Improved Unbiased Estimation Through Partial James–Stein Shrinkage

Jann Spiess

Stanford University

January 4, 2021

Abstract

Shrinkage estimation usually reduces variance at the cost of bias. But when we care only about some parameters of a model, I show that we can reduce variance without incurring bias if we have additional information about the distribution of covariates. In a linear regression model with homoscedastic Normal noise, I consider shrinkage estimation of the nuisance parameters associated with control variables. For at least three control variables and exogenous treatment, I establish that the standard least-squares estimator is dominated with respect to squared-error loss in the treatment effect even among unbiased estimators and even when the target parameter is low-dimensional. I construct the dominating estimator by a variant of James–Stein shrinkage in a high-dimensional Normal-means problem. It can be interpreted as an invariant generalized Bayes estimator with an uninformative (improper) Jeffreys prior in the target parameter.

Jann Spiess, Graduate School of Business, Stanford University, jspiess@stanford.edu. First version: August 21, 2017. I thank Gary Chamberlain, Maximilian Kasy, Carl Morris, and Jim Stock for insightful conversations, and seminar participants at Harvard for helpful comments.

Introduction

Many inference tasks have the following structure: the researcher wants to obtain a high-quality estimate of a set of target parameters (for example, a set of treatment effects in an RCT), but also estimates a number of nuisance parameters she does not care about separately (for example, coefficients on control variables). In these cases, can we reduce variance in the estimation of a target parameter without inducing bias by shrinking in the estimation of possibly high-dimensional nuisance parameters? In a linear regression model with homoscedastic, Normal noise, I show that the answer is yes: a natural application of James–Stein shrinkage to the parameters associated with at least three control variables reduces loss in the possibly low-dimensional treatment effect parameter without producing bias, provided that treatment is random.

In this article I consider how we can improve over linear least-squares estimation of the (treatment) parameter β in the simple linear regression model

$$Y_i = X_i'\beta + W_i'\gamma + U_i$$

with control variables W_i and Normal, homoscedastic noise U_i. As a direct application of the classical James and Stein (1961) shrinkage estimator, the linear-least squares estimator $\hat{\beta}^{\text{OLS}}$ is inadmissible with respect to mean-squared error when β has at least three components. But in addition to adding bias, this result applies only for high-dimensional target parameters, and linear least-squares remains admissible for one or two target parameters and among unbiased estimators.

In this article, I focus on the case where the target parameter itself may be low-dimensional and we may care about unbiased estimation, so that standard shrinkage improvements do not apply. I show that when we have additional information about

the distribution of covariates – such as in the case of a randomized experiment, where treatment X and controls W are orthogonal – we can still improve estimation of the treatment-effect parameter β by employing James–Stein shrinkage in the estimation of the high-dimensional nuisance parameter γ. The resulting estimator first estimates γ by a shrinkage estimator $\hat{\gamma}$, and then estimates β by regressing $Y_i - W_i'\hat{\gamma}$ on X_i. It remains unbiased, while also reducing variance relative to $\hat{\beta}^{\text{OLS}}$.

The estimator considered in this article effectively averages between regression models with and without control variables, similar to the Hansen (2016) model-averaging estimator and coinciding up to a degrees-of-freedom correction with the corresponding Mallows estimator from Hansen (2007). For the specific choice of shrinkage, I contribute four finite-sample properties: First, I note that by averaging over the distribution of controls we obtain dominance of the shrinkage estimator even for low-dimensional target parameters, unlike other available results that require a loss function that is at least three-dimensional. Second, I establish that the resulting estimator remains unbiased under exogeneity of treatment. Third, I conceptualize it as a two-step estimator with a first-stage prediction component. Fourth, I show that the estimator can be interpreted as a natural, invariant generalized Bayes estimator with respect to a partially improper prior corresponding to uninformativeness in the target parameter.

The estimator is also related to Sclove's (1968) partial-shrinkage estimator, which shrinks some of the coefficients in a linear-regression model. Yet this classical literature, going back to Stein (1956) and James and Stein (1961), considers cases where the parameter entering the loss function is itself high-dimensional, while I care only about the loss in a (potentially) low-dimensional target parameter and leverage additional information about the covariate distribution, obtaining improvements in unbiased estimation from shrinking in a high-dimensional nuisance param-

eter. Although an extensive literature has expanded James–Stein shrinkage, I am not aware that the finite-sample results in this article, which combine exogeneity, a low-dimensional target parameter, and a high-dimensional nuisance parameter, were previously available.

The linear regression model is set up in Section 1. Section 2 previews the main inadmissibility result in the univariate case. Section 3 derives a general class of estimators and establishes loss improvement relative to a benchmark OLS estimator provided treatment is exogenous. Section 4 motivates the two-step estimator as an invariant generalized Bayes estimator (with respect to an improper prior) in a suitably transformed many-means problem. Section 5 discusses the properties of the estimator in a simulation exercise.

1 Linear Regression Setup

In this article, I consider estimation of the structural parameter $\beta \in \mathbb{R}^k$ in the canonical linear regression model

$$Y_i = \alpha + X_i'\beta + W_i'\gamma + U_i \tag{1}$$

from n iid observations (Y_i, X_i, W_i), where $X_i \in \mathbb{R}^m$ are the regressors of interest, $W_i \in \mathbb{R}^k$ control variables, and $U_i \in \mathbb{R}$ is homoscedastic, Normal noise. α is an intercept,[1] and γ is a nuisance parameter. To obtain identification of β in Equation 1, I assume that U_i is orthogonal to X_i and W_i (no omitted variables).

Throughout, I write upper-case letters for random variables (such as Y_i) and lower-case letters for fixed values (such as when I condition on $X_i = x_i$). When I suppress indices, I refer to the associated vector or matrix of observations, e.g.

[1] We could alternatively include a constant regressor in X_i and subsume α in β. I choose to treat α separately since I will focus on the loss in estimating β, ignoring the performance in recovering the intercept α.

$Y \in \mathbb{R}^n$ is the vector of outcome variables Y_i and $X \in \mathbb{R}^{n \times m}$ is the matrix with rows X_i'. All results are derived conditional on the value of the regressors X, and extend to the case where X is itself random.

