum risk equivariant over this ball.

The Bayesian approach starts from the perspective of a subjectivist Bayesian and models the coefficients as i.i.d. draws from some subjective prior distribution G. Under quadratic loss, the Bayes estimator is the posterior mean which, in the Gaussian model with known error variance, depends only on the OLS estimators and the prior. The Gaussian empirical Bayes estimator is this subjectivist Bayes estimator, constructed using an estimate of G and the regression error variance.

We analyze these estimators under an asymptotic nesting in which $K = \rho T$ (where $0<\rho<1$). So the R^2 of the regression does not approach one, the true coefficients are modeled as being in a $T^{-1/2}$ neighborhood of zero. Under this nesting, the estimation risk, both frequentist and Bayesian, has a O(1) asymptotic limit. In some applied settings, both T and K are large, but K/T is small. For example, Section 6 considers a prediction problem using the well-known Angrist-Krueger (1991) data set in which T = 329,509 and K = 178. To accommodate such empirical situations we further consider the nesting $K = \rho T^{\delta}$, $0<\delta\leq 1$, and analyze the relevant risk functions scaled by (T/K).

This paper makes three main contributions. The first concerns the Bayes risk. In the Gaussian model, we show that a Gaussian empirical Bayes estimator asymptotically achieves the same Bayes risk as the subjectivist Bayes estimator, which treats G as known. This is shown both in a nonparametric framework, in which G is treated as an infinite-dimensional nuisance parameter, and in a parametric framework, in which G is known up to a finite dimensional parameter vector. Thus this Gaussian empirical Bayes estimator is asymptotically optimal in the Gaussian model in the sense of Robbins (1964), and the Gaussian empirical Bayes estimator is admissible asymptotically. Moreover, the same Bayes risk is attained under weaker, non-Gaussian assumptions on the distribution of the error term and regressors. Thus, the Gaussian empirical Bayes estimator is minimax (as measured by the Bayes risk) against a large class of distributional deviations from the assumptions of the Gaussian model.

The second contribution concerns the frequentist risk. In the Gaussian model, the Gaussian empirical Bayes estimator is shown to be asymptotically the uniformly minimum risk equivariant estimator. Moreover, the same frequentist risk is attained under weaker, non-Gaussian assumptions. Thus, the Gaussian empirical Bayes estimator is minimax (as

measured by the frequentist risk) among equivariant estimators against these deviations from the Gaussian model.

Third, because the same estimator solves both the Bayes and the frequentist problems, it makes sense that the problems themselves are asymptotically related. We show that this is so. In the Gaussian model, the limiting frequentist risk of permutation-equivariant estimators and the limiting Bayes risk are shown to share a lower bound which is the risk of the subjectivist Bayes estimator constructed using a "prior" that equals the limiting empirical distribution of the true regression coefficients. This bound is achieved asymptotically by the empirical Bayes estimators laid out in this paper. The empirical Bayes estimators use the large number of estimated regression coefficients to estimate this "prior." These results differ in an important way from the usual asymptotic analysis of Bayes estimators in finite dimensional settings, in which the likelihood dominates the prior distribution. Here the number of parameters can grow proportionally to the sample size so that the prior affects the posterior asymptotically.

This paper also makes several contributions within the context of the empirical Bayes literature. Although we do not have repeated experiments, under our asymptotic nesting in the Gaussian model the regression problem becomes formally similar to the Gaussian compound decision problem. Also, results for the compound decision problem are extended to the non-Gaussian model by exploiting Berry-Esseen type results for the regression coefficients; this leads to our minimax results. Finally, permutation arguments are used to extend an insight of Edelman (1988) in the Gaussian compound decision problem to show that the empirical Bayes estimator is also minimum risk equivariant. As far as we know, the work closest to the present paper is George and Foster (2000), who consider an empirical Bayes approach to variable selection. However, their setup is fully parametric and their results refer to model selection, a different objective than ours.

The remainder of the paper is organized as follows. Section 2 presents the Gaussian model, the risk functions, and estimators. Section 3 presents the asymptotic efficiency results for the Gaussian model. Section 4 extends the analysis of these Gaussian estimators to (a) non-Gaussian regression errors and (b) the presence of a small number of "base" regressors with non-local (fixed) coefficients. Section 5 investigates the finite-sample

efficiency of the estimators using a Monte Carlo experiment, and Section 6 is a brief application to the Angrist-Krueger (1991) data set.

2. Risk and Estimators in the Gaussian Model

2.1 Model, Risk and Asymptotic Nesting

Let y denote the $T\times1$ vector of observations on the regressand and let X denote the $T\times K$ matrix of observations on the regressors. In this section we consider the Gaussian regression model,

$$y = X\beta + \varepsilon,\ \varepsilon|X \sim \mathrm{N}(0, \sigma_\varepsilon^2 I_T),\ \sigma_\varepsilon^2 > 0, \tag{2.1}$$

where I_T is the $T\times T$ identity matrix. We assume that the regressors have been transformed so that they are orthonormal:

Assumption 1: $T^{-1}X'X = I_K$

We consider the estimation of β under the (frequentist) risk function $\mathrm{tr}(V_{\tilde\beta})$, where $V_{\tilde\beta} = E(\tilde\beta - \beta)(\tilde\beta - \beta)'$, $\tilde\beta$ is an estimator of β, and the expectation is taken conditional on the value of β.

We adopt an asymptotic nesting that formalizes the notion that the number of regressors is large, specifically, that $K = \rho T$, $0 < \rho < 1$, and all limits are taken as $T \to \infty$. (This is generalized in section 4 to allow $K = \rho T^\delta$ for $0 < \delta \le 1$.) Under the nesting $K = \rho T$, if β is nonzero and fixed, the population R^2 tends to one, which is unrepresentative of the empirical problems of interest. We therefore adopt a nesting in which each regressor makes a small but potentially nonzero contribution, specifically we adopt the local parameterization

$$\beta = b/\sqrt{T} \text{ and } \tilde\beta = \tilde b/\sqrt{T}, \tag{2.2}$$

where $\{b_i\}$ is fixed as $T \to \infty$. Because K and T are linked, various objects are doubly indexed arrays, and b and its estimators are sequences indexed by K, but to simplify notation this indexing is usually suppressed.

Using this change of variables, the frequentist risk $\mathrm{tr}(V_{\tilde{\beta}})$ becomes

$$R(b, \tilde{b}) = \left(\frac{K}{T}\right) K^{-1} \sum_{i=1}^{K} E(\tilde{b}_i - b_i)^2 , \tag{2.3}$$

where b_i is the i^{th} element of b, etc.

2.3 OLS and Bayes Estimators in the Gaussian Model

Using the change of variables (2.2) and the orthonormality assumption, the OLS estimators of b and σ_ε^2 are

$$\hat{b} = T^{-1/2} \sum_{t=1}^{T} X_t y_t \text{ and } \hat{\sigma}_\varepsilon^2 = \frac{1}{T-K} \sum_{t=1}^{T} (y_t - \hat{\beta}' X_t)^2 . \tag{2.4}$$

In the Gaussian model, $\hat{b} - b \sim N(0, \sigma_\varepsilon^2 I_k)$ and $(T-K)\hat{\sigma}_\varepsilon^2 / \sigma_\varepsilon^2 \sim \chi^2_{T-K}$.

If the distribution of $\{X_t\}$ does not depend on $(b, \sigma_\varepsilon^2)$, then $(\hat{b}, \hat{\sigma}_\varepsilon^2)$ are sufficient for $(b, \sigma_\varepsilon^2)$, and the Rao-Blackwell theorem implies that we may restrict attention to estimators $\tilde{b}$ that are functions of $(\hat{b}, \hat{\sigma}_\varepsilon^2)$. Let $\phi_K(x; \sigma_\varepsilon^2) = \prod_{i=1}^{K} \phi(x_i; \sigma_\varepsilon^2)$, where $\phi(\bullet; \sigma_\varepsilon^2)$ is the $N(0, \sigma_\varepsilon^2)$ density; the density of $\hat{b}$ is $\phi_K(\hat{b} - b; \sigma_\varepsilon^2)$.

In the Bayesian formulation, we model $\{b_i\}$ as i.i.d. draws from the prior distribution G. We suppose that the subjectivist Bayesian knows σ_ε^2. (One motivation for this simplification is that $\hat{\sigma}_\varepsilon^2$ is L_2-consistent for σ_ε^2, so that a proper prior for σ_ε^2 with support on $(0, \infty)$ would be dominated by the information in $\hat{\sigma}_\varepsilon^2$.) Accordingly, the Bayes risk is the frequentist risk (2.3), integrated against the prior distribution G. Setting $K/T = \rho$ and $G_K(b) = G(b_1) \cdots G(b_K)$, the Bayes risk is

(2.5) $$r_G(\tilde{b}) = \int R(b,\tilde{b})dG_K(b) = \rho \int K^{-1}\sum_{i=1}^{K}\int (\tilde{b}_i(\hat{b}) - b_i)^2 \phi_K(\hat{b}-b)\mathrm{d}\hat{b}\mathrm{d}G_K(b).$$

Because loss is quadratic, the Bayes risk (2.5) is minimized by the "normal Bayes" estimator,

(2.6) $$\hat{b}^{NB} = \frac{\int x\phi_K(\hat{b}-x;\sigma_\varepsilon^2)\mathrm{d}G_K(x)}{\int \phi_K(\hat{b}-x;\sigma_\varepsilon^2)\mathrm{d}G_K(x)}.$$

Because the likelihood is Gaussian, $\hat{b}^{NB}$ can be rewritten as a function of the score of the marginal distribution of $\hat{b}$ (see for example Maritz and Lwin [1989, p. 73]). Let m_K denote the marginal distribution of $(\hat{b}_1,...,\hat{b}_K)$, $m_K(x;\sigma_\varepsilon^2) = \int \phi_K(x-b;\sigma_\varepsilon^2)dG_K(b)$, and let $\ell_K(x;\sigma_\varepsilon^2) = m_K'(x;\sigma_\varepsilon^2)/m_K(x;\sigma_\varepsilon^2)$ be its score; then $\hat{b}^{NB}$ can be written as

(2.7) $$\hat{b}^{NB} = \hat{b} + \sigma_\varepsilon^2 \ell_K(\hat{b};\sigma_\varepsilon^2).$$

The Gaussian empirical Bayes estimators studied here are motivated by (2.7). In the empirical Bayes approach to this problem, G is unknown, as is σ_ε^2. Thus the score ℓ_K is unknown, and the estimator (2.7) is infeasible. However, both the score and σ_ε^2 can be estimated. The resulting estimator is referred to as the simple Gaussian empirical Bayes estimator ("simple" because G does not appear explicitly in (2.7)). Throughout, σ_ε^2 is estimated by $\hat{\sigma}_\varepsilon^2$, the usual degrees-of-freedom-adjusted OLS estimator. Both parametric and nonparametric approaches to estimating the score are considered. These respectively yield parametric and nonparametric empirical Bayes estimators.

Parametric Gaussian empirical Bayes estimator. The parametric Gaussian empirical Bayes (PEB) estimator is based on adopting a parametric specification for the distribution of b, which we denote by $G_K(b;\theta)$, where θ is a finite-dimensional parameter

vector. The marginal distribution of $\hat{b}$ is then $m_K(x;\theta,\sigma_\varepsilon^2) = \int \phi_K(x-b;\sigma_\varepsilon^2)dG_K(b;\theta)$. The PEB estimator is computed by substituting estimates of σ_ε^2 and θ into $m_K(x;\theta,\sigma_\varepsilon^2)$ using this parametrically estimated marginal and its derivative to estimate the score, and substituting this parametric score estimator into (2.7). The specific parametric score estimator used here is,

$$\hat{\ell}_K(x;\hat{\theta},\hat{\sigma}_\varepsilon^2) = \frac{m_K'(x;\hat{\theta},\hat{\sigma}_\varepsilon^2)}{m_K(x;\hat{\theta},\hat{\sigma}_\varepsilon^2)+s_K}, \tag{2.8}$$

where $\{s_K\}$ is a sequence of small positive numbers such that $s_K \to 0$. (The sequence $\{s_K\}$, specified below, is a technical device used in the proof to control the rate of convergence.)

