Confidence regions for averaging estimators

Tom Boot*

July 24, 2020

Abstract

While averaging unrestricted with restricted estimators is known to reduce estimation risk, it is an open question whether this reduction can in turn improve inference. To analyze this question, we construct confidence regions centered at James-Stein averaging estimators in a linear regression model. We show the validity of the regions allowing the number of restrictions on the parameters of interest to increase proportionally with the sample size. When used for hypothesis testing, the recentered confidence regions enable a power enhancement compared to the standard F-test.

1 Introduction

Averaging unrestricted with restricted estimators has been shown to reduce estimation risk for least squares estimators (Hansen, 2014; Liu and Kuo, 2016), two-stage least squares estimators (Hansen, 2017), and for GMM estimators (Cheng et al., 2019). With averaging weights of the form suggested by Stein (1956) and James and Stein (1961), averaging estimators are shown by Hansen (2016) to achieve a local minimax efficiency bound when the number of restrictions on the parameters of interest is large.

These favorable risk properties raise the question whether averaging can also be used to improve inference. To answer this question, we develop

*University of Groningen, t.boot@rug.nl

I thank Bruce Hansen, Artūras Juodis, Peter C.B. Phillips and Tom Wansbeek for helpful comments.

joint confidence regions centered at James-Stein averaging estimators in a homoskedastic linear regression model. Given the known effectiveness of averaging when the number of restrictions is large, we provide asymptotic results allowing the number of restrictions to increase, possibly proportionally, with the sample size.

When used for joint hypothesis testing, we find that suitable restrictions can enhance power compared to the usual F-test. For example, suppose we test a high-dimensional parameter vector which only has a single nonzero element. A t-test on the nonzero element can pick up alternatives of $O(n^{-1/2})$, where n denotes the sample size. However, the standard F-test has trivial power against such alternatives due to the large number of parameters under the test. If we average with a restricted estimator that correctly identifies the location of the nonzero element, we find that the recentered confidence regions restore asymptotic power to that of the t-test on the nonzero element.

Technically, the proposed confidence regions are based on the observation by Stein (1981) that the difference between the mean squared error of the averaging estimator and an unbiased risk estimate satisfies a central limit theorem in the number of parameters of interest. Beran (1995) formalizes this in a set-up where a normally distributed vector is averaged with a fixed vector.

We extend the results by Beran (1995) to a linear regression context by deriving the limiting distribution of the scaled difference between the mean squared error of the averaging estimator and a suitable risk estimate. This extension requires a joint asymptotic limit theory in the sample size and the number of restrictions that are imposed on the parameters of interest. To facilitate this analysis, we turn to the many-instrument literature and adapt a central limit theorem by Chao et al. (2012).

We numerically analyze the confidence regions in a setting where a researcher has a primary variable of interest in mind, and performs a joint test including a number of secondary variables that are expected to have small effects. The coverage rate of the developed confidence regions is close to nominal, both when the number of restrictions is small and large relative to the sample size. Conform the theory, we observe substantial power improvements over a standard F-test, especially when the number

of parameters under the test increases.

Related literature Recentered confidence regions for multiple parameters have been discussed for the case where the restricted estimator is a fixed vector. Casella and Hwang (2012) provide an overview of the literature on recentered confidence regions. If the same radius is used as for the standard confidence region, Casella and Hwang (1982) prove that recentering increases the coverage rate under known variance, and Hwang and Ullah (1994) for the unknown variance case. Confidence sets with reduced volume are developed for example by Casella and Hwang (1983) and Samworth (2005). In our numerical evaluation, we find these confidence regions to be conservative, especially when the number of parameters increases.

Confidence intervals for individual parameters after model averaging are proposed by Hjort and Claeskens (2003). Based on this suggestion, Liu (2015) develops confidence intervals for the Mallows model averaging estimator of Hansen (2007) and the jackknife model averaging estimator of Hansen and Racine (2012). Simulation-based approaches are considered by Claeskens and Hjort (2008), DiTraglia (2016) and Zhang and Liu (2019). Leeb and Kabaila (2017) show that for one-dimensional intervals, length reductions do not occur uniformly over the parameter space.