2 Main Inadmissibility Result in the Univariate Case

The main result of this article is that the linear-least squares estimator $\hat{\beta}^{\text{OLS}}$ in Equation 1 is inadmissible among unbiased estimators of β if we assume that treatment X is independent of controls W (as would be the case in an RCT) and controls W are themselves Normally distributed. Specifically, in Section 3, we obtain an estimator based on first-stage shrinkage in the estimation of $\hat{\gamma}$ that implies dominance for the univariate case:

Proposition 1 (Inadmissibility of OLS among unbaised estimators under exogeneity)**.** *For Normally distributed control variables W of dimension $k \geq 3$ that are unrelated to X, $m = 1$ target parameters, and sample size $n \geq k + 3$, there exists an estimator $\hat{\beta}$ with $\text{E}[\hat{\beta}|X{=}x] = \beta$ and $\text{Var}(\hat{\beta}|X{=}x) < \text{Var}(\hat{\beta}^{\text{OLS}}|X{=}x)$.*

The assumption of exogenous treatment is essential for this result, as restricting interest to β would not suffice to break optimality of linear least-squares. Indeed, for one- or two-dimensional estimands, OLS remains admissible if we do not have any information about the distribution of covariates, even relative to biased estimators. The dominating estimator in this proposition works as follows:

1. Residualize the outcome variables Y and the control variables W with respect to the target regressor X (and a constant);

2. From the resulting residuals, estimate the nuisance parameter γ by a James–Stein type shrinkage estimator $\hat{\gamma}$;

3. Estimate β by a regression of $Y - W\hat{\gamma}$ on X (and a constant).

Because we are able to produce a better fit in the second step are we able to provide an estimator with lower variance in the third step. The orthogonality between control and target covariates, along with the Normality assumptions, ensure that the resulting estimator has no bias.

In addition to the "long" linear least-squares regression of Y on (a constant and) X and W, in the exogenous case the "short" regression of Y on (a constant and) X only also provides an unbiased estimator for β. The partial-shrinkage estimator can be interpreted as an interpolation between those two unbiased estimators, where the relative weight is estimated in a way that does not contribute any bias.

While the results here and throughout are formulated conditional on X, they also hold unconditionally, and independently assigned treatment X implies the orthogonality of W we leverage for this result. Conditioning on X and imposing Normality only for W has the advantage that we do not have to restrict the distribution of X, allowing e.g. for binary treatment. Finally, the restriction to univarariate treatment is relaxed in the general result in Theorem 1 below.

3 General Two-Step Partial Shrinkage Estimator

In this section, we derive a two-step estimator for the target (treatment) parameter β based on shrinkage in the nuisance (control) parameter γ. By assumption there are control variables W available with

$$Y|X{=}x, W{=}w \sim \mathcal{N}(\mathbf{1}\alpha + x\beta + w\gamma, \sigma^2\mathbb{I}_n) \tag{2}$$

where σ^2 need not be known. We care about the (possibly high-dimensional) nuisance parameter γ only in so far as it helps us to estimate the (typically low-dimensional) target parameter β, which is our object of interest.

3.1 A canonical form that preserves structure

We first transform Equation 2 into a canonical Normal-means problem. Given $x \in \mathbb{R}^{n\times m}$ and $w \in \mathbb{R}^{n\times k}$, where we assume that $(\mathbf{1}, x, w)$ has full rank $1+m+k \leq n$, let $q = (q_{\mathbf{1}}, q_x, q_w, q_r) \in R^{n\times n}$ orthonormal where $q_{\mathbf{1}} \in \mathbb{R}^n, q_x \in \mathbb{R}^{n\times m}, q_w \in \mathbb{R}^{n\times k}$ such that $\mathbf{1}$ is in the linear subspace of $\mathbb{R}^n$ spanned by $q_{\mathbf{1}} \in \mathbb{R}^n$ (that is, $q_{\mathbf{1}} \in \{\mathbf{1}/\sqrt{n}, -\mathbf{1}/\sqrt{n}\}$), the columns of $(\mathbf{1}, x)$ are in the space spanned by the columns of $(q_{\mathbf{1}}, q_x)$, and the columns of $(\mathbf{1}, x, w)$ are in the space spanned by the columns of $(q_{\mathbf{1}}, q_x, q_w)$. (Such a basis exists, for example, by an iterated singular value decomposition.) Then,

$$Y^* = q'Y | X{=}x, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} q_{\mathbf{1}}'\mathbf{1}\alpha + q_{\mathbf{1}}'x\beta + q_{\mathbf{1}}'w\gamma \\ q_x'x\beta + q_x'w\gamma \\ q_w'w\gamma \\ \mathbf{0}_{n-1-m-k} \end{pmatrix}, \sigma^2\mathbb{I}_n\right).$$

In transforming linear regression to this Normal-means problem, as well as in partitioning the coefficient vector into two groups, for only one of which I will apply shrinkage, I follow Sclove (1968).

Writing Y_x^*, Y_w^*, Y_r^* for the appropriate subvectors of Y^*, we find, in particular, that

$$\begin{pmatrix} Y_x^* \\ Y_w^* \\ Y_r^* \end{pmatrix} | X{=}x, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} \mu_x + a\mu_w \\ \mu_w \\ \mathbf{0}_{n-1-m-k} \end{pmatrix}, \sigma^2\mathbb{I}_{n-1}\right) \tag{3}$$

where $\mu_x = q_x'x\beta \in \mathbb{R}^m$, $\mu_w = q_w'w\gamma \in \mathbb{R}^k$, and $a = q_x'w(q_w'w)^{-1} \in \mathbb{R}^{m\times k}$.[2] The

[2] Alternatively, we could have denoted by μ_x the mean of Y_x^*. However, by separating out μ_x from $a\mu_w$ I feel that the role of μ_w as a relevant nuisance parameter becomes more transparent.

first part of the mean corresponds to a combination of the target parameter μ_x and the nuisance parameter μ_w. The second part corresponds to the nuisance parameter only. The last part collects parts that are uninformative about either parameter.