The parametric Gaussian empirical Bayes estimator, $\hat{b}^{PEB}$, is obtained by combining (2.7) and (2.8) and using $\hat{\sigma}_\varepsilon^2$,

$$\hat{b}^{PEB} = \hat{b} + \hat{\sigma}_\varepsilon^2 \hat{\ell}_K(\hat{b};\hat{\theta},\hat{\sigma}_\varepsilon^2). \tag{2.9}$$

Nonparametric Gaussian simple empirical Bayes estimator. The nonparametric Gaussian simple empirical Bayes (NSEB) estimator estimates the score without assuming a parametric form for *G*. The score estimator used for the theoretical results uses a construction similar to Bickel et. al. (1993) and van der Vaart (1988). Let $w(z)$ be a symmetric bounded kernel with $\int z^4 w(z)dz < \infty$ and with bounded derivative $w'(z) = dw(z)/dz$, and let h_K denote the kernel bandwidth sequence. Define

$$\hat{m}_{iK}(x) = \frac{1}{(K-1)h_K}\sum_{j\neq i} w\left(\frac{\hat{b}_j - x}{h_K}\right) \tag{2.10}$$

$$\hat{m}_{iK}'(x) = -\frac{1}{(K-1)h_K^2}\sum_{j\neq i} w'\left(\frac{\hat{b}_j - x}{h_K}\right), \text{ and} \tag{2.11}$$

$$\breve{\ell}_{iK}(x) = \frac{\hat{m}_{iK}'(x)}{\hat{m}_{iK}(x)+s_K}. \tag{2.12}$$

The nonparametric score estimator considered here is,

$$\hat{\ell}_{iK}(x) = \left\{ \begin{array}{l} \breve{\ell}_{iK}(x) \text{ if } |x| < \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K} \text{ and } |\breve{\ell}_{iK}(x)| \le q_K, \\ 0 \text{ otherwise} \end{array} \right\}, \tag{2.13}$$

where $\{q_K\}$ is a nonrandom sequence (specified below) with $q_K \to \infty$.

The nonparametric Gaussian simple empirical Bayes estimator, $\hat{b}^{NSEB}$, obtains by substituting $\hat{\sigma}_\varepsilon^2$ and (2.13) into (2.7); thus,

$$\hat{b}_i^{NSEB} = \hat{b}_i + \hat{\sigma}_\varepsilon^2 \hat{\ell}_{iK}(\hat{b}_i), \; i=1,\ldots, K. \tag{2.14}$$

2.5 Equivariant Estimators in the Gaussian Model

As in the Bayesian case and motivated by sufficiency, we restrict analysis to estimators, $\tilde{b}$, that are functions of the OLS estimators, $\hat{b}$. We further restrict the estimators so that they do not depend on the ordering of the regressors. To motivate this restriction, let $\underline{X}_i$ denote the i^{th} column of X, and note that the value of the likelihood ϕ_K does not change under a simultaneous reordering of the index i on $(\underline{X}_i, b_i)$. More precisely, let **P** denote the permutation operator, so that $\mathbf{P}(\underline{X}_1, \underline{X}_2, ..., \underline{X}_K) = (\underline{X}_{i_1}, \underline{X}_{i_2}, ..., \underline{X}_{i_K})$, where $\{i_j, j=1,\ldots,K\}$ is a permutation of $\{1,\ldots,K\}$. The collection of all such permutations is a group, where the group operation is composition. The induced permutation of the parameters is $\mathbf{P}b$. The likelihood constructed using $(\mathbf{P}X, \mathbf{P}b)$ equals the likelihood constructed using (X, b); that is, the likelihood is invariant to **P**. Since the likelihood does not depend on the ordering of the regressors, we consider estimators that do not depend on the ordering.

Following the theory of equivariant estimation (e.g. Lehmann and Casella (1998, ch. 3)), this leads us to consider the set of estimators that are equivariant under any such permutation. An estimator $\tilde{b}(\hat{b})$ is equivariant under **P** if the permutation of the estimator

constructed using $\hat{b}$ equals the (non-permuted) estimator constructed using the same permutation applied to $\hat{b}$. The set $\mathcal{B}$ of all estimators that are functions of $\hat{b}$ and are equivariant under the permutation group is,

$$\mathcal{B} = \{ \tilde{b}(\hat{b}) : \mathbf{P}[\tilde{b}(\hat{b})] = \tilde{b}(\boldsymbol{P}\hat{b}) \}, \tag{2.15}$$

and we study optimal estimators in this set.

3. Efficiency Results in the Gaussian Model

In this section we present efficiency results for the empirical Bayes and equivariant estimators in the Gaussian model. We begin by listing two properties of the OLS estimators, which are easily derived in the Gaussian model with $K = \rho T$:

$$\mathrm{E}[(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)^2 | b, \sigma_\varepsilon^2] = \frac{2\sigma_\varepsilon^4}{T-K} \to 0, \tag{3.1}$$

$$R(\hat{b}, b) = r_G(\hat{b}) = \rho\sigma_\varepsilon^2. \tag{3.2}$$

3.2 Asymptotic Efficiency of Empirical Bayes Estimators

We start with assumptions on the family of distributions for b. Assumption 2 is used for the nonparametric estimator, and assumption 3 is used for the parametric estimator.

Assumption 2: $\{b_i\}$ are i.i.d. with distribution G and $\mathrm{var}(b_i) = \sigma_b^2 < \infty$.

Assumption 3:

(a) G belongs to a known family of distributions $G(b;\theta)$ indexed by the finite-dimensional parameter θ contained in a compact Euclidean parameter space Θ;

(b) G has density $g(b;\theta)$ that satisfies $\sup_{b,\theta\in\Theta} g(b;\theta) < C$ and $\sup_{b,\theta\in\Theta} \left\| \partial g(b;\theta)/\partial\theta \right\| < C$.

(c) There exists an estimator $\hat{\theta} = \theta(\hat{b}, \hat{\sigma}_\varepsilon^2)$ such that, for all K sufficiently large, $\mathrm{E}[K\left\|\hat{\theta} - \theta\right\|^2] \leq C < \infty$.

The final assumption provides conditions on the rates of the various sequences of constants used to construct the estimators.

Assumption 4: The sequences $\{s_K\}$, $\{q_K\}$, and $\{h_K\}$ satisfy: $s_K \to 0$ and $s_K^2 \log K \to \infty$; $q_K \to \infty$, $K^{-1/2} q_K \to 0$, and $K^{-1/6} q_K \to \infty$; and $h_K \to 0$, $K^{1/24} h_K \log K \to 0$, and $K^{2/9} h_K \to \infty$.

For example, Assumption 4 is satisfied by $s_K = (\log K)^{-1/3}$, $q_K = K^{1/3}$, and $h_K = K^{-1/9}$.

The efficiency properties of the empirical Bayes estimators are summarized in the following theorem.

Theorem 1: In the Gaussian regression model,

(a) given Assumptions 1-4, $r_G(\hat{b}^{PEB}) - r_G(\hat{b}^{NB}) \to 0$; and

(b) given Assumptions 1, 2 and 4, $r_G(\hat{b}^{NSEB}) - r_G(\hat{b}^{NB}) \to 0$.

Theorem 1 states that the Bayes risk of the EB estimators asymptotically equals the Bayes risk of the infeasible estimator, $\hat{b}^{NB}$. Thus the theorem implies that, in the Gaussian model, the empirical Bayes estimators are asymptotically optimal in the sense of Robbins (1964).

3.3 Results for Equivariant Estimators

The next theorem characterizes the asymptotic limit of the frequentist risk of the minimum risk equivariant estimator for the class of equivariant estimators, $\mathcal{B}$, given in (2.15). Let $\tilde{G}_K$ denote the (unknown) empirical c.d.f. of the true coefficients $\{b_i\}$ for fixed K, that is, the one-dimensional distribution assigning point mass of $1/K$ to each b_i. Let

$\hat{b}^{NB}_{\tilde{G}_K}$ denote the normal Bayes estimator constructed using this distribution as a prior, and let $\|x\|_q = (K^{-1}\sum_{i=1}^{K} x_i^q)^{1/q}$ for the K-vector x.

Theorem 2: Given Assumptions 1 and 4,

(a) $\inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b}) \geq R(b,\hat{b}^{NB}_{\tilde{G}_K}) = r_{\tilde{G}_K}(\hat{b}^{NB}_{\tilde{G}_K})$ for all $b \in \mathbb{R}^K$ and for all K; and

(b) $\sup_{\|b\|_2 \leq M} |R(b,\hat{b}^{NSEB}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})| \to 0$ for all $M < \infty$.

Part (a) of this theorem provides a device for calculating a lower bound on the frequentist risk of any equivariant estimator in the Gaussian model. This lower bound is the Bayes risk of the subjectivist normal Bayes estimator computed using the "prior" that equals the empirical c.d.f. of the true coefficients. Part (b) states that, in the Gaussian model, the optimal risk is achieved asymptotically by $\hat{b}^{NSEB}$. Moreover, this bound is achieved uniformly for coefficient vectors in a normalized ball (of arbitrary radius) around the origin. Thus, in the Gaussian model, $\hat{b}^{NSEB}$ is asymptotically uniformly (over the ball) minimum risk equivariant.

3.4 Connecting the Frequentist and Bayesian Problems

The fact that $\hat{b}^{NSEB}$ is the optimal estimator in both the Bayes and frequentist estimation problems suggests that the problems themselves are related. It is well known that in conventional, fixed-dimensional parametric settings, by the Bernstein – von Mises argument (e.g. Lehman and Casella [1998, section 6.8]), Bayes estimators and efficient frequentist estimators can be asymptotically equivalent. In those settings, a proper prior is dominated asymptotically by the likelihood. This is not, however, what is happening here. Instead, because the number of coefficients is increasing with the sample size and the coefficients are local to zero, the local coefficients $\{b_i\}$ cannot be estimated consistently. Indeed, Stein's (1955) result that the OLS estimator is inadmissible holds asymptotically here, and the Bayes risks of the OLS and subjectivist Bayes estimators differ asymptotically. Thus the standard argument does not apply here.

Instead, the reason that these two problems are related is that the frequentist risk for equivariant estimators is in effect the Bayes risk, evaluated at the empirical c.d.f. $\tilde{G}_K$. As shown in the appendix, for equivariant estimators, the i^{th} component of the frequentist risk, $E[\tilde{b}_i(\hat{b}) - b_i]^2$, depends on b only through b_i and $\tilde{G}_K$. Thus we might write, $R(b,\tilde{b}) = \rho K^{-1}\sum_{i=1}^{K} E[\tilde{b}_i(\hat{b}) - b_i]^2 = \rho \int E[\tilde{b}_1(\hat{b}) - b_1]^2 d\tilde{G}_K(b_1)$. If the sequence of empirical c.d.f.s $\{\tilde{G}_K\}$ has the weak limit G that is, $\tilde{G}_K \Rightarrow G$, and if the integrand is dominated by an integrable function, then $R(b,\tilde{b}) = \rho \int E[\tilde{b}_1(\hat{b}) - b_1]^2 d\tilde{G}_K(b_1) \Rightarrow \rho \int E[\tilde{b}_1(\hat{b}) - b_1]^2 dG(b_1)$, which is the Bayes risk of $\tilde{b}$. This reasoning extends Edelman's (1988) argument linking the compound decision problem and the Bayes problem (for a narrow class of estimators) in the problem of estimating multiple means under a Gaussian likelihood.