Organization This paper is structured as follows. Section 2 introduces the model, defines the averaging estimator and the construction of the confidence regions. The theoretical validity of the confidence regions is discussed in Section 3. Section 4 provides numerical evidence for the coverage rate and power properties of associated hypothesis tests. Section 5 concludes.

2 Averaging estimators and confidence regions

Consider the homoskedastic linear regression model

$$y_i = \boldsymbol{x}_{i,k}'\boldsymbol{\theta}_k + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{x}_{i,k} \in \mathbb{R}^{k\times 1}$. We define $\boldsymbol{y}_n = (y_1, \ldots, y_n)'$, $\boldsymbol{X}_{n,k} = (\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{n,k})'$ and $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$. Throughout, we subscript vectors and matrices by their respective dimensions.

The goal of the paper is to construct a confidence region for the parameter vector of interest $\boldsymbol{\beta}_p = \boldsymbol{G}'_{k,p}\boldsymbol{\theta}_k$, where $\boldsymbol{G}_{k,p} \in \mathbb{R}^{k\times p}$ and $p \leq k$. The leading case is where $\boldsymbol{G}_{k,p}$ selects a subset of parameters of interest from $\boldsymbol{\theta}_k$. We also define a set of restrictions on the parameters of the model by

$$\boldsymbol{R}'_{k,r}\boldsymbol{\theta}_k = \boldsymbol{c}_r, \qquad \boldsymbol{R}_{k,r} \in \mathbb{R}^{k\times r}. \tag{2}$$

The number of restrictions equals $\text{rank}(\boldsymbol{R}_{k,r}) = r$. We assume that these restrictions are imposed directly on the parameters of interest, in the sense that

$$\boldsymbol{R}_{k,r} = \boldsymbol{G}_{k,p}\boldsymbol{B}_{p,r}, \quad r \leq p. \tag{3}$$

2.1 Estimators

The averaging estimator for $\boldsymbol{\beta}_p$ is a linear combination of an unrestricted estimator $\hat{\boldsymbol{\beta}}_p$ and a restricted estimator $\tilde{\boldsymbol{\beta}}_p$,

$$\hat{\boldsymbol{\beta}}^a_p = \hat{\omega}\tilde{\boldsymbol{\beta}}_p + (1-\hat{\omega})\hat{\boldsymbol{\beta}}_p. \tag{4}$$

To obtain $\hat{\boldsymbol{\beta}}_p$ and $\tilde{\boldsymbol{\beta}}_p$, we first estimate $\boldsymbol{\theta}_k$ without and with imposing (2),

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k &= (\boldsymbol{X}'_{n,k}\boldsymbol{X}_{n,k})^{-1}\boldsymbol{X}'_{n,k}\boldsymbol{y}_n, \\ \tilde{\boldsymbol{\theta}}_k &= \hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\Sigma}}_{\theta,k}\boldsymbol{R}_{k,r}(\boldsymbol{R}'_{k,r}\hat{\boldsymbol{\Sigma}}_{\theta,k}\boldsymbol{R}_{k,r})^{-1}(\boldsymbol{R}'_{k,r}\hat{\boldsymbol{\theta}}_k - \boldsymbol{c}_r), \end{aligned} \tag{5}$$

where,

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\theta,k} &= \hat{\sigma}^2\left(n^{-1}\boldsymbol{X}'_{n,k}\boldsymbol{X}_{n,k}\right)^{-1}, \\ \hat{\sigma}^2 &= \frac{1}{n-k}\boldsymbol{y}'_n\boldsymbol{M}_{X_{n,k}}\boldsymbol{y}_n, \quad \boldsymbol{M}_{X_{n,k}} = \boldsymbol{I}_n - \boldsymbol{X}_{n,k}(\boldsymbol{X}'_{n,k}\boldsymbol{X}_{n,k})^{-1}\boldsymbol{X}'_{n,k}. \end{aligned} \tag{6}$$