3.2 Two-step estimator

We now consider two-step estimators that estimate the target parameter μ_x (a transformation of β) in the canonical Normal means problem (Equation 3) by estimating the nuisance parameter μ_w (a transformation of γ) in a first step. Conditional on $X{=}x, W{=}w$ and given an estimator $\hat{\mu}_w = \hat{\mu}_w(Y_w^*, Y_r^*)$ of μ_w, a natural estimator of μ_x is $\hat{\mu}_x = \hat{\mu}_x(Y_x^*, Y_w^*, Y_r^*) = Y_x^* - a\hat{\mu}_w$. An estimator of β is obtained by setting $\hat{\beta} = (q_x' x)^{-1}\hat{\mu}_x$. (The linear least-squares estimator for β is obtained from $\hat{\mu}_w = Y_w^*$.) A natural loss function for $\hat{\beta}$ that represents prediction loss units is the weighted loss $(\hat{\beta} - \beta)'(x' q_x q_x' x)(\hat{\beta} - \beta) = \|\hat{\mu}_x - \mu_x\|^2$. We can therefore focus on the (conditional) expected squared-error loss in estimating μ_x, for which we find

$$\mathrm{E}[\|\hat{\mu}_x - \mu_x\|^2 | X{=}x, W{=}w] = m\sigma^2 + \mathrm{E}[\|\hat{\mu}_w - \mu_w\|_{a'a}^2 | X{=}x, W{=}w]$$

with the seminorm $\|v\|_{a'a} = \sqrt{v'a'av}$ on $\mathbb{R}^k$.

For high-dimensional μ_w ($k \geq 3$), a natural estimator $\hat{\mu}_w$ with low expected squared-error loss is a shrinkage estimator of the form $\hat{\mu}_w = CY_w^*$ with scalar C, such as the James and Stein (1961) estimator for which $C = 1 - \frac{(k-2)\|Y_r^*\|^2}{(n-m-k+1)\|Y_w^*\|^2}$ (or its positive part). While improving with respect to expected squared-error loss ($a'a = \text{const.} \cdot \mathbb{I}_k$), this specific estimator may yield higher (conditional) expected loss in μ_x when the implied loss function for μ_w deviates from squared-error loss ($a'a \neq \text{const.} \cdot \mathbb{I}_k$, i.e. the loss function is not invariant under rotations). We will show below that it is still appropriate in the case of independence of treatment and control.

3.3 From conditional to unconditional loss

So far, we have considered the distribution of the data conditional on the realization of the control variables. We now assume that the control variables are themselves distributed Normally, and derive the (expected) loss in the target parameter. To this end, assume that

$$\text{vec}(W)|X{=}x \sim \mathcal{N}(\text{vec}(\mathbf{1}\alpha_W + x\beta_W), \Sigma_W \otimes \mathbb{I}_n) \tag{4}$$

(that is, $W_i|X{=}x \overset{\text{iid}}{\sim} \mathcal{N}(1\alpha_W + x_i\beta_W, \Sigma_W)$). Here, $\Sigma_W \in \mathbb{R}^{k\times k}$ is symmetric positive-definite (but not necessarily known). $\alpha_W \in \mathbb{R}^{1\times k}, \beta_W \in \mathbb{R}^{m\times k}$ describe the conditional expectation of control variables given the regressors $X{=}x$. The case where x and W are orthogonal ($\beta_W = \mathbb{O}_{m\times k}$) and controls W thus not required for identification will play a special role below.

For conditional inference it is known that the least-squares estimator is admissible for estimating β provided $m \leq 2$ and inadmissible provided $m \geq 3$ no matter what the dimensionality k of the nuisance parameter γ is (James and Stein, 1961), as the rank of the loss function is decisive. The above construction does not provide a counter-example to this result: the rank of $a'a = (w'q_w)^{-1}w'q_xq_x'w(q_w'w)^{-1}$ is at most m, so for $m \leq 2$, $\hat{\mu}_w = Y_w^*$ remains admissible for the loss function on the right. While we could achieve improvements for $m \geq 3$ – through shrinkage in $\hat{\mu}_w$ and/or directly in $\hat{\mu}_x$ – our interest is in the case where m is low and k is high, or when we want to obtain an unbiased estimate of β. Conditional on $X{=}x, W{=}w$ we can thus not hope to achieve improvements that hold for any (β, γ), but we can still hope that shrinkage estimation of μ_w yields better estimates of β on average over draws of W.

In the case of random control variables, I consider the bias and variance proper-

ties of a two-step estimator where the nuisance parameter γ is estimated only from data Y_w^* and Y_r^* in Equation 3, which does not include information about the target parameter μ_x (or equivalently, β). Since the control covariates W are now random, the orthonormal transformation itself may depend on the data. We avoid this complication by assuming that, conditional on $X{=}x$, $(q_{\mathbf{1}}, q_x)$ is deterministic, and fix $q_\perp$ such that $\tilde{q} = (q_{\mathbf{1}}, q_x, q_\perp) \in \mathbb{R}^{n\times n}$ is orthonormal. In that case, $q_x'(Y,W) \perp\!\!\!\perp q_\perp'(Y,W)$. We obtain the following characterization of conditional bias and squared-error loss of the implied estimator $\hat{\beta}$:

Lemma 1 (Properties of the two-step estimator)**.** *For any measurable estimator* $\hat{\gamma} : \mathbb{R}^{n-m-1} \times \mathbb{R}^{n-m-1\times k} \to \mathbb{R}^k$ *with* $\mathrm{E}[\|\hat{\gamma}(q_\perp' Y, q_\perp' W)\|^2] < \infty$*, the estimator* $\hat{\beta}(y,w) = (q_x'x)^{-1}q_x'y - (q_x'x)^{-1}q_x'w\,\hat{\gamma}(q_\perp'y, q_\perp'w)$ *defined for convenience for fixed* x*, has conditional bias*

$$\mathrm{E}[\hat{\beta}(Y,W)|X{=}x] - \beta = -\beta_W(\mathrm{E}[\hat{\gamma}(q_\perp'Y, q_\perp'W)] - \gamma)$$

and expected (prediction-norm) loss

$$\mathrm{E}[\|\hat{\beta}(Y,W) - \beta\|^2_{x'q_xq_x'x}|X{=}x] = m\sigma^2 + \mathrm{E}[\|\hat{\gamma}(q_\perp'Y, q_\perp'W) - \gamma\|^2_\phi]$$

for $\phi = \beta_W' x' q_x q_x' x \beta_W + m\Sigma_W$*, where*

$$\begin{aligned} \mathrm{vec}(q_\perp'W) &\sim \mathcal{N}(\mathbf{0}_{k(n-1-m)}, \Sigma_W \otimes \mathbb{I}_{n-1-m}), \\ q_\perp'Y | q_\perp'W = \tilde{w} &\sim \mathcal{N}(\tilde{w}\gamma, \sigma^2\mathbb{I}_{n-1-m}). \end{aligned}$$