This heuristic argument is made precise in the next theorem. Let $\hat{b}^{NB}_{\tilde{G}_K}$ denote the Normal Bayes estimator constructed using the prior $\tilde{G}_K$, and let $\hat{b}^{NB}_{G}$ denote the Normal Bayes estimator constructed using the prior G, then

Theorem 3: If $\tilde{G}_K \Rightarrow G$ and $\sup_K \|b\|_{2+\delta} \leq M$ for some $\delta > 0$, then $|R(b,\hat{b}^{NB}_{\tilde{G}_K}) - r_G(\hat{b}^{NB}_{G})| \to 0$.

4. Results under Alternative Assumptions

4.1 Alternative Asymptotic Nesting

While the analysis in the last section required large K and T, the only purpose of the restriction that $K=\rho T$ was to provide a convenient asymptotic limit of the risk functions. An alternative formulation relaxes this proportionality restriction and rescales the risk. Thus, we now adopt

Assumption 5: $K = \rho T^{\delta}$ with $0 < \delta \leq 1$ and $0 < \rho < 1$.

4.2 Relaxing Assumptions on the Errors and Regressors

The efficiency results in section 3 relied on two related implications of the Gaussian model: first with σ_ε^2 known, that $\hat{b}$ was sufficient for b, and second that $\hat{b}_i - b_i$ was distributed i.i.d. N(0, σ_ε^2). The first implication yielded the efficiency of the normal Bayes estimator $\hat{b}^{NB}$, and the second implication made it possible to show that the empirical Bayes estimators achieved the same asymptotic risk as $\hat{b}^{NB}$.

In this section we relax the Gaussian assumption and show that the empirical Bayes estimators asymptotically achieve the same risk as $\hat{b}^{NB}$. With non-Gaussian errors, $\hat{b}^{NB}$ is no longer the Bayes estimator, but it is robust in the sense that achieves the minimax risk over all error distributions with the same first and second moment. We will show that the empirical Bayes estimator inherits this minimax property, asymptotically.

The logic underlying these results is straightforward. Let f_K denote the distribution of $\hat{b}$. In the non-Gaussian model $f_K \neq \phi_K$. If K were fixed, then the central limit theorem would imply that f_K converges to ϕ_K. The analysis here is complicated by the fact that $K \to \infty$, but under the assumptions listed below Berry-Esseen results can be used to bound the differences between f_K and ϕ_K yielding the required asymptotic results.

The assumptions explicitly admit weak dependence across observations so that the results of the Gaussian model can be extended to time series applications with X strictly exogenous. Throughout, we adopt the notation that C is a finite constant, possibly different in each occurrence. Let $Z_t = (X_{1t}, \ldots, X_{Kt}, \varepsilon_t)$.

The first assumption restricts the moments of $\{X_t, \varepsilon_t\}$.

Assumption 6:

(a) $E(\varepsilon_t \mid X_t, Z_{t-1}, \ldots, Z_1) = 0$;

(b) $\sup_{it} EX_{it}^{12} \leq C < \infty$ and $\sup_t E\varepsilon_t^{12} \leq C < \infty$;

(c) $E(\varepsilon_t^2 \mid X_t, Z_{t-1}, \ldots, Z_1) = \sigma_\varepsilon^2 > 0$; and

(d) $\sup_t \sup_{Z_1, \ldots, Z_{t-1}} E(\varepsilon_t^4 \mid X_t, Z_{t-1}, \ldots, Z_1) \leq C < \infty$.

The next assumption is that the maximal correlation coefficient of Z decays geometrically (cf. Hall and Heyde [1980], p. 147). Let $\{\nu_n\}$ denote the maximal correlation

coefficients of Z, that is, let $\{\nu_n\}$ be a sequence such that $\sup_m \sup_{y\in L^2(\mathcal{H}_1^m),x\in L^2(\mathcal{H}_{m+n}^\infty)} |\mathrm{corr}(x,y)| = \nu_n$, where $\mathcal{H}_a^b$ is the σ-field generated by the random variables $\{Z_s, s=a,\ldots,b\}$, and $L^2(\mathcal{H}_a^b)$ denotes the set of $\mathcal{H}_a^b$-measurable functions with finite variance.

Assumption 7: $\{Z_t, t=1,\ldots,T\}$ has maximal correlation coefficient ν_n that satisfies $\nu_n \le De^{-\lambda n}$ for some positive finite constants D and λ.

The next assumption places smoothness restrictions on the (conditional) densities of $\{X_{it}\}$ and $\{\varepsilon_{it}\}$.

Assumption 8:

(a) The distributions of $\{X_{it}\}$ and $\{\varepsilon_t\}$ do not depend on $\{b_i\}$.

(b) (i) $\sup_{it} \int_{-\infty}^{\infty} \left|\frac{d^2}{d\varepsilon_t^2} p_{it}^{\varepsilon}(\varepsilon_t \mid \eta_{i,t-1},\ldots,\eta_{i1})\right| d\varepsilon_t \le C$

(ii) $\sup_{ijt} \int_{-\infty}^{\infty} \left|\frac{d^2}{d\varepsilon_t^2} p_{ijt}^{\varepsilon}(\varepsilon_t \mid \eta_{ij,t-1},\ldots,\eta_{ij1})\right| d\varepsilon_t \le C$

(iii) $\sup_{ijt} p_{ijt}^{X}(x_{jt} \mid \eta_{ij,t-1},\ldots,\eta_{ij1}) \le C$

(iv) $\sup_{ijt} p_{ijt}^{X}(x_{it} \mid x_{jt}, \eta_{ij,t-1},\ldots,\eta_{ij1}) \le C$

where $i \ne j$, $\eta_{it} = \frac{X_{it}\varepsilon_t}{\sigma_\varepsilon}$, $\eta_{ijt} = \frac{1}{\sigma_\varepsilon}\begin{pmatrix} X_{it}\varepsilon_t \\ X_{jt}\varepsilon_t \end{pmatrix}$, $p_{it}^{\varepsilon}(\varepsilon_t \mid \eta_{i,t-1},\ldots,\eta_{i1})$ denotes the conditional density of ε_t given $(\eta_{i,t-1}, \ldots, \eta_{i1})$, and so forth.

The final assumption restricts the dependence across the different regressors $\{X_{it}\}$ using a conditional maximal correlation coefficient condition. Let $\underline{X}_i = (X_{i1},\ldots, X_{iT})$ and let $\mathcal{F}_a^b$ be the σ-field generated by the random variables $\{\underline{X}_i, i = a,\ldots,b\}$, and define the cross-sectional conditional maximal correlation coefficients $\{\tau_n\}$ to be a sequence satisfying $\sup_m \sup_{y\in L^2(\mathcal{F}_1^m),x\in L^2(\mathcal{F}_{m+n}^\infty)} |\mathrm{corr}(x,y|\underline{X}_j)| \le \tau_n$ for all j.

Assumption 9: There exists a sequence of cross sectional maximal correlation coefficients $\{\tau_n\}$ such that $\sum_{n=1}^{\infty}\tau_n < \infty$.

With these assumptions, we now state the analogues of Theorems 1 and 2 for the empirical Bayes and equivariant estimators. We begin with a result for OLS. (Proofs for these theorems are contained in Knox, Stock and Watson (2003).) Since the risk functions depend on f_K, the theorems are stated using the notation $R(b,\tilde{b}\,; f_K)$ and $r_G(\tilde{b}; f_K)$.

Theorem 4 : Under Assumptions 1, 5, 6 and 7

(a) $E[(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)^2 | b, \sigma_\varepsilon^2] \le C/K$.

(b) $(T/K)\, R(\hat{b}, b; f_K) = \sigma_\varepsilon^2$, $(T/K) r_G(\hat{b}, f_K) = \sigma_\varepsilon^2$,

The results in Theorem 4 parallel the OLS results in (3.1) and (3.2). Part (a) provides a rate for the consistency of $\hat{\sigma}_\varepsilon^2$ and part (b) repeats (3.2) for the rescaled risk functions.

Theorem 5:

(a) Given Assumptions 1-9:

(i) $(T/K)[r_G(\hat{b}^{PEB}, f_K) - r_G(\hat{b}^{NB}, \phi_K)] \to 0$, and

(ii) $\inf_{\tilde{b}} \sup_{f_K} (T/K)[r_G(\tilde{b}, f_K) - r_G(\hat{b}^{NB}, \phi_K)] \to 0$, where the supremum is taken over the class of likelihoods f_K that satisfy assumptions 6-9 with fixed constants.

(b) Given Assumption 1, 2, 4-9:

(i) $(T/K)[r_G(\hat{b}^{NSEB}, f_K) - r_G(\hat{b}^{NB}, \phi_K)] \to 0$, and

(ii) $\inf_{\tilde{b}} \sup_{f_K} (T/K)[r_G(\tilde{b}, f_K) - r_G(\hat{b}^{NB}, \phi_K)] \to 0$, where the supremum is taken over the class of likelihoods f_K that satisfy assumptions 6-9 with fixed constants.

Parts a(i) and b(i) of Theorem 5 states that the Bayes risk of the EB estimators asymptotically equals the Gaussian Bayes risk of the infeasible estimator. The theorem

states that the Bayes risk of the infeasible estimator $\hat{b}^{NB}$ is achieved even if the conditions for $\hat{b}^{NB}$ to be optimal (Gaussianity) are not met. Moreover, according to part (ii), this risk is achieved uniformly over distributions satisfying the stated assumptions. If $\{\varepsilon_t\}$ has a nonnormal distribution, then the OLS estimators are no longer sufficient, and generally a lower Bayes risk can be achieved by using the Bayes estimator based on the true error distribution. Together these observations imply that $r_G(\hat{b}^{NB}, \phi_K)$ is an upper bound on the Bayes risk of the Bayes estimator under the prior G when $\{\varepsilon_t\}$ is known but nonnormal. However, the EB estimators are asymptotically optimal in the Gaussian model, with Bayes risk that does not depend on the true error distribution asymptotically, so the EB estimators are asymptotically minimax.

Theorem 6: Given Assumptions 1 and 4-9,

$\sup_{\|b\|_2 \le M} \{\sup_{f_K} |(T/K)R(b, \hat{b}^{NSEB}; f_K) - \inf_{\tilde{b} \in \mathcal{B}} (T/K)R(b, \tilde{b}; \phi_K)|\} \to 0$ for all $M < \infty$,

where the supremum over f_K is taken over the class of likelihoods f_K which satisfy assumptions 6-9 with fixed constants.

This theorem parallels part b of Theorem 2. In addition, it shows that even outside the Gaussian model the frequentist risk of $\hat{b}^{NSEB}$ does not depend on f_K for f_K satisfying assumptions 6-9. Because $\hat{b}^{NSEB}$ is optimal among equivariant estimators in the Gaussian model, and because its asymptotic risk does not depend on f_K, it is minimax among equivariant estimators.

4.3 Including Additional Regressors

Often, as in the empirical application in Section 6, a small "base" set of regressors could have large (non-local) coefficients and the remaining many regressors could have small (local) coefficients. Specifically, let U be a $T \times L$ matrix of observations on these base regressors and let γ be a $L \times 1$ vector of non-local coefficients. Incorporating these base regressors yields,

$$y = U\gamma + X\beta + \varepsilon. \tag{4.1}$$

The foregoing results can be extended to (4.1) in two ways. First, in the Gaussian model, it is possible to extend Theorems 1 and 2 to the case that L is held fixed as $K,T \to \infty$; details are given in Knox, Stock and Watson (2003). Second, the method that we use in the empirical analysis of section 6, is to transform the regressors so that $T^{-1}U'X = 0$, so that the results follow directly.