We then have the following estimators for the parameters of interest $\boldsymbol{\beta}_p$,

$$\hat{\boldsymbol{\beta}}_p = \boldsymbol{G}'_{k,p}\hat{\boldsymbol{\theta}}_k, \quad \tilde{\boldsymbol{\beta}}_p = \boldsymbol{G}'_{k,p}\tilde{\boldsymbol{\theta}}_k. \tag{7}$$

We will later see that for hypothesis testing it can be beneficial to add a fixed vector to $\tilde{\boldsymbol{\theta}}_k$. Since the vector is fixed, this can be done without affecting our results.

We consider averaging weights $\hat{\omega}$ in (4) closely related to the shrinkage factor of James and Stein (1961). They can be expressed in terms of the standard F-statistic as

$$\hat{\omega} = \frac{r-2}{r}\frac{1}{\hat{F}}, \quad \hat{F} = \frac{n(\hat{\boldsymbol{\beta}}_p - \tilde{\boldsymbol{\beta}}_p)'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\hat{\boldsymbol{\beta}}_p - \tilde{\boldsymbol{\beta}}_p)}{r}, \quad \hat{\boldsymbol{\Sigma}}_{\beta,p} = \hat{\sigma}^2 \boldsymbol{G}_{k,p}' \hat{\boldsymbol{\Sigma}}_{\theta,k} \boldsymbol{G}_{k,p}. \tag{8}$$

The inverse F-statistic emphasizes that the weight on the restricted estimator is large when there is no clear evidence to reject the restrictions. The weights aim to minimize estimation risk $\rho(\bar{\boldsymbol{\beta}}_p, \boldsymbol{\beta}_p)$, defined as

$$\rho(\bar{\boldsymbol{\beta}}_p, \boldsymbol{\beta}_p) = \mathrm{E}[\ell(\bar{\boldsymbol{\beta}}_p, \boldsymbol{\beta}_p)], \quad \ell(\bar{\boldsymbol{\beta}}_p, \boldsymbol{\beta}_p) = n(\bar{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\bar{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p). \tag{9}$$

Hansen (2016) shows that the averaging estimator (4) with weights (8) is locally asymptotically minimax efficient.

2.2 Confidence regions

The confidence regions we consider are based on the following scaled difference between the loss of the averaging estimator and an estimator for its risk,

$$D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) = p^{-1/2}\left[\ell(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) - \hat{\rho}(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)\right], \tag{10}$$

with the loss $\ell(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$ as defined in (9) and

$$\hat{\rho}(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) = p - (r-2)\hat{\omega}. \tag{11}$$

The risk estimator is motivated in Section A.1.1, where we show that under normally distributed errors, we have $\mathrm{E}[\hat{\rho}(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)] = \rho(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$.

Confidence regions for $\boldsymbol{\beta}_p$ follow from Theorem 1 below, which shows that (10) has the limiting distribution $N(0, \tau^2)$. When a consistent estimator $\hat{\tau}^2$ is available for τ^2, confidence regions with coverage rate $1-\alpha$ are

readily constructed as

$$C(\hat{\boldsymbol{\beta}}_p^a) = \left\{ \boldsymbol{t} \colon D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{t}) \leq \Phi^{-1}(1-\alpha)\hat{\tau} \right\}. \tag{12}$$

The asymptotic variance τ^2 is discussed in detail in Section 3. Here we only provide the estimator that is used to operationalize (12), which is

$$\hat{\tau}^2 = 2 + 2\frac{p}{n+k} - 4\frac{(r-2)^2}{pr}\frac{\hat{\lambda}^2}{(\hat{\lambda}^2+1)^2} - 8\frac{r}{n+k}\frac{1}{\hat{\lambda}^2+1}\left(1 - \frac{r}{p}\frac{1}{\hat{\lambda}^2+1}\right), \tag{13}$$

where $\hat{\lambda}^2$ estimates the non-centrality parameter of the F-test appearing in (8) as

$$\hat{\lambda}^2 = \max\left(0, \hat{F} - 1\right). \tag{14}$$

The trimming is suggested by Beran (1995) to ensure that the noncentrality parameter estimate is positive. It will not affect the asymptotic results, as $\hat{F} - 1$ converges in probability to a nonnegative constant.