This lemma relates the estimation properties of a two-step estimator $\hat{\beta}$ that first estimates the nuisance parameter γ from the residualized data $q_\perp'(Y,W)$ (which can be obtained by residualizing Y and W with respect to a constant and X) to the

properties of the first-step estimator $\hat{\gamma}$ according to a specific norm. Note that this lemma does not rely on $n \geq 1+m+k$, and indeed generalizes to the case $n > 1+m$ for any $k \geq 1$, including $k > n$.

3.4 Exogenous treatment

In the previous section, we related the estimation properties of the two-step estimator $\hat{\beta}$ to the fit of the nuisance parameter $\hat{\gamma}$. We now consider the special case where treatment is exogenous, and thus $\beta_W = \mathbb{O}_{m\times k}$ in Equation 4. This assumption could be justified, for example, in a randomized trial. Note that in this case in addition to the linear least-squares estimator in the "long" regression that includes controls W another natural unbiased (conditional on $X{=}x$) estimator is available, namely the coefficient $(q_x'x)^{-1}q_x'Y$ in the "short" regression without controls. The "long" and "short" regression represent special (edge) cases in the class of two-step estimators introduced above, which are all unbiased in that sense under the exogeneity assumption:

Corollary 1 (A class of unbiased two-step estimators)**.** *If $\beta_W = \mathbb{O}_{m\times k}$ then for any $\hat{\gamma}$ and $\hat{\beta}$ as in Lemma 1,* $\mathrm{E}[\hat{\beta}(Y,W)|X{=}x] = \beta$*. Furthermore,*

$$\mathrm{E}[\|\hat{\beta}(Y,W) - \beta\|^2_{x'q_xq_x'x}|X{=}x] = m\,\mathrm{E}[(\tilde{Y}_0 - \tilde{W}_0'\hat{\gamma}((\tilde{Y}_i,\tilde{W}_i)_{i=1}^{n-1-m}))^2]$$

for $(\tilde{Y}_i,\tilde{W}_i)_{i=0}^{n-1-m}$ iid with $\tilde{W}_i \sim \mathcal{N}(\mathbf{0}_k,\Sigma_W)$, $\tilde{Y}_i|\tilde{W}_i = \tilde{w}_i \sim \mathcal{N}(\tilde{w}_i'\gamma,\sigma^2)$ (here, $(\tilde{Y}_i,\tilde{W}_i)_{i=1}^{n-1-m}$ is the training sample and $(\tilde{Y}_0,\tilde{W}_0)$ an additional test point drawn from the same distribution).

This corollary clarifies that the class of natural estimators derived above are unbiased conditional on $X{=}x$ (but not necessarily on $X{=}x, W{=}w$ jointly), with expected loss equal to the expected out-of-sample prediction loss in a prediction problem where

the prediction function $\tilde{w}_0 \mapsto \tilde{w}_0'\hat{\gamma}$ is trained on $n-1-m$ iid draws, and evaluated on an additional, independent draw $(\tilde{Y}_0, \tilde{W}_0)$ from the same distribution. The "long" and "short" regressions are included as the special cases $\hat{\gamma}(\tilde{w}, \tilde{y}) = (\tilde{w}'\tilde{w})^{-1}\tilde{w}'\tilde{y}$ and $\hat{\gamma} \equiv \mathbf{0}_k$, respectively.

In the prediction problem implicit to the two-step estimator $\hat{\beta}$, the covariates in the training and test samples follow the same distribution, which suggests an estimator that is invariant to rotations in the corresponding k-means problem. Indeed, using Lemma 1, we can directly leverage a result in Baranchik (1973) that shows that a James–Stein estimator of the form[3]

$$\hat{\mu}_w = \left(1 - \frac{p\|Y_r^*\|^2}{\|Y_w^*\|^2}\right) Y_w^*$$

dominates the OLS estimator in out-of-sample prediction to obtain inadmissibility of the "long" regression:

Theorem 1 (Inadmissibility of OLS among unbiased estimators)**.** *Maintain* $\beta_W = \mathbb{O}_{m\times k}$. *Denote by* $(\hat{\alpha}^{\mathrm{OLS}}, \hat{\beta}^{\mathrm{OLS}}, \hat{\gamma}^{\mathrm{OLS}})$ *the coefficients and by* $\mathrm{SSR} = \|Y - \mathbf{1}\hat{\alpha}^{\mathrm{OLS}} - X\hat{\beta}^{\mathrm{OLS}} - W\hat{\gamma}^{\mathrm{OLS}}\|^2$ *the sum of squared residuals in a linear least-squares regression of* Y *on* $\mathbf{1}$, X, *and* W. *Write* $h = \mathbb{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ *(the annihilator matrix with respect to the intercept). Assume that* $k \geq 3$ *and* $n \geq m + k + 2$. *Then, the two-step estimator* $\hat{\beta} = (X'hX)^{-1}X'h(Y - W\hat{\gamma})$ *with*

$$\hat{\gamma} = \left(1 - \frac{p\,\mathrm{SSR}}{\|\hat{\gamma}^{\mathrm{OLS}}\|^2_{W'h(\mathbb{I}-X(X'hX)^{-1}X')hW}}\right)\hat{\gamma}^{\mathrm{OLS}}$$

where $p \in \left(0, \frac{2(k-2)}{n-m-k+2}\right)$ *is unbiased for* β *given* $X{=}x$ *and dominates* $\hat{\beta}^{\mathrm{OLS}}$ *in the*

[3]The standard James and Stein (1961) estimator (for unknown σ^2) is recovered at $p = \frac{k-2}{n-m-k+1}$.

sense that

$$\mathrm{E}[\|\hat{\beta} - \beta\|^2_{X'hX}|X{=}x] < \mathrm{E}[\|\hat{\beta}^{\mathrm{OLS}} - \beta\|^2_{X'hX}|X{=}x].$$

While the estimator is overall unbiased, it is generally biased conditional on W and X – indeed, the improvement provided by this estimator comes from trading off bias and variance conditional on W and X. Note that the result extends to the positive-part analog for which the shrinkage factor is set to zero whenever the expression is negative. Proposition 1 is an immediate corollary for the case $m = 1$, and the estimator in the general case can be obtained in the same three-step way as outlined below that proposition.