5. Monte Carlo Analysis

5.1. Estimators

Parametric Gaussian EB estimator. The parametric Gaussian EB estimator examined in this Monte Carlo study is based on the parametric specification that $\{b_i\}$ are i.i.d. N(μ, τ^2). Using the normal approximating distribution for the likelihood, the marginal distribution of $\hat{b}_i$ is thus N($\mu, \sigma_{\hat{b}}^2$), where $\sigma_{\hat{b}}^2 = \sigma_{\varepsilon}^2 + \tau^2$. The parameters μ and $\sigma_{\hat{b}}^2$ are consistently estimated by $\hat{\mu} = K^{-1}\sum_{i=1}^{K}\hat{b}_i$ and $\hat{\sigma}_{\hat{b}}^2 = (K-1)^{-1}\sum_{i=1}^{K}(\hat{b}_i - \hat{\mu})^2$. For the Monte Carlo analysis, we treat the sequence of constants s_K as a technical device and thus drop this term from (2.8). Accordingly, the parametric Gaussian empirical Bayes estimator, $\hat{b}^{PEB}$, is given by (2.9) with $\hat{\ell}_K(\hat{b};\hat{\theta}) = -(\hat{b} - \hat{\mu})/\hat{\sigma}_{\hat{b}}^2$.

Nonparametric simple EB estimator. The nonparametric Gaussian simple EB estimator is computed as in (2.10)-(2.14), with some modifications. Following Härdle et. al. (1992), the score function is estimated using the bisquare kernel with bandwidth proportional to $(T/100)^{-2/7}$. Preliminary numerical investigation found advantages to shrinking the nonparametric score estimator towards the parametric Gaussian score estimator. We therefore use the modified score estimator, $\hat{\ell}_{iK}^{s}(x) = \lambda_T(x)\hat{\ell}_{iK}(x) + [1-\lambda_T(x)]\hat{\ell}_K(x;\hat{\theta})$, where $\hat{\ell}_{iK}(x)$ is (2.12) implemented using the bisquare kernel nonparametric score estimator and $s_K = 0$, and, and $\hat{\ell}_K(x;\hat{\theta})$ is given in (2.8). The shrinkage weights are $\lambda_T(x) = \exp[-0.78(x-\hat{\mu})^2/\sigma_{\hat{b}}^2]$. The nonparametric simple EB estimators occasionally produced extremely large estimates, and some results were

sensitive to these outliers. We therefore implemented the upper truncation $|\hat{b}_i^{SNEB}| \leq \max_i|\hat{b}_i|$.

Other benchmark estimators. Results are also reported for some estimators that serve as benchmarks: the infeasible Bayes estimator, the OLS estimator, and the BIC estimator. The infeasible Bayes estimator is the Bayes estimator based on the true G and σ_ε^2; this is feasible only in a controlled experimental setting and provides a lower bound on the Bayes risk. The BIC estimator estimates b_i either by $\hat{b}_i$ or by zero, depending on whether this regressor is included in the regression according to the BIC criterion (enumeration of all possible models is computationally feasible because of the orthonormality of the X's).

5.2 Experimental Designs

We considered two designs. In the first, the data were generated according to (2.1), with ε_t i.i.d. N(0,1), where X_t are the K principal components of $\{W_t, t=1,...,T\}$, where W_{it} are i.i.d. N(0,1) and independent of $\{\varepsilon_t\}$; X_t was rescaled to be orthonormal. The number of regressors was set at $K = \rho T$. The parameters β_i were drawn from a mixture of normal distributions,

(5.1) $$\beta_i \text{ i.i.d. } N(\mu_1, \sigma_1^2) \text{ w.p. } \omega \text{ and } N(\mu_2, \sigma_2^2) \text{ w.p. } 1-\omega.$$

Six configurations of the parameters, taken from Marron and Wand (1992), were chosen to generate a wide range of distribution shapes. The densities are shown are shown in figure 1. The first sets $\omega = 1$, so that the β's are normally distributed. The second and third are symmetric and bimodal, and the fourth is skewed. The fifth density is heavy tailed, and the sixth is extremely so.

The second design is identical to the first except that β_i was set according to

$$\beta_i = \begin{cases} \alpha, i = 1,...,\frac{\zeta}{2}K \\ -\alpha, i = (\frac{\zeta}{2}K)+1,...,\zeta K \\ 0, i = (\zeta K)+1,...,K \end{cases} \tag{5.2}$$

where α and ζ are design parameters. The values of the coefficients were held fixed across the simulations to evaluate the frequentist risk. The regression coefficients in (5.2), while fixed across simulations, have a mean of zero when averaged over *i*; the parameter α was chosen so that $\lambda = (\sigma_b^2 / \sigma_b^2 + \sigma_\varepsilon^2)$ was equal to 0.5, where $\sigma_b^2 = \frac{1}{K}\sum_i b_i^2$.

5.3 Results and Discussion

Results for the first design are shown in Table 1. Panel A shows results for $\lambda = \sigma_b^2 /(\sigma_b^2 + \sigma_\varepsilon^2)$ = 0.5 and various values of *T* and for ρ = 0.25 and ρ = 0.50, where ρ = *K/T*; the entries are the Bayes risk relative to the risk of the OLS estimator. Asymptotically these relative risks do not depend on ρ, and this is evident by comparing the results for the two different values of ρ in panel A. A calculation shows that the Bayes estimator constructed from the Gaussian prior has a relative risk of λ in this design and that λ is the maximum risk over all priors with the same σ_b^2. This is evident in the first column of the table. The BIC estimator generally performs worse than OLS; the exception to this is when the β_i's are generated by the outlier distribution. The PEB estimator, which uses a Gaussian prior, has a performance quite close the minimax bound of 0.5. Interestingly, the SNEB estimator has a similar performance, even when the prior is decidedly non-Gaussian, such as the outlier distribution. In this case the SNEB estimator offers little improvement over the PEB estimator and has a relative performance that is markedly worse than the efficiency bound.

Panel B shows results for different value of λ with *T* = 400 and ρ = 0.25. For small values of λ, the PEB estimator dominates the nonparametric estimator, even when the β_i's are generated by the outlier distribution. In this case most of the variability in the OLS

estimators arise from sampling error (σ_ε^2) and not from variability in the coefficients (σ_b^2) making it difficult for the nonparametric estimator to implicitly estimate the β distribution. In contrast, when λ is large, the nonparametric estimator dominates the parametric estimator, particularly for the the extreme non-Gaussian distributions (the separated bimodal and the outlier distributions). However in these cases, the risk of the nonparametric estimator is still significantly larger than the efficiency bound achieved by the exact Bayes estimator. Panel C repeats the experiment for λ =0.9, but with a sample size of T =4000. The Bayes risk of the nonparametric estimator moves closer to the bound, consistent with the implications of Theorems 1 and 5.

The frequentist risk results for the second experiment are given in table 2. No prior is specified so the exact Bayes estimator is not relevant here. When ζ is small, there are only a few non-zero (and large) coefficients, much like the β_i's generated by the outlier distribution. Thus, the results for ζ = 0.10 are much like those for the outlier distribution in table 1; BIC does well selecting the few non-zero coefficients. However, the performance of BIC drops sharply as ζ increases; BIC and OLS are roughly comparable when ζ = .30. In contrast, the empirical Bayes estimators work well for all values of ζ. The (frequentist) risk of both empirical Bayes estimator offer a 50% improvement on OLS for all values of ζ.

Taken together, these results suggest that the proposed empirical Bayes methods have good finite-sample performance for various distributions of the regression coefficients.

6. Application to Earnings and Education

In a well-known paper, Angrist and Krueger (1991) investigated the causal effect of education on earnings using instrumental variables regression, using as instruments a person's quarter and place of birth. The dataset is large, and this motivated them to use a large number of instruments formed by interacting indicators of time and place of birth. Here we investigate the predictive ability of these variables for education and earnings using the estimators studied in the Monte Carlo experiments.

We use Angrist and Krueger's 1980 Census dataset, which has T = 329,509 observations on men born between 1930 and 1939. The model is of the form (4.1), where y denotes an individual's years of schooling or the logarithm of earnings, U is a vector of 9 regressors that include a constant, age and age^2, and indicators for race, region, SMSA, marital status, and year and state for birth. The vector X includes 178 interactions of quarter of birth with the state and year of birth indicators. These variables have been orthogonalized with respect to U and orthonormalized using principal components. We present results for the estimators studied in the Monte Carlo section: the OLS estimator, an estimator that uses BIC to select elements of X for inclusion in the regression, and parametric and nonparametric empirical Bayes estimators implemented as described in the last section. We also present results for a restricted estimator that sets $\beta = 0$. In all cases γ is estimated by OLS.

To estimate the risk, let $\tilde{\beta}^{(t)}$ denote an estimator computed omitting the t^{th} observation and note that the risk of the forecast of y_t using $\tilde{\beta}^{(t)}$ is $\mathrm{E}(y_t - \tilde{\beta}^{(t)\prime}X_t)^2 = \sigma_\varepsilon^2 + \mathrm{E}[(\tilde{\beta}^{(t)} - \beta)'H_T(\tilde{\beta}^{(t)} - \beta)]$, where $H_T = \mathrm{E}(X_t X_t' | \{X_s, y_s\}_{t=1,s\neq t}^{T}) \cong I_T$ (because X_t is orthonormal); thus $\mathrm{E}(y_t - \tilde{\beta}^{(t)\prime}X_t)^2 \cong \sigma_\varepsilon^2 + R(\tilde{b}^{(t)}, b)$. This suggests the estimator of the risk, $\tilde{\sigma}_{cv}^2 - \hat{\sigma}_\varepsilon^2$, where $\tilde{\sigma}_{cv}^2$ is the leave-one-out cross-validation estimator of the forecast risk and $\hat{\sigma}_\varepsilon^2$ is the OLS estimator of σ_ε^2. As in the Monte Carlo section, we present estimates of the estimation risk relative to risk for OLS.

The results are shown in Table 3. The first row of the table shows the F statistic for testing the hypothesis $\beta = 0$. The next row shows an estimate of $\lambda = \sigma_b^2/(\sigma_b^2 + \sigma_\varepsilon^2)$ constructed as $\hat{\lambda} = F/(F-1)$. Also shown is a 95% confidence interval for λ constructed using the value of regression F–statistic and quantiles of non-central F distribution. The final four rows show the estimated relative estimation risk.

Three aspects of the results are noteworthy. First, the empirical Bayes estimators outperform OLS for both education and for earnings. Second, the empirical Bayes estimators outperform the restricted estimator for education, which has a relatively large and statistically significant F-statistic, and perform as well as the restricted estimator for earnings, where the F-statistic is small and insignificant. Finally, the estimator performance is broadly consistent with the predictions of the analysis in Sections 3 and 4.

For example, that analysis predicts that the limiting relative estimation risk of the restricted estimator is $\lambda/(1-\lambda)$ and is equal to λ for the parametric Bayes estimator. The point estimates of relative risk for the restricted estimators are contained in the 95% confidence intervals for $\lambda/(1-\lambda)$; the point estimates for the parametric empirical Bayes risk are slightly larger than expected, but this may be a reflection of sampling error in the cross-validation estimates of the relative risk.

Appendix

Definitions and Notation: Unless otherwise noted, all limits of integration are from -∞ to ∞, and limits of summation are from 1 to *K*. The densities $\phi(u)$ and $\phi_K(u)$ denote $\phi(u;\sigma_\varepsilon^2)$ and $\phi_K(u;\sigma_\varepsilon^2)$; $m(\hat{b}_i)$ and $m_K(\hat{b})$ denote the marginal densities of $\hat{b}_i$ and $\hat{b}$. Let

$$\hat{b}_i^{INB} = \hat{b}_i + \sigma_\varepsilon^2 \frac{m'(\hat{b}_i)}{m(\hat{b}_i) + s_K} \tag{A.1}$$

be an infeasible estimator constructed using the actual marginal distribution. Finally, $\hat{b}_{-i} = (\hat{b}_1, ..., \hat{b}_{i-1}, \hat{b}_{i+1}, ..., \hat{b}_K)$ and $d_K = \sqrt{(\hat{\sigma}_\varepsilon^2 / 64)\log K}$.