The factor $(r-2)^2$ in (13) is motivated by the variance of $\hat{D}(\hat{\boldsymbol{\beta}}_p^a, \hat{\boldsymbol{\beta}}_p)$ for finite r under exact normality of the errors as derived in Section A.1.2.

2.3 Geometric intuition

The averaging weights (8) are selected to achieve a low risk (9). Figure 1 displays the parameter vectors $\boldsymbol{\beta}_p$, $\hat{\boldsymbol{\beta}}_p$, and $\tilde{\boldsymbol{\beta}}_p$ where the subscript s indicates that they are rescaled by $(n^{-1}\hat{\boldsymbol{\Sigma}}_{\beta,p})^{-1/2}$. The averaging estimator $\hat{\boldsymbol{\beta}}_{p,s}^a$ closest to $\boldsymbol{\beta}_{p,s}$ is given by the orthogonal projection of $\boldsymbol{\beta}_{p,s}$ on the line segment joining $\hat{\boldsymbol{\beta}}_{p,s}$ and $\tilde{\boldsymbol{\beta}}_{p,s}$. Defining $\hat{\boldsymbol{\delta}}_{p,s} = \hat{\boldsymbol{\beta}}_{p,s} - \tilde{\boldsymbol{\beta}}_{p,s}$, this suggests

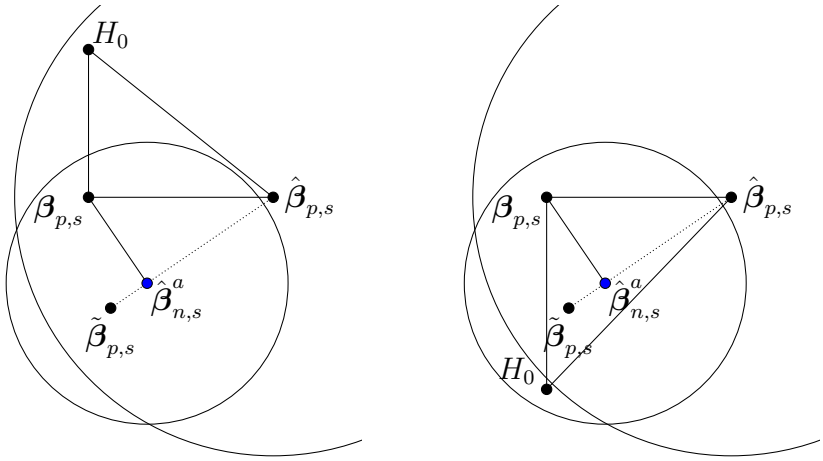
$$\hat{\boldsymbol{\beta}}_{p,s}^a = \tilde{\boldsymbol{\beta}}_{p,s} + \frac{\hat{\boldsymbol{\delta}}_{p,s}'(\boldsymbol{\beta}_s - \tilde{\boldsymbol{\beta}}_{p,s})}{\hat{\boldsymbol{\delta}}_{n,s}'\hat{\boldsymbol{\delta}}_{p,s}}\hat{\boldsymbol{\delta}}_{p,s} = \tilde{\boldsymbol{\beta}}_{p,s} + \left[1 - \frac{\hat{\boldsymbol{\delta}}_{p,s}'(\hat{\boldsymbol{\beta}}_{p,s} - \boldsymbol{\beta}_{p,s})}{\hat{\boldsymbol{\delta}}_{p,s}'\hat{\boldsymbol{\delta}}_{p,s}}\right]\hat{\boldsymbol{\delta}}_{p,s}. \tag{15}$$