4 Invariance Properties and Bayesian Interpretation

In this section, I provide an alternative motivation of the partial-shrinkage estimator from Theorem 1 as an empirical Bayes estimator, mirroring the Bayesian derivation of the James–Stein estimator in Efron and Morris (1973). Starting with the transformations in Section 3.3, I consider the decision problem of estimating β (equivalently, μ_x). Guided by the treatment of a linear panel-data model in Chamberlain and Moreira (2009), I develop the specific estimator as an invariant empirical Bayes estimator with respect to a partially uninformative (improper) Jeffreys prior.

4.1 Decision problem set-up

First, I set up a formal decision problem in the canonical Normal-means problem. As before, we condition on X throughout and assume that covariates W are Normally distributed given X. Writing $W^*_x = q'_x W, W^*_\perp = q'_\perp W, Y^*_x = q'_x Y, Y^*_\perp = q'_\perp Y$, the

transformation developed in Section 3.3 yields the joint distribution

$$\begin{aligned}\begin{pmatrix} W_x^* \\ W_\perp^* \end{pmatrix} &= \begin{pmatrix} \mu_W \\ \mathbb{O}_{s\times k} \end{pmatrix} + V_W \Sigma_W^{1/2} \qquad (V_W)_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ \begin{pmatrix} Y_x^* \\ Y_\perp^* \end{pmatrix} &= \begin{pmatrix} \mu_x + W_x^* \gamma \\ W_\perp^* \gamma \end{pmatrix} + V_Y \sigma^2 \qquad (V_Y)_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)\end{aligned} \tag{5}$$

where $\Sigma_W^{1/2}$ is the unique symmetric positive-definite square-root of the symmetric positive-definite matrix Σ_W, and V_W and V_Y are independent. Here, in addition to $\mu_x = q_x' x\beta$, also $\mu_W = q_x' x\beta_W$, and $s = n - m - 1$. I write $\mathcal{Z} = \mathbb{R}^{m+s} \times \mathbb{R}^{(m+s)\times k}$ for the sample space from which (Y^*, W^*) is drawn according to this P_θ, where I parametrize $\theta = (\mu_x, \gamma) \in \Theta = \mathbb{R}^m \times \mathbb{R}^k$. (I take $\sigma^2, \Sigma_W, \mu_W$ to be constants.)

Given data from this distribution, the analyst chooses an estimate of μ_x from the action space $\mathcal{A} = \mathbb{R}^m$ to minimize expected loss $L(\theta, a) = \|\mu_x - a\|^2$, which is a loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$. An estimator $\hat{\beta} : \mathcal{Z} \to \mathcal{A}$ from the previous section is a feasible decision rule in this decision problem.

4.2 A group of transformations tied to exogenous treatment

Within this statistical decision problem, I develop a group of transformations that leave distribution and loss invariant in the case of exogenous treatment. Specifically, for an element $g = (g_\mu, g_x, g_W, g_\perp)$ in the (product) group $G = \mathbb{R}^m \times O(m) \times O(k) \times O(s)$, where $\mathbb{R}^m$ denotes the group of real numbers with addition (neutral element 0) and $O(k)$ the group of orthonormal matrices in $\mathbb{R}^{k\times k}$ with matrix multiplication (neutral element $\mathbb{I}_k$), consider the following set of transformations (which are actions of G on $\mathcal{Z}, \Theta, \mathcal{A}$):

– Sample space: $m_{\mathcal{Z}} : G \times \mathcal{Z} \to \mathcal{Z}$,

$$\begin{aligned}&(g, (y_x, y_\perp, w_x, w_\perp))\\ &\mapsto (g_x y_x + g_\mu, g_\perp y_\perp, g_x w_x \Sigma_W^{-1/2} g_W' \Sigma_W^{1/2}, g_\perp w_\perp \Sigma_W^{-1/2} g_W' \Sigma_W^{1/2})\end{aligned}$$

– Parameter space: $m_\Theta : G \times \Theta \to \Theta$,

$$(g, (\mu_x, \gamma)) \mapsto (g_x \mu_x + g_\mu, \Sigma_W^{-1/2} g_W \Sigma_W^{1/2} \gamma)$$

– Action space: $m_{\mathcal{A}} : G \times \mathcal{A} \to \mathcal{A}, (g, a) \mapsto g_x a + g_\mu$

These transformations correspond to translations with respect to the target parameter and orthonormal transformations of target parameter, nuisance parameter, and additional data. For exogenous treatment, they are tied together by leaving model and loss invariant. Indeed, the following is immediate from Equation 5:[4]

Proposition 2 (Invariance of model and loss)**.** *For $\mu_W = \mathbb{O}_{m\times k}$:*

1. *The model is invariant: $m_{\mathcal{Z}}(g, (Y^*, W^*)) \sim \mathrm{P}_{m_\Theta(g,\theta)}$ for all $g \in G$.*
2. *The loss is invariant: $L(m_\Theta(g, \theta), m_{\mathcal{A}}(g, a)) = L(\theta, a)$ for all $g \in G$.*

These invariances make a prior with similar invariances a natural choice for solving the decision problem.