Lemma 1: Under assumptions 1, 2 and 4, the following results hold uniformly in *K*:

$$\int \frac{(m'(x))^2}{m(x)} dx < \infty \tag{A.2}$$

$$\int \left(\frac{m'(x)}{m(x) + s_k} \right)^2 m(x) dx \to \int \frac{(m'(x))^2}{m(x)} dx \tag{A.3}$$

$$r_G(\hat{b}^{NB}) < \infty \tag{A.4}$$

$$r_G(\hat{b}^{INB}) < \infty \tag{A.5}$$

$$\sup_x |m'(x)| < \infty . \tag{A.6}$$

Further, if assumption 3 holds

$$\sup_{x,\theta\in\Theta} |m'(x;\theta,\sigma_\varepsilon^2)| < \infty \tag{A.7}$$

$$\sup_{x,\theta} \left\| \frac{\partial \dfrac{\sigma_\varepsilon^2 m'(x;\theta,\sigma_\varepsilon^2)}{m(x;\theta,\sigma_\varepsilon^2) + s_K}}{\partial\theta} \right\| \le \sigma_\varepsilon^2 \frac{C}{s_K^2} \tag{A.8}$$

$$\sup_{x,\theta,\hat{\sigma}_\varepsilon^2 \in [0.5\sigma_\varepsilon^2, 1.5\sigma_\varepsilon^2]} \left\| \frac{\partial \dfrac{\hat{\sigma}_\varepsilon^2 m'(x;\theta,\hat{\sigma}_\varepsilon^2)}{m(x;\theta,\hat{\sigma}_\varepsilon^2) + s_K}}{\partial\hat{\sigma}_\varepsilon^2} \right\| \le \frac{C}{s_K^2} \tag{A.9}$$

Proof: Knox, Stock and Watson (2003).

Lemma 2: Under assumptions 1, 2 and 4, $\exists\ C, K_0 < \infty$ such that $\forall K \geq K_0$,

$$\begin{aligned}&\sup_{i,|\hat{b}_i|<d_K} E\{[\hat{m}_{iK}(\hat{b}_i) - m(\hat{b}_i)]^2 \mid \hat{b}_i\} \leq \\ &C\left(\frac{1}{h_K(K-1)} + h_K^2 + K^{-13/48}\log^2 K + 2K^{-13/96}h_K \log K + K^{-1/2}\log^2 K\right)\end{aligned} \tag{A.10}$$

$$\begin{aligned}&\sup_{i,|\hat{b}_i|<d_K} E\{[\hat{m}'_{iK}(\hat{b}_i) - m'(\hat{b}_i)]^2 \mid \hat{b}_i\} \leq \\ &C\left(\frac{1}{h_K^3(K-1)} + h_K^2 + K^{-5/48}\log^2 K + 2K^{-5/96}h_K \log K + K^{-1/4}\log^2 K\right)\end{aligned} \tag{A.11}$$

$$\sup_{i,|\hat{b}_i|\leq d_K} \Pr[\left|\breve{l}_i(\hat{b}_i)\right| > q_K \mid \hat{b}_i] \to 0\,. \tag{A.12}$$

Proof: Knox, Stock and Watson (2003).

Proof of Theorem 1

Part (a): Write

$$r_G(\hat{b}^{PEB}) - r_G(\hat{b}^{NB}) = [r_G(\hat{b}^{INB}) - r_G(\hat{b}^{NB})] + [r_G(\hat{b}^{PEB}) - r_G(\hat{b}^{INB})] \tag{A.13}$$

where $\hat{b}^{INB}$ is defined in (A.1). The proof shows that both bracketed terms converges to zero. The first bracketed term in (A.13) is

$$\begin{aligned}&[r_G(\hat{b}^{INB}) - r_G(\hat{b}^{NB})] \\ &= \rho\frac{1}{K}\sum\int\int\left[\left(\hat{b}_i^{INB}(\hat{b}_i) - b_i\right)^2 - \left(\hat{b}_i^{NB}(\hat{b}_i) - b_i\right)^2\right]\phi(\hat{b}_i - b_i)d\hat{b}_i dG(b_i,\theta) \\ &= \rho\int\int\left[\left(\hat{b}_1^{INB}(\hat{b}_1) - b_1\right)^2 - \left(\hat{b}_1^{NB}(\hat{b}_1) - b_1\right)^2\right]\phi(\hat{b}_1 - b_1)d\hat{b}_1 dG(b_1,\theta) \\ &\to 0\end{aligned} \tag{A.14}$$

where the second equality follows because the summands are identical. The convergence follows from the dominated convergence theorem since $\hat{b}_1^{INB}(x) - \hat{b}_1^{NB}(x) = \sigma_\varepsilon^2\{m'(x)/[m(x)+s_K] - m'(x)/m(x)\} \to 0$ pointwise by Assumption 4, and

$(\hat{b}_1^{INB}(\hat{b}_1) - b_1)^2 \le 2(b_1 - \hat{b}_1)^2 + 2\sigma_\varepsilon^2 \left(\frac{m'(\hat{b}_1; \theta, \sigma_\varepsilon^2)}{m(\hat{b}_1; \theta, \sigma_\varepsilon^2) + s_K} \right)^2$ (since $(a + b)^2 \le 2a^2 + 2b^2$), which is integrable because $\hat{b}_1 - b \sim \text{N}(0, \sigma_\varepsilon^2)$ and Lemma 1.

Now consider $[r_G(\hat{b}^{PEB}) - r_G(\hat{b}^{INB})]$, the second bracketed term in (A.13). Because $r_G(\hat{b}^{INB})$ is bounded (Lemma 1), $[r_G(\hat{b}^{PEB}) - r_G(\hat{b}^{INB})] \to 0$ is implied by $\frac{1}{K}\sum E(\hat{b}_i^{PEB} - \hat{b}_i^{INB})^2 \to 0$. Let $\gamma = (\theta, \sigma_\varepsilon^2)$ and $h(x, \gamma) = \sigma_\varepsilon^2 \frac{m'(x; \theta, \sigma_\varepsilon^2)}{m(x; \theta, \sigma_\varepsilon^2) + s_K}$, so that $\hat{b}_i^{INB} - \hat{b}_i^{PEB} = h(\hat{b}_i, \gamma) - h(\hat{b}_i, \hat{\gamma})$. Let $F(\hat{\gamma} \mid \hat{b}, b)$ denote the conditional c.d.f. of $\hat{\gamma}$, and note that this can be written as $F(\hat{\gamma} \mid \hat{b})$. Thus

$$
\begin{aligned}
&\frac{1}{K}\sum E(\hat{b}_i^{PEB} - \hat{b}_i^{INB})^2 \\
&= \frac{1}{K}\sum \int\int (h(\hat{b}_i, \gamma) - h(\hat{b}_i, \hat{\gamma}))^2 \, dF(\hat{\gamma} \mid \hat{b}) m_K(\hat{b}; \gamma) d\hat{b} \\
&= \frac{1}{K}\sum \int\int\int\int (h(\hat{b}_i, \gamma) - h(\hat{b}_i, \hat{\gamma}))^2 \, dF(\hat{\theta} \mid \hat{b}, \hat{\sigma}_\varepsilon^2) dJ(\hat{\sigma}_\varepsilon^2) m_K(\hat{b}; \gamma) d\hat{b} \\
&\le \frac{2}{K}\sum \int\int\int\int (h(\hat{b}_i, \theta, \sigma_\varepsilon^2) - h(\hat{b}_i, \hat{\theta}, \sigma_\varepsilon^2))^2 \, dF(\hat{\theta} \mid \hat{b}, \hat{\sigma}_\varepsilon^2) dJ(\hat{\sigma}_\varepsilon^2) m_K(\hat{b}; \gamma) d\hat{b} \\
&+ \frac{2}{K}\sum \int\int\int\int (h(\hat{b}_i, \hat{\theta}, \sigma_\varepsilon^2) - h(\hat{b}_i, \hat{\theta}, \hat{\sigma}_\varepsilon^2))^2 \, dF(\hat{\theta} \mid \hat{b}, \hat{\sigma}_\varepsilon^2) dJ(\hat{\sigma}_\varepsilon^2) m_K(\hat{b}; \gamma) d\hat{b} \\
&\le \frac{2C^2\sigma_\varepsilon^4}{Ks_K^4}\sum \int \left\| \hat{\theta} - \theta \right\|^2 dF(\hat{\theta}) \\
&+ \frac{2}{K}\sum \int \int_{\hat{\sigma}_\varepsilon^2 \in [0.5\sigma_\varepsilon^2, 1.5\sigma_\varepsilon^2]} \int (h(\hat{b}_i, \hat{\theta}, \sigma_\varepsilon^2) - h(\hat{b}_i, \hat{\theta}, \hat{\sigma}_\varepsilon^2))^2 \, dF(\hat{\theta} \mid \hat{b}, \hat{\sigma}_\varepsilon^2) dJ(\hat{\sigma}_\varepsilon^2) m_K(\hat{b}; \gamma) d\hat{b} \qquad \text{(A.15)} \\
&+ \frac{2}{K}\sum \int \int_{\hat{\sigma}_\varepsilon^2 \notin [0.5\sigma_\varepsilon^2, 1.5\sigma_\varepsilon^2]} \int (h(\hat{b}_i, \hat{\theta}, \sigma_\varepsilon^2) - h(\hat{b}_i, \hat{\theta}, \hat{\sigma}_\varepsilon^2))^2 \, dF(\hat{\theta} \mid \hat{b}, \hat{\sigma}_\varepsilon^2) dJ(\hat{\sigma}_\varepsilon^2) m_K(\hat{b}; \gamma) d\hat{b}
\end{aligned}
$$

where the first equality is by definition, the second introduces the notation J for the marginal distribution of $\hat{\sigma}_\varepsilon^2$, the first inequality uses $(a+b)^2 \le 2a^2+2b^2$, and the first term in the final inequality uses the mean value thereom and the derivative bound in (A.8).

We now consider each of the terms in the final inequality. The first term converges to zero by assumption 3(c). As for the second term,

$$\frac{2}{K}\sum\int\int_{\hat{\sigma}_\varepsilon^2\in[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2]}\int(h(\hat{b}_i,\hat{\theta},\sigma_\varepsilon^2)-h(\hat{b}_i,\hat{\theta},\hat{\sigma}_\varepsilon^2))^2\,dF(\hat{\theta}\mid\hat{b},\hat{\sigma}_\varepsilon^2)dJ(\hat{\sigma}_\varepsilon^2)m_K(\hat{b};\gamma)d\hat{b}$$

$$\leq\frac{2}{K}\sum\int\{\sup_{\hat{b}_i,\hat{\theta},\sigma_\varepsilon^2}\left|\frac{\partial h(\hat{b}_i,\hat{\theta},\sigma_\varepsilon^2)}{\partial\sigma_\varepsilon^2}\right|^2\}(\hat{\sigma}_\varepsilon^2-\sigma_\varepsilon^2)^2\,dJ(\hat{\sigma}_\varepsilon^2)$$

$$\leq\frac{C}{Ks_K^4}\to 0\,,$$

where the first inequality follows from the mean value theorem and the second inequality follows from (A.9) and (3.1). As for the third term, let $\lambda(z)$ denote the standard normal density, then

$$\frac{2}{K}\sum\int\int_{\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2]}\int(h(\hat{b}_i,\hat{\theta},\sigma_\varepsilon^2)-h(\hat{b}_i,\hat{\theta},\hat{\sigma}_\varepsilon^2))^2\,dF(\hat{\theta}\mid\hat{b},\hat{\sigma}_\varepsilon^2)dJ(\hat{\sigma}_\varepsilon^2)m_K(\hat{b};\gamma)d\hat{b}$$

$$\leq\frac{4}{Ks_K^2}\sum\int\int_{\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2]}\sup_{\hat{\theta}}\{(\hat{\sigma}_\varepsilon^2 m'(\hat{b}_i,\hat{\theta},\hat{\sigma}_\varepsilon^2))^2+(\sigma_\varepsilon^2 m'(\hat{b}_i,\hat{\theta},\sigma_\varepsilon^2))^2\}dJ(\hat{\sigma}_\varepsilon^2)m_K(\hat{b};\gamma)d\hat{b}$$

$$\leq\frac{4}{Ks_K^2}\sum\int_{\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2]}\left\{\begin{array}{l}\hat{\sigma}_\varepsilon^2\sup_{\hat{\theta},\hat{b}_i}(\int z\lambda(z)g(\hat{b}_i-\hat{\sigma}_\varepsilon^2 z;\hat{\theta})dz)^2\\+\sigma_\varepsilon^2\sup_{\hat{\theta},\hat{b}_i}(\int z\lambda(z)g(\hat{b}_i-\sigma_\varepsilon^2 z;\hat{\theta})dz)^2\end{array}\right\}dJ(\hat{\sigma}_\varepsilon^2)$$

$$\leq\frac{4}{Ks_K^2}\sum\int_{\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2]}\left\{\begin{array}{l}\hat{\sigma}_\varepsilon^2\int\lambda(z)z^2\sup_{\hat{\theta},\hat{b}_i}\{g(\hat{b}_i-\hat{\sigma}_\varepsilon^2 z;\hat{\theta})\}^2dz\\+\sigma_\varepsilon^2(\int\lambda(z)z^2\sup_{\hat{\theta},\hat{b}_i}\{g(\hat{b}_i-\sigma_\varepsilon^2 z;\hat{\theta})\}^2dz\end{array}\right\}dJ(\hat{\sigma}_\varepsilon^2)$$

$$\leq\frac{4}{Ks_K^2}\sum\int_{\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2]}(\hat{\sigma}_\varepsilon^2C_1+\sigma_\varepsilon^2C_2)dJ(\hat{\sigma}_\varepsilon^2)$$

$$=\frac{4}{Ks_K^2}\sum\{\hat{\sigma}_\varepsilon^2\mathbf{1}(\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2])+\sigma_\varepsilon^2\Pr(\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2])\}$$

$$\leq\frac{4C}{Ks_K^2}\sum\{\sqrt{E(\hat{\sigma}_\varepsilon^4)}\sqrt{\Pr(\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2])}+\sigma_\varepsilon^2\Pr(\hat{\sigma}_\varepsilon^2\notin[0.5\sigma_\varepsilon^2,1.5\sigma_\varepsilon^2])\}$$

$$\leq\frac{4C_3}{s_K^2\sqrt{K}}+\frac{4C_4}{s_K^2K}\to 0\,,$$

where the first inequality uses the definition of $h(.)$, $m(.)\geq 0$ and replaces an expectation with the sup; the second inequality uses the definition of m' and the change of variables

$z=(\hat{b}_i - b_i)/\hat{\sigma}_\varepsilon$; the third inequality uses Jensen's inequality; the fourth inequality follows from assumption 3(b); the fifth inequality uses Cauchy-Schwartz, and the final inequality uses Chebyshev's inequality. Thus, the three terms making up the final inequality in (A.15) converge to 0, and this completes the proof to part (a) of Theorem 1.

Part (b):

Using the same steps as in part(a), the result in part(b) is implied by $\frac{1}{K}\sum E(\hat{b}_i^{NSEB} - \hat{b}_i^{INB})^2 \to 0$. To show this, we use an argument similar to Bickel *et al.* (1993, p. 405) and van der Vaart (1988, p. 169):

$$\begin{aligned}
&\frac{1}{K}\sum E(\hat{b}_i^{PEB} - \hat{b}_i^{INB})^2 \\
&= \frac{1}{K}\sum\int\int\left(\hat{\sigma}_\varepsilon^2 \hat{l}(\hat{b}_i;\hat{b}_{-i}) - \sigma_\varepsilon^2 \frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 \phi_K(\hat{b}-b)d\hat{b}dG_K(b) \\
&\le \frac{2}{K}\sum\int\int(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)^2 \hat{l}^2(\hat{b}_i;\hat{b}_{-i})\phi_K(\hat{b}-b)d\hat{b}dG_K(b) \qquad \text{(Term A)} \\
&+ \frac{2\sigma_\varepsilon^4}{K}\sum\int\int\left(\hat{l}(\hat{b}_i;\hat{b}_{-i}) - \frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 \phi_K(\hat{b}-b)d\hat{b}dG_K(b) \qquad \text{(Term B)}
\end{aligned} \tag{A.16}$$

where the inequality uses $(a+b)^2 \le 2a^2 + 2b^2$.

Term A satisfies

$$\begin{aligned}
&\frac{2}{K}\sum\int\int(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)^2 \hat{l}^2(\hat{b}_i;\hat{b}_{-i})\phi_K(\hat{b}-b)d\hat{b}dG_K(b) \\
&\le \frac{2q_K^2}{K}\sum\int\int(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)^2 \phi_K(\hat{b}-b)d\hat{b}dG_K(b) \\
&= 2q_K^2 \operatorname{var}(\hat{\sigma}_\varepsilon^2) \to 0
\end{aligned} \tag{A.17}$$

where the inequality comes from the truncation of the score estimator and the convergence follows from (3.1) and Assumption 4.

Now consider Term B. Let $D_i = \{\hat{b}: |\hat{b}_i| \le \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K}$ and $|\breve{l}(\hat{b}_i;\hat{b}_{-i})| \le q_K\}$, and let $E_{D_i}[(\cdot)] \equiv \int_{\hat{b}\in D_i}(\cdot)m_K(\hat{b})d\hat{b}$. Now

$$\frac{2\sigma_\varepsilon^4}{K}\sum\int\int\left(\hat{l}(\hat{b}_i;\hat{b}_{-i})-\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2\phi_K(\hat{b}-b)d\hat{b}dG_K(b) \tag{A.18}$$

$$=\frac{2\sigma_\varepsilon^4}{K}\sum\int\left(\hat{l}(\hat{b}_i;\hat{b}_{-i})-\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 m_K(\hat{b})d\hat{b}$$

$$=\frac{2\sigma_\varepsilon^4}{K}\sum\int E\left(\hat{l}(\hat{b}_i;\hat{b}_{-i})-\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\,|\,\hat{b}_i\right)^2 m(\hat{b}_i)d\hat{b}_i$$

$$\le\frac{2\sigma_\varepsilon^4}{K}\sum\int\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 m(\hat{b}_i)\times \qquad \text{(Term Bi)}$$

$$\left(\Pr\left(\left|\hat{b}_i\right|>\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K}\,|\,\hat{b}_i\right)+\Pr\left(\left|\breve{l}(\hat{b}_i;\hat{b}_{-i})\right|>q_K\,|\,\hat{b}_i\right)\right)d\hat{b}_i$$

$$+\frac{2\sigma_\varepsilon^4}{K}\sum E_{D_i}\left[\left(\hat{l}(\hat{b}_i;\hat{b}_{-i})-\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2\right]. \qquad \text{(Term Bii)}$$

Consider Term Bi:

$$\frac{2\sigma_\varepsilon^4}{K}\sum\int\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 m(\hat{b}_i)\times$$

$$\left(\Pr\left(\left|\hat{b}_i\right|>\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K}\,|\,\hat{b}_i\right)+\Pr\left(\left|\breve{l}(\hat{b}_i;\hat{b}_{-i})\right|>q_K\,|\,\hat{b}_i\right)\right)d\hat{b}_i \tag{A.19}$$

$$\le\frac{2\sigma_\varepsilon^2}{Ks_K^2}C\sum\Pr\left(\left|\hat{b}_i\right|>\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K}\right) \qquad \text{(Term Bia)}$$

$$+2\sigma_\varepsilon^2\int_{|\hat{b}_1|>d_K}\left(\frac{m'(\hat{b}_1)}{m(\hat{b}_1)+s_K}\right)^2 m(\hat{b}_1)d\hat{b}_1 \qquad \text{(Term Bib)}$$

$$+2\sigma_\varepsilon^4\left[\sup_{i,|\hat{b}_i|\le d_K}\Pr\left(\left|\breve{l}(\hat{b}_i;\hat{b}_{-i})\right|>q_k\,|\,\hat{b}_i\right)\right]\int_{-d_K}^{d_K}\left(\frac{m'(\hat{b}_1)}{m(\hat{b}_1)+s_K}\right)^2 m(\hat{b}_1)d\hat{b}_1. \quad \text{(Term Bic)}$$

where d_k is defined in the first paragraph of this appendix. Term Bib converges to 0 by the integrability of the integrand (Lemma 1) and $d_K\to\infty$. Similarly, the integral in Term Bic is bounded (Lemma 1) and the bracketed term converges to 0 (Lemma 2) so that Term Bic converges to 0. For Term Bia:

$$\frac{2\sigma_\varepsilon^2}{Ks_K^2}C\sum \Pr\left(\left|\hat{b}_i\right| > \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K}\right)$$
$$\leq \frac{2\sigma_\varepsilon^2}{Ks_K^2}C\sum\left[\Pr(\left|\hat{b}_i\right| > \sqrt{\frac{\sigma_\varepsilon^2}{256}\log K}) + \Pr(\hat{\sigma}_\varepsilon^2 \leq \frac{1}{2}\sigma_\varepsilon^2)\right] \quad \text{(A.20)}$$
$$\leq C_2 s_K^{-2}\left\{\frac{1}{\log K} + \frac{1}{K}\right\} \to 0$$

where the first inequality follows from the observation that $\left|\hat{b}_i\right| > \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{128}\log K}$ implies that $\left|\hat{b}_i\right| > \sqrt{\frac{\sigma_\varepsilon^2}{256}\log K}$ or $\hat{\sigma}_\varepsilon^2 \leq \frac{1}{2}\sigma_\varepsilon^2$ and the second inequality follows from Chebyshev's inequality. Thus Term Bi converges to 0.