4.3 Interpretation as an invariant empirical Bayes estimator

By Proposition 2, a natural (generalized) Bayes estimator of μ_x is derived from an improper prior on θ that is invariant under the action of G on Θ, as this will yield

[4] Alternatively, we could have treated μ_W as an element of the parameter space and extend the analysis to the case of endogenous treatment. Adding $(g, \mu_W) \mapsto g_x \mu_W \Sigma_W^{-1/2} g_W' \Sigma_W^{1/2}$ to the action on the parameter space would have retained invariance.

a decision rule $d : \mathcal{Z} \to \mathcal{A}$ that is invariant in the sense that $d(m_{\mathcal{Z}}(g,(y,w))) = m_{\mathcal{A}}(g, d((y,w)))$ for all $(g,(y,w)) \in G \times \mathcal{Z}$. This implies for μ_x as an improper prior the Haar measure with respect to the translation action (i.e. up to a multiplicative constant the σ-finite Lebesgue measure on $\mathbb{R}^m$), and for γ a prior that is uniform on ellipsoids $\gamma'\Sigma_W\gamma = \omega$. Taking $\frac{\omega}{\tau^2} \sim \chi^2_m$ with some $\tau > 0$ yields the prior $\gamma \sim \mathcal{N}(\mathbf{0}, \tau^2\Sigma_W^{-1})$. With a product prior for θ, the resulting generalized Bayes estimator for μ_x – which minimizes posterior loss conditional on the data – is

$$\begin{aligned}\mathrm{E}[\mu_x | Y^* = y, W^* = w] &= y_x - w_x \,\mathrm{E}[\gamma | Y^* = y, W^* = w] \\ &= y_x - w_x(w_\perp' w_\perp + \sigma^2\Sigma_W/\tau^2)^{-1} w_\perp' y_\perp.\end{aligned}$$

This Bayes estimator uses the variance matrix Σ_W, which is typically unknown. Replacing Σ_W by the specific invariant sample analog $W_\perp' W_\perp / s$, we obtain the estimator $Y_x^* - \frac{s\tau^2}{s\tau^2+\sigma^2} W_x^*((W_\perp^*)'W_\perp^*)^{-1}(W_\perp^*)'Y_\perp^*$. Similarly assuming that $\gamma \sim \mathcal{N}(\mathbf{0}, s\tau^2 W_\perp' W_\perp)$, an unbiased invariant estimator of $\frac{s\tau^2}{s\tau^2+\sigma^2}$ (given W) is

$$C = 1 - \frac{(Y_\perp^*)'(\mathbb{I}_s - W_\perp^*((W_\perp^*)'W_\perp^*)^{-1}(W_\perp^*)')Y_\perp^*/(s-k)}{(Y_\perp^*)'W_\perp^*((W_\perp^*)'W_\perp^*)^{-1}(W_\perp^*)'Y_\perp^*/(k-2)}.$$

This empirical Bayes estimator corresponds to the estimator from Theorem 1 at $p = \frac{k-2}{s-k} = \frac{k-2}{n-m-k-1}$. By construction, it retains the invariance of the associated generalized Bayes estimator. This is not specific to this value of p:

Proposition 3 (Invariance of estimator)**.** *For any p, the estimator $\hat{\beta}$ from Theorem 1 is invariant with respect to the above actions of G.*

Together, this construction provides an alternative interpretation of the above partial James–Stein estimator: it is a specific empirical Bayes estimator that imposes a proper prior on the nuisance components, while remaining uninformative about the target parameter. The invariances developed in this section make this

uninformativeness precise.

5 Simulation

In this section, I study the performance of the shrinkage estimator introduced in Section 3 in a simulation exercise. I generate data according to Equation 1, where I normalize the variance of the error term to one and the target parameter to $\beta = 1$ (in particular, X_i is uni-dimensional). The (X_i, W_i) are drawn independently from a multivariate standard Normal distribution. I fix the sample size to $n = 80$. I vary the size $\|\gamma\|$ of the control-variable parameter, as well as the number k of controls. On this data, I compare the performance of estimates of β from the short OLS regression (Y_i on X_i and a constant, "short"), the long OLS regression (Y_i on X_i, W_i and a constant, "long"), and the partial-shrinkage estimator introduced in Section 3 ("shrink"). For each parameter setting and estimator I obtain the root mean-squared error from 100,000 Monte-Carlo draws.

$k \to$	5			15			40		
$\|\gamma\| \downarrow$	Short	Long	Shrink	Short	Long	Shrink	Short	Long	Shrink
0.0	0.132	0.138	0.134	0.132	0.154	0.135	0.132	0.243	0.149
0.5	0.148	0.138	0.138	0.148	0.155	0.144	0.148	0.242	0.168
1.0	0.187	0.139	0.139	0.188	0.154	0.150	0.187	0.243	0.198

Table 1: Root mean-squared error in estimating β from 100,000 Monte-Carlo draws for the short linear regression ("short", on constant and X_i), long regression ("long", on constant, X_i, and W_i), and partial-shrinkage estimator ("shrink").

Table 1 reports the results of the simulation exercise for $\|\gamma\| \in \{0.0, 0.5, 1.0\}$ and $k \in \{5, 15, 40\}$. The short, long, and partial-shrinkage estimators are all unbiased. As predicted by the theory, the partial-shrinkage estimator persistently outperforms the long OLS estimator, with higher gains when the control coefficient is small or its dimension high. Unsurprisingly, the short OLS estimator performs better than the

long and partial-shrinkage estimators when the control variables matter very little or the dimension of the nuisance parameters is very large, but even in these cases does the partial-shrinkage estimator still perform comparably well.

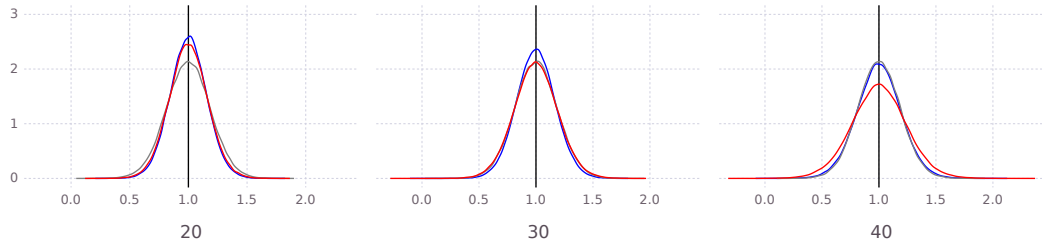

Figure 1: Kernel-density estimates of the distribution of estimates of $\beta = 1$ from 100,000 Monte-Carlo draws for the short linear regression (grey, on constant and X_i), long regression (red, on constant, X_i, and W_i), and partial-shrinkage estimator (blue) for varying dimension $k \in \{20, 30, 40\}$ and $\|\gamma\| = 1$.

For further illustration of this finding, Figure 1 plots Kernel-density estimates of the respective distributions of estimates in the high-dimensional cases $k \in \{20, 30, 40\}$ with $\|\gamma\| = 1$. The partial-shrinkage estimator outperforms the long and the short regression for $k = 20$ and $k = 30$, while still performing comparably well relative to the short regression for $k = 40$ and decisively outperforming the (inadmissible) long regression in that case.

Conclusion

A natural application of James–Stein shrinkage to control variables in a Normal linear model consistently reduces expected prediction error without introducing bias in the treatment parameter of interest provided treatment is random. In this case, the linear least-squares estimator is thus inadmissible even among unbiased estimators.

The results in this article contribute an inadmissibility result for linear regression. The exact finite-sample results come at the cost of realism. Specifically, I

assume Normally distributed control covariates and homoscedastic, Normal error terms. While these can be seen as an approximation, a fully non-parametric approach could involve sample splitting in order to avoid bias and still provide improvement in variance, such as proposed for the analysis of experimental data with binary treatment by Wu and Gagnon-Bartsch (2018), but such a treatment is beyond the scope of the current article.

The case of endogenous treatment offers avenues for future research. When treatment is not exogeneous, Lemma 1 shows that efficient first-stage estimation of the nuisance parameter γ involves minimizing a loss function that depends on the relationship of treatment to controls, adding an additional estimation problem related to the secondary prediction problem in doubly-robust estimation in linear models (Robins et al., 1994; Belloni et al., 2014). It would be interesting to see whether generalizations of the James–Stein estimator to general weight matrices (Bhattacharya, 1966) could improve estimation in that case.

In a companion article (Spiess, 2017), I show how shrinkage in at least four instrumental variables in a canonical structural form provides consistent bias improvement over the two-stage least-squares estimator. Together, these results suggests different roles of overfitting in control and instrumental variable coefficients, respectively: while overfitting to control variables induces variance, overfitting to instrumental variables in the first stage of a two-stage least-squares procedure induces bias. In both cases, James–Stein shrinkage can improve performance.

References

Baranchik, A. J. (1973). Inadmissibility of Maximum Likelihood Estimators in Some Multiple Regression Problems with Three or More Independent Variables. *The Annals of Statistics*, 1(2):312–321.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bhattacharya, P. K. (1966). Estimating the Mean of a Multivariate Normal Population with General Quadratic Loss Function. *The Annals of Mathematical Statistics*, 37(6):1819–1824.

Chamberlain, G. and Moreira, M. J. (2009). Decision Theory Applied to a Linear Panel Data Model. *Econometrica*, 77(1):107–133.

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

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Fourth Berkeley Symposium*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American statistical Association*, 89(427):846–866.

Sclove, S. L. (1968). Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 63(322):596.

Spiess, J. (2017). Bias Reduction in Instrumental Variable Estimation through First-Stage Shrinkage. *arXiv preprint arXiv:1708.06443*.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Third Berkeley Symposium*.

Wu, E. and Gagnon-Bartsch, J. A. (2018). The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation review*, 42(4):458–488.

Appendix

Proof of Lemma 1. Conditional on $W{=}w$, $a\hat{\mu}_w - a\mu_w = q_x'w\hat{\gamma} - q_x'w\gamma$ for $\hat{\gamma} = (q_w'w)^{-1}\hat{\mu}_w$ a function of $q_\perp'w$ and $(Y_w^*, Y_r^*) = (q_w'q_\perp, q_r'q_\perp)(q_\perp'Y)$, so $\hat{\gamma} = \hat{\gamma}(q_\perp'Y, q_\perp'w)$. Assuming measurability, $\hat{\gamma}(q_\perp'Y, q_\perp'W) \perp\!\!\!\perp (q_x'Y, q_x'W)$. Now writing $\hat{\gamma} = \hat{\gamma}(q_\perp'y, q_\perp'w)$ this implies that

$$\mathrm{E}[\|\hat{\mu}_w - \mu_w\|_{a'a}^2 | X{=}x, q_\perp'(Y,W) = q_\perp'(y,w)] = \|\hat{\gamma} - \gamma\|^2_{\mathrm{E}[W'q_xq_x'W|X=x]}$$

with $\mathrm{E}[W'q_xq_x'W|X{=}x] = \beta_W'x'q_xq_x'x\beta_W + m\Sigma_W$ of full rank k. For the expectation of the implied $\hat{\beta}$, we find

$$\mathrm{E}[\hat{\beta}|X{=}x, q_\perp'(Y,W) = q_\perp'(y,w)] = \beta - \beta_W(\hat{\gamma} - \gamma),$$

and the result follows by taking expectations conditional on $X{=}x$. □

Proof of Theorem 1. The OLS estimator in the theorem corresponds to $\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w}) = (\tilde{w}'\tilde{w})^{-1}\tilde{y}'\tilde{w}$ in Lemma 1, which yields the maximum-likelihood estimator $\hat{\gamma}^{\mathrm{OLS}}(q_\perp'Y, q_\perp'W)$

for γ given data

$$\begin{aligned}\text{vec}(q_\perp' W) &\sim \mathcal{N}(\mathbf{0}_{k(n-1-m)}, \Sigma_W \otimes \mathbb{I}_{n-1-m}),\\ q_\perp' Y | q_\perp' W = \tilde{w} &\sim \mathcal{N}(\tilde{w}\gamma, \sigma^2 \mathbb{I}_{n-1-m}).\end{aligned}$$

By Baranchik (1973), this maximum-likelihood estimator is inadmissible with respect to the risk $\mathrm{E}[\|\hat{\gamma} - \gamma\|^2_{\Sigma_W}]$ and thus for $\mathrm{E}[\|\hat{\gamma} - \gamma\|^2_\phi$ in Lemma 1, as $\phi = m\Sigma_W$ for $\beta_W = \mathbb{O}_{m\times k}$. However, Baranchik (1973) also includes an intercept that is estimated, but does not enter the loss function. To formally use the result for our case without intercept in the first-step prediction exercise, I construct an augmented problem such that the dominance result in the augmented problem implies the theorem.