Now, consider Term Bii. Define $E_{D_i}^{out}[(\cdot)]$ as $\int_{\hat{b}\in D_i \cap \{\hat{\sigma}_\varepsilon^2 > 2\sigma_\varepsilon^2\}} (\cdot) m_K(\hat{b}) d\hat{b}$ and $E_{D_i}^{cond}[(\cdot)\,|\,\hat{b}_i]$ as $\int_{\hat{b}\in D_i} (\cdot) m_K(\hat{b}_{-i}\,|\,\hat{b}_i) d\hat{b}_{-i}$. Now, the i^{th} term in the sum in Term Bii is

$$E_{D_i}\left[\left(\hat{l}(\hat{b}_i;\hat{b}_{-i}) - \frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2\right] \quad \text{(A.21)}$$
$$\leq 2E_{D_i}^{out}\left[\left(\hat{l}(\hat{b}_i;\hat{b}_{-i})\right)^2\right] + 2E_{D_i}^{out}\left[\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2\right]$$
$$+\int_{-d_K}^{d_K} E_{D_i}^{cond}\left[\left(\hat{l}(\hat{b}_i;\hat{b}_{-i}) - \frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 |\,\hat{b}_i\right] m(\hat{b}_i) d\hat{b}_i$$
$$\leq 2C(q_K^2 + s_K^{-2})\Pr(\hat{\sigma}_\varepsilon^2 > 2\sigma_\varepsilon^2)$$
$$+2\int_{-d_K}^{d_K}\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 E_{D_i}^{cond}\left[\left(\frac{m(\hat{b}_i)-\hat{m}_{iK}(\hat{b}_i)}{\hat{m}_{iK}(\hat{b}_i)+s_K}\right)^2 |\,\hat{b}_i\right] m(\hat{b}_i) d\hat{b}_i$$
$$+2\int_{-d_K}^{d_K} E_{D_i}^{cond}\left[\left(\frac{\hat{m}'_{iK}(\hat{b}_i)-m'(\hat{b}_i)}{\hat{m}_{iK}(\hat{b}_i)+s_K}\right)^2 |\,\hat{b}_i\right] m(\hat{b}_i) d\hat{b}_i$$

$$\leq 2(q_K^2 + s_K^{-2})\frac{C_2}{K}$$

$$+\frac{2}{s_K^2}\int_{-d_K}^{d_K}\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 E_{D_i}^{cond}\left[\left(m(\hat{b}_i)-\hat{m}_{iK}(\hat{b}_i)\right)^2 \mid \hat{b}_i\right] m(\hat{b}_i)d\hat{b}_i \qquad \text{(Term Biia)}$$

$$+\frac{2}{s_K^2}\int_{-d_K}^{d_K} E_{D_i}^{cond}\left[\left(\hat{m}'_{iK}(\hat{b}_i)-m'(\hat{b}_i)\right)^2 \mid \hat{b}_i\right] m(\hat{b}_i)d\hat{b}_i \qquad \text{(Term Biib)}$$

where the first inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$, and the second inequality follows from the truncation of the score estimator, the boundedness of $(m'(x))^2$ (Lemma 1) and the result $\left(\frac{a}{b}-\frac{c}{d}\right)^2 = \left(\frac{a}{b}\left(\frac{d-b}{d}\right)+\frac{a-c}{d}\right)^2 \leq 2\left(\frac{a}{b}\right)^2\left(\frac{d-b}{d}\right)^2 + 2\left(\frac{a-c}{d}\right)^2$. The final inequality is by Chebychev's equality. The first term in the final expression converges to 0 uniformly in *i* by Assumption 4. Term Biia satisfies

$$\frac{2}{s_K^2}\int_{-d_K}^{d_K}\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 E_{D_i}^{cond}\left[\left(m(\hat{b}_i)-\hat{m}_{iK}(\hat{b}_i)\right)^2 \mid \hat{b}_i\right] m(\hat{b}_i)d\hat{b}_i$$

$$\leq \frac{2}{s_K^2} C\left(\frac{1}{h_K(K-1)} + h_K^2 + K^{-13/48}\log^2 K + 2K^{-13/96}h_K \log K + K^{-1/2}\log^2 K\right)$$

$$\times \int_{-d_K}^{d_K}\left(\frac{m'(\hat{b}_i)}{m(\hat{b}_i)+s_K}\right)^2 m(\hat{b}_i)d\hat{b}_i \;\to 0$$

where the inequality follows from Lemma 2, and the convergence follows from Assumption 4 and Lemma 1. Simlarly, Term Biib satisfies

$$\frac{2}{s_K^2}\int_{-d_K}^{d_K} E_{D_i}^{cond}\left[\left(\hat{m}'_{iK}(\hat{b}_i)-m'(\hat{b}_i)\right)^2 \mid \hat{b}_i\right] m(\hat{b}_i)d\hat{b}_i$$

$$\leq \frac{2}{s_K^2} C\left(\frac{1}{h_K^3(K-1)} + h_K^2 + K^{-5/48}\log^2 K + 2K^{-5/96}h_K \log K + K^{-1/4}\log^2 K\right)$$

$$\to 0$$

Thus, Term Bii converges to 0.

Proof of Theorem 2

Part (a): The proof uses two characteristics of equivariant estimators. To derive the first, let $\boldsymbol{P}_{-i}$ denote a permutation operator that is restricted to leave the i^{th} index unchanged (but may permute the other indexes). Let $\tilde{b} \in \mathcal{B}$, and note that $\tilde{b}_i(\boldsymbol{P}_{-i}\hat{b}) = (\boldsymbol{P}_{-i}\tilde{b})_i(\hat{b}) = \tilde{b}_i(\hat{b})$. Thus, while $\tilde{b}_i$ may depend on the values of the elements in $\hat{b}_{-i}$, it cannot depend on their order. Equivalently $\tilde{b}_i$ may depend on $\hat{b}_{-i}$ only through the empirical c.d.f. of $\hat{b}_{-i}$. Denote this c.d.f. by $\hat{G}_{K,-i}$. The second characteristic is $\tilde{b}_i(\hat{b}) = \tilde{b}_i(\hat{b}_1, \hat{b}_2, ..., \hat{b}_{i-1}, \hat{b}_i, \hat{b}_{i+1}, ..., \hat{b}_K) = \tilde{b}_1(\hat{b}_i, \hat{b}_2, ..., \hat{b}_{i-1}, \hat{b}_1, \hat{b}_{i+1}, ..., \hat{b}_K)$. Together, these two characteristics mean that an equivariant estimator can be written as

$$\tilde{b}_i(\hat{b}) = \tilde{b}_1(\hat{b}_i; \hat{G}_{K,-i}) \tag{A.22}$$

Now, consider the i^{th} component of the risk function for $\tilde{b} \in \mathcal{B}$:

$$\begin{aligned} R_i(b, \tilde{b}) &= E\left[\left(\tilde{b}_i(\hat{b}) - b_i\right)^2 \mid b\right] \\ &= E\left[\left(\tilde{b}_1(\hat{b}_i; \hat{G}_{K,-i}) - b_i\right)^2 \mid b\right] \\ &= E\left[\left(\tilde{b}_1(\hat{b}_i; \hat{G}_{K,-i}) - b_i\right)^2 \mid b_i, \tilde{G}_{K,-i}\right] \end{aligned} \tag{A.23}$$

where $\tilde{G}_{K,-i}$ denotes the empirical c.d.f. of b_{-i}. The second equality in (A.23) uses (A.22), and the third uses the fact that b_{-i} affects the risk only through $\tilde{G}_{K,-i}$. Thus,

$$\begin{aligned} R(b, \tilde{b}) &= \rho \frac{1}{K} \sum E\left[\left(\tilde{b}_1(\hat{b}_i; \hat{G}_{K,-i}) - b_i\right)^2 \mid b_i, \tilde{G}_{K,-i}\right] \\ &= \rho \int E\left[\left(\tilde{b}_1(\hat{b}_{ind(z)}, \hat{G}_{K,-ind(z)}) - z\right)^2 \mid z, \tilde{G}_{K,-ind(z)}\right] d\tilde{G}_K(z) \end{aligned} \tag{A.24}$$

where z is the variable of integration and $\hat{b}_{ind(z)}$ is the value is the coordinate of $\hat{b}$ corresponding to the value of b that z takes on.

The equivariant estimator $\tilde{b}_1$ depends on $\hat{b}$ in a restricted way. Removing this restriction cannot increase the risk, so

$$\inf_{\tilde{b} \in \mathcal{B}} R(b, \tilde{b}) \geq \inf_{b^* \in \mathcal{B}^*} \left\{ \rho \int E\left[\left(b^*(\hat{b}) - z\right)^2 \mid z, \tilde{G}_{K,-i}\right] d\tilde{G}_K(z) \right\}, \tag{A.25}$$

where $\mathcal{B}^*$ is the set of regular estimators of *b* that are functions of only $\hat{b}$. The usual Bayesian calculations show that the inf of the right hand side of (A.25) is obtained using the posterior mean of *b* constructed from the normal likelihood and the prior $\tilde{G}_K$. In the notation of this paper, this is normal Bayes estimator using $\tilde{G}_K$ as the prior, say $\hat{b}^{NB}_{\tilde{G}_K}$. Thus (A.25) implies

$$\inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b}) \geq R(b,\hat{b}^{NB}_{\tilde{G}_K}) = r_{\tilde{G}_K}(\hat{b}^{NB}) \tag{A.26}$$

Part (b): Since

$$\limsup_{K\to\infty} \sup_{\|b\|_2\leq M} \left|R(b,\hat{b}^{NSEB}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right| \tag{A.27}$$

$$\leq \max\left\{\begin{array}{l} -\liminf_{K\to\infty} \inf_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NSEB}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right\}, \\ \limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NSEB}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right\} \end{array}\right\}$$

the theorem follows by showing that each of these terms is non-positive.

For the first term, observe that $\hat{b}^{NSEB}$ is an equivariant estimator, so that

$$-\liminf_{K\to\infty} \inf_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NSEB}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right\} \leq 0 \tag{A.28}$$

follows immediately.

To show the corresponding result for the second term, we first note that

$$\lim_{K\to\infty} \sup_{\|b\|_2\leq M} \left|R(b,\hat{b}^{NSEB}) - R(b,\hat{b}^{NB}_{\tilde{G}_K})\right| = 0 \tag{A.29}$$

follows from an argument like that used to prove Theorem 2b. (See Knox, Stock and Watson(2003) for the details.) Now,

$$\limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NSEB}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right\} \tag{A.30}$$

$$= \limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\{[R(b,\hat{b}^{NSEB}) - R(b,\hat{b}^{NB}_{\tilde{G}_K})] + [R(b,\hat{b}^{NB}_{\tilde{G}_K}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})]\}$$

$$\leq \limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NSEB}) - R(b,\hat{b}^{NB}_{\tilde{G}_K})\right\}$$

$$+ \limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NB}_{\tilde{G}_K}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right\}$$

$$\leq \limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\left|R(b,\hat{b}^{NSEB}) - R(b,\hat{b}^{NB}_{\tilde{G}_K})\right|$$

$$+ \limsup_{K\to\infty} \sup_{\|b\|_2\leq M}\left\{R(b,\hat{b}^{NB}_{\tilde{G}_K}) - \inf_{\tilde{b}\in\mathcal{B}} R(b,\tilde{b})\right\}$$

$$\leq \limsup_{K\to\infty} \sup_{\|b\|_2 \leq M} \left\{ R(b, \hat{b}^{NB}_{\tilde{G}_K}) - \inf_{\tilde{b}\in\mathcal{B}} R(b, \tilde{b}) \right\}$$
$$\leq 0$$

where the third inequality follows from (A.29) and the final inequality follows from part(a) of the theorem.

Proof of Theorem 3

The frequentist risk is $R(b, \hat{b}^{NB}_{\tilde{G}_k}) = \rho \int\int (\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1) - b_1)^2 \phi(\hat{b}_1 - b_1) d\hat{b}_1 d\tilde{G}_K(b_1) =$ $\rho E_{\hat{b}_1|b_1,b_1} (\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1) - b_1)^2 = \rho E_{b_1|\hat{b}_1,\hat{b}_1} (\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1) - b_1)^2 = \rho E_{\hat{b}_1}[\text{var}(b_1 \mid \hat{b}_1)] = \rho E_{\hat{b}_1}[E(b_1^2 \mid \hat{b}_1)] -$ $\rho E_{\hat{b}_1}[\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1)]^2$, where $E_{\hat{b}_1|b_1,b_1}$ denotes expectation with respect to the conditional distribution of $\hat{b}_1$ given b_1 and with respect to the marginal distribution of b_1 (which is $\tilde{G}_K$); the other subscripts on *E* are defined analogously. Thus

$$R(b, \hat{b}^{NB}_{\tilde{G}_k}) = \rho \int b_1^2 d\tilde{G}_K(b_1) - \rho \int (\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1))^2 m(\hat{b}_1; \tilde{G}_K) d\hat{b}_1 . \qquad \text{(A.31)}$$

The Bayes risk, $r_G(\hat{b}^{NB}_{G_k})$, can be written in the same way, with *G* replacing $\tilde{G}_K$. The theorem is proved by showing the convergence of each of the terms on the right hand side of (A.31) to the corresponding term in $r_G(\hat{b}^{NB}_{G_k})$.