To this end, let

$$\begin{aligned}\text{vec}(W^a) &\sim \mathcal{N}(\mathbf{0}_{k(n-m)}, \Sigma_W \otimes \mathbb{I}_{n-m}),\\ Y^a | W^a = w^a &\sim \mathcal{N}(w^a\gamma, \sigma^2 \mathbb{I}_{n-m}).\end{aligned}$$

(which has one additional sample point, and could without loss include intercepts in W^a, Y^a). By Baranchik (1973, Theorem 1), the estimator

$$\hat{\gamma}^a = \left(1 - p\frac{(Y^a)' h^a Y^a - \|\hat{\gamma}^{a,\mathrm{OLS}}\|^2_{(W^a)' h^a W^a}}{\|\hat{\gamma}^{a,\mathrm{OLS}}\|^2_{(W^a)' h^a W^a}}\right)\hat{\gamma}^{a,\mathrm{OLS}}$$

strictly dominates $\hat{\gamma}^{a,\mathrm{OLS}} = ((W^a)' h^a W^a)^{-1}(W^a)' h^a Z^a$, where $h^a = \mathbb{I}_{n-m} - \mathbf{1}_{n-m}\mathbf{1}_{n-m}'/(n-m)$, in the sense that

$$\mathrm{E}^a[(\hat{\gamma}^a - \gamma)'\Sigma_W(\hat{\gamma}^a - \gamma)] < \mathrm{E}^a[(\hat{\gamma}^{a,\mathrm{OLS}} - \gamma)'\Sigma_W(\hat{\gamma}^{a,\mathrm{OLS}} - \gamma)]$$

for any $\gamma \in \mathbb{R}^k$, provided that $p \in \left(0, \frac{2(k-2)}{n-m-k+2}\right)$ with $k \geq 3$ and $n - m \geq k + 2$.

We now show that this implies dominance of $\hat{\gamma}(q_\perp' Y, q_\perp' W)$ for

$$\hat{\gamma}(\tilde{y}, \tilde{w}) = \left(1 - p\frac{\tilde{y}'\tilde{y} - \|\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w})\|^2_{\tilde{w}'\tilde{w}}}{\|\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w})\|^2_{\tilde{w}'\tilde{w}}}\right)\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w})$$

in the original problem. Let $q^a \in \mathbb{R}^{(n-m)\times(n-m-1)}$ be such that $(q^a, \mathbf{1}_{n-m}/(n-m))$ is orthonormal (that is, the columns of q^a complete $\mathbf{1}_{n-m}/(n-m)$ to an orthonormal basis of $\mathbb{R}^{m-n}$). This implies that $q^a(q^a)' = h^a$ and $(q^a)'q^a = \mathbb{I}_{n-m-1}$. Then, $(q^a)'(Y^a, W^a) \stackrel{d}{=} (q_\perp' Y, q_\perp' W)$. In particular,

$$((Y^a)'h^a(Y^a), (Y^a)'h^a(W^a), (W^a)'h^a(W^a)) \stackrel{d}{=} (q_\perp' Y' q_\perp' Y, q_\perp' Y' q_\perp' W, q_\perp' W' q_\perp' W)$$

and thus $(\hat{\gamma}^a, \hat{\gamma}^{a,\mathrm{OLS}}) \stackrel{d}{=} (\hat{\gamma}(q_\perp' Y, q_\perp' W), \hat{\gamma}^{\mathrm{OLS}}(q_\perp' Y, q_\perp' W))$. We have thus established

$$\begin{aligned}
&\mathrm{E}[(\hat{\gamma}(q_\perp' Y, q_\perp' W) - \gamma)'\Sigma_W(\hat{\gamma}(q_\perp' Y, q_\perp' W) - \gamma)] \\
&< \mathrm{E}[(\hat{\gamma}^{\mathrm{OLS}}(q_\perp' Y, q_\perp' W) - \gamma)'\Sigma_W(\hat{\gamma}^{\mathrm{OLS}}(q_\perp' Y, q_\perp' W) - \gamma)].
\end{aligned}$$

Note that $\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w}) = (\tilde{w}'\tilde{w})^{-1}\tilde{y}'\tilde{w}$ in Lemma 1 does indeed yield $\hat{\gamma}^{\mathrm{OLS}}$ and $\hat{\beta}^{\mathrm{OLS}}$ in the theorem, and that this extends to $\hat{\gamma}$ and $\hat{\beta}$ by

$$\hat{\gamma}(q_\perp' y, q_\perp' w) = \left(1 - p\frac{\|y\|^2_{q_\perp q_\perp'} - \|\hat{\gamma}^{\mathrm{OLS}}(\ldots)\|^2_{w' q_\perp q_\perp' w}}{\|\hat{\gamma}^{\mathrm{OLS}}(\ldots)\|^2_{w' q_\perp q_\perp' w}}\right)\hat{\gamma}^{\mathrm{OLS}}(\ldots)$$

with $q_\perp q_\perp' = h(\mathbb{I} - x(x'hx)^{-1}x')h$ and

$$\begin{aligned}
\mathrm{SSR} &= \|Y - \mathbf{1}\hat{\alpha}^{\mathrm{OLS}} - X\hat{\beta}^{\mathrm{OLS}} - W\hat{\gamma}^{\mathrm{OLS}}\|^2 \\
&= \|Y - W\hat{\gamma}^{\mathrm{OLS}}\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} \\
&= \|Y\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} - \|W\hat{\gamma}^{\mathrm{OLS}}\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} \\
&= \|Y\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} - \|\hat{\gamma}^{\mathrm{OLS}}\|^2_{W'h(\mathbb{I}-X(X'hX)^{-1}X')hW}.
\end{aligned}$$

Unbiasedness and dominance follow with $\beta_W = \mathbb{O}_{m\times k}$ in Lemma 1. $\square$