For the first term, the assumption $\sup_K \|b\|_{2+\delta} \leq M$ implies that $b^2_{1,K}$ (where $b^2_{1,K}$ has the distribution $\tilde{G}_K$, and we temporarily make the implicit double subscripting explicit) is uniformly integrable along the *K* sequence (Davidson (1994), Theorem 12.10). Because $\tilde{G}_K \Rightarrow G$, $\int b_1^2 d\tilde{G}_K(b_1) \to \int b_1^2 dG(b_1)$ (Davidson (1994), Theorem 22.16).

Now for the second term in (A.31), we first show point-wise convergence of the integrand. For all $x \in \Re$, $b_1\phi(x - b_1)$ and $\phi(x - b_1)$ are uniformly bounded in b_1, so that $\int b_1 \phi(\hat{b}_1 - b_1) d\tilde{G}_K(b_1) \to \int b_1 \phi(\hat{b}_1 - b_1) dG(b_1)$ and $\int \phi(\hat{b}_1 - b_1) d\tilde{G}_K(b_1) \to \int \phi(\hat{b}_1 - b_1) dG(b_1)$. Using continuity and the definition of $\hat{b}^{NB}_G(\bullet)$ and $m(\bullet; \tilde{G}_K)$, $(\hat{b}^{NB}_{\tilde{G}_k}(x))^2 m(x; \tilde{G}_K) \to$ $(\hat{b}^{NB}_G(x))^2 m(x; G)$ for all $x \in \Re$.

To see that the integral converges, note that $E_{\hat{b}_1}[\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1)]^2 = E_{\hat{b}_1}[E(b_1 \mid \hat{b}_1)^2] \leq E(b_1^2)$, so that $\int (\hat{b}^{NB}_{\tilde{G}_k}(\hat{b}_1))^2 m(\hat{b}_1; \tilde{G}_K) d\hat{b}_1 \leq \int b_1^2 d\tilde{G}_K(b_1) \rightarrow \int b_1^2 dG(b_1)$. Convergence of the second term in (A.31) then follows from the dominated convergence theorem in Royden (1988, Chapter 4, Theorem 17).

References

Angrist, J. and A. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979-1014.

Bickel, P., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford: Oxford University Press.

Edelman, D. (1988), "Estimation of the Mixing Distribution for a Normal Mean with Applications to the Compound Decision Problem," *Annals of Statistics* 16, 1609-1622.

George, E.I. (1999), "Comment on *Bayesian Model Averaging*," *Statistical Science* 14, no. 382, 409-412.

George, E.I. and D.P. Foster (2000), "Calibration and Empirical Bayes Variable Selection," manuscript, University of Texas – Austin.

Hall, P. and C.C. Heyde (1980), *Martingale Limit Theory and its Application*. New York: Academic Press.

Härdle, W., J. Hart, J.S. Marron, and A.B. Tsybakov (1992), "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association* 87, 218-226.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science* 14, no. 38, 382-401.

James, W. and C. Stein (1960), "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 361-379.

Knox, T., J.H. Stock and M.W. Watson (2003), "Supplement to 'Empirical Bayes Regression With Many Regressors'," unpublished manuscript.

Lehmann, E.L. and G. Casella (1998), *Theory of Point Estimation, Second Edition*. New York: Springer-Verlag.

Maritz, J.S. and T. Lwin (1989), *Empirical Bayes Methods, Second Edition*. London: Chapman and Hall.

Marron, J.S. and M.P. Wand (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712-736.

Robbins, H. (1955), "An Empirical Bayes Approach to Statistics," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 157-164.

Robbins, H. (1964), "The Empirical Bayes Approach to Statistical Problems," *Annals of Mathematical Statistics*, 35, 1-20.

Royden, H.L. (1988), *Real Analysis*, New York: Macmillan.

Stein, C. (1955), "Inadmissibility of the Usual Estimator for the Mean of Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 197-206.

Stock, J.H. and M.W. Watson (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, Vol. 97, No. 960, pp 1167-1179.

Van der Vaart, A.W. (1988), *Statistical Estimation in Large Parameter Spaces*. Stichting Mathematisch Centrum, Amsterdam.

Table 1
Bayes Estimation Risk of Various Estimators, Relative to OLS

A. $\lambda = 0.5$

	$\rho = 0.25$					$\rho = 0.50$			
	Exact Bayes	**BIC**	**PEB**	**SNEB**		**Exact Bayes**	**BIC**	**PEB**	**SNEB**
	T=100								
Gaussian	0.50	1.00	0.58	0.58		0.50	0.99	0.55	0.56
Bimodal	0.49	1.10	0.58	0.58		0.49	1.10	0.56	0.56
Separated Bimodal	0.47	1.19	0.58	0.58		0.47	1.19	0.56	0.55
Asymmetric Bimodal	0.47	1.09	0.56	0.56		0.47	1.08	0.52	0.53
Kurtotic	0.49	0.88	0.58	0.59		0.49	0.87	0.55	0.56
Outlier	0.27	0.44	0.50	0.49		0.27	0.44	0.51	0.49
	T=200								
Gaussian	0.50	1.00	0.54	0.55		0.50	1.00	0.53	0.54
Bimodal	0.49	1.10	0.54	0.55		0.49	1.10	0.53	0.53
Separated Bimodal	0.47	1.18	0.54	0.54		0.47	1.18	0.53	0.52
Asymmetric Bimodal	0.47	1.08	0.51	0.52		0.47	1.09	0.50	0.51
Kurtotic	0.49	0.88	0.54	0.55		0.49	0.87	0.53	0.54
Outlier	0.27	0.40	0.49	0.47		0.27	0.40	0.50	0.47
	T=400								
Gaussian	0.50	1.00	0.52	0.53		0.50	1.00	0.51	0.53
Bimodal	0.49	1.09	0.52	0.53		0.49	1.09	0.51	0.52
Separated Bimodal	0.47	1.17	0.52	0.52		0.47	1.16	0.51	0.51
Asymmetric Bimodal	0.47	1.08	0.49	0.50		0.47	1.08	0.48	0.49
Kurtotic	0.49	0.88	0.52	0.53		0.49	0.88	0.51	0.52
Outlier	0.27	0.38	0.50	0.46		0.27	0.38	0.50	0.46

Notes: The table gives the Bayes risk of the estimator indicated in the column heading, $r_G(\tilde{\beta})$, relative to the Bayes risk of the OLS estimator, $r_G(\hat{\beta})$, where *G*, the distribution of β is given in the first column. Results are shown for $\lambda = \sigma_b^2/(\sigma_b^2 + \sigma_\varepsilon^2)$. The parameter $\rho = K/T$. The estimators are the Bayes estimator using the true value of *G* (Exact Bayes), BIC models selection over all possible regressions, the parametric Gaussian simple empirical Bayes estimator (PEB), and the nonparametric Gaussian simple empirical Bayes estimator (NSEB). Results for the Exact Bayes estimator are based on an analytic calculation; results for the other estimators are based on 5,000 simulated values.

Table 1
Continued

B. ρ = 0.25, T = 400

	Exact Bayes	BIC	PEB	SNEB
	λ = 0.10			
Gaussian	0.10	0.24	0.13	0.16
Bimodal	0.10	0.24	0.13	0.16
Separated Bimodal	0.10	0.24	0.13	0.16
Asymmetric Bimodal	0.09	0.24	0.12	0.15
Kurtotic	0.10	0.24	0.13	0.16
Outlier	0.09	0.21	0.13	0.15
	λ = 0.25			
Gaussian	0.25	0.48	0.28	0.30
Bimodal	0.25	0.48	0.28	0.30
Separated Bimodal	0.25	0.49	0.28	0.30
Asymmetric Bimodal	0.23	0.49	0.26	0.28
Kurtotic	0.25	0.46	0.28	0.30
Outlier	0.19	0.30	0.27	0.28
	λ = 0.50			
Gaussian	0.50	1.00	0.52	0.53
Bimodal	0.49	1.09	0.52	0.53
Separated Bimodal	0.47	1.17	0.52	0.52
Asymmetric Bimodal	0.47	1.08	0.49	0.50
Kurtotic	0.49	0.88	0.52	0.53
Outlier	0.27	0.38	0.49	0.46
	λ = 0.75			
Gaussian	0.75	1.63	0.76	0.76
Bimodal	0.72	1.97	0.76	0.75
Separated Bimodal	0.58	2.40	0.76	0.68
Asymmetric Bimodal	0.70	1.90	0.74	0.73
Kurtotic	0.71	1.24	0.76	0.75
Outlier	0.38	0.54	0.73	0.62
	λ = 0.90			
Gaussian	0.90	1.83	0.90	0.91
Bimodal	0.86	2.09	0.90	0.89
Separated Bimodal	0.61	2.65	0.90	0.77
Asymmetric Bimodal	0.86	1.88	0.90	0.89
Kurtotic	0.82	1.28	0.90	0.87
Outlier	0.56	0.94	0.88	0.76

C. ρ = 0.25, T = 4000, λ = 0.90

	Exact Bayes	BIC	PEB	SNEB
Gaussian	0.90	2.26	0.90	0.90
Bimodal	0.86	2.70	0.90	0.88
Separated Bimodal	0.61	3.60	0.90	0.70
Asymmetric Bimodal	0.85	2.42	0.89	0.88
Kurtotic	0.82	1.55	0.90	0.84
Outlier	0.56	0.94	0.90	0.68

Table 2
Frequentist Estimation Risk of Various Estimators, Relative to OLS

λ=0.50, T = 400, ρ = 0.25

$$\beta_i = \alpha \times \mathbf{1}(i \le \frac{\zeta}{2} K) - \alpha \times \mathbf{1}(\frac{\zeta}{2} K < i \le \zeta K)$$

ζ	BIC	PEB	NSEB
0.10	0.39	0.52	0.47
0.20	0.78	0.52	0.52
0.30	0.95	0.52	0.53
0.40	1.05	0.52	0.54
0.50	1.10	0.52	0.53
0.60	1.14	0.52	0.53
0.70	1.16	0.52	0.52
0.80	1.18	0.52	0.52
0.90	1.20	0.52	0.52

Notes: The table gives the frequentist risk of the estimator indicated in the column, $R(\tilde{\beta}, \beta)$, relative to the frequentist risk of the OLS estimator, $R(\hat{\beta}, \beta)$. The parameter α is chosen so that $\lambda = \sigma_b^2 / (\sigma_b^2 + \sigma_\varepsilon^2) = 0.5$, and $K = \rho T$. The estimators are BIC model selection over all possible regressions, the parametric Gaussian simple empirical Bayes estimator (PEB), and the nonparametric Gaussian simple empirical Bayes estimator (NSEB).

Table 3
Application: Predicting Education and Earnings

T = 329,509, K = 178

	Education	**Earnings**
F Statistic	1.87	1.12
$\hat{\lambda}$ (95% Conf. Interval)	0.46 (0.35,0.56)	0.11 (0.00, 0.28)
Relative Frequentist Risk		
OLS	1.00	1.00
Restricted	0.91	0.38
BIC	0.90	0.38
PEB	0.63	0.38
NSEB	0.67	0.30

Notes: Results based on (4.1) using the Angrist and Krueger (1991) dataset from the 1980 Census. See the text for description of the variables. The first row is the F statistic for testing the hypothesis that $\beta = 0$. The estimate of λ in the next row is given by $(F-1)/F$, and the 95% confidence interval is obtained from F and the quantiles of the non-central F distribution. The estimation risk is estimated by $\tilde{\sigma}^2_{cv} - \hat{\sigma}^2_{\varepsilon}$ where $\tilde{\sigma}^2_{cv}$ is the leave-one-out cross-validation estimator of the forecast risk and $\hat{\sigma}^2_{\varepsilon}$ is the degrees of freedom adjusted estimator of σ^2_{ε} computed from the OLS residuals. The risk values are relative to the risk of OLS.