Multiplying from the left with $(n^{-1}\hat{\boldsymbol{\Sigma}}_{\beta,p})^{\frac{1}{2}}$, we get the averaging estimator

$$\hat{\boldsymbol{\beta}}_p^a = \tilde{\boldsymbol{\beta}}_p + \frac{n\hat{\boldsymbol{\delta}}_p'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\boldsymbol{\beta}_p - \tilde{\boldsymbol{\beta}}_p)}{n\hat{\boldsymbol{\delta}}_p'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\hat{\boldsymbol{\delta}}_p}\hat{\boldsymbol{\delta}}_p = \tilde{\boldsymbol{\beta}}_p + \left[1 - \frac{n\hat{\boldsymbol{\delta}}_p'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)}{n\hat{\boldsymbol{\delta}}_p'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\hat{\boldsymbol{\delta}}_p}\right]\hat{\boldsymbol{\delta}}_p. \tag{16}$$

where $\hat{\boldsymbol{\delta}}_p = \hat{\boldsymbol{\beta}}_p - \tilde{\boldsymbol{\beta}}_p$.

Figure 1: Power resulting from recentered confidence regions



Note: $\boldsymbol{\beta}_{p,s} = (n^{-1}\hat{\boldsymbol{\Sigma}}_{\beta,p})^{-\frac{1}{2}}\boldsymbol{\beta}_p$, and similar for the other vectors. H_0 denotes the parameter vector under the null hypothesis.

The denominator equals that in the averaging weights (8). In the numerator, $\mathrm{E}[n\hat{\boldsymbol{\delta}}_p'\hat{\boldsymbol{\Sigma}}^{-1}_{\beta,p}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)|\boldsymbol{X}_{n,k}] = \mathrm{tr}[n\hat{\boldsymbol{\Sigma}}^{-1}_{\beta,p}\mathrm{cov}(\hat{\boldsymbol{\beta}}_p, \hat{\boldsymbol{\delta}}_p|\boldsymbol{X}_{n,k})] = r$, corresponding to the leading term in the numerator of (8). As such, the weights (8) estimate the projection that minimizes the loss $\ell(\hat{\boldsymbol{\beta}}^a_p, \boldsymbol{\beta}_p)$.

Figure 1 also shows a particular realization of a confidence region centered at the unrestricted estimator $\hat{\boldsymbol{\beta}}_{p,s}$ given by the larger circle, and one recentered at the averaging estimator $\hat{\boldsymbol{\beta}}^a_{p,s}$ given by the smaller circle. While the reduction in volume appears attractive, Efron (2006) points out that this reduction will not necessarily improve the power of corresponding tests. We illustrate this by comparing both panels of Figure 1. On the left, the restricted estimator $\tilde{\boldsymbol{\beta}}_{p,s}$ is further away from the null hypothesis than the true parameter vector $\boldsymbol{\beta}_{p,s}$. In this case, recentering shifts the confidence region away from the parameter vector under the null and we gain power against H_0. On the right, the restricted estimator is close to the parameter vector under the null. The recentered confidence region now does not reject the null, while the standard confidence region would. The choice of the restrictions is studied in Section 3.2.

3 Theoretical results

3.1 Assumptions

Let $M > 0$ denote a generic finite constant that can differ across equations. We follow (Chao et al., 2012) in writing *a.s.* for almost surely and *a.s.n.* for almost surely for n large enough.[1]

Assumption 1 *For all n, $k/n \leq \bar{\kappa} < 1$. As $(r, n) \to \infty$, (a) $\frac{k}{n} \to \kappa$, (b) $\frac{p}{n} \to \pi$, (c) $\frac{r}{n} \to \rho$ with $\kappa \in [0, 1)$ and $(\pi, \rho) \in [0, 1] \times [0, 1]$.*

The first part restricts the number of parameters in the model to be strictly smaller than the number of observations. Assumption 1 provides the rate conditions, which allow for the number of parameters in the model (k), the number of parameters of interest (p), and the number of restrictions imposed on the parameters of interest (r) to increase possibly proportionally to the sample size (n). Note that we write $(r, n) \to \infty$, but as $r \leq p \leq k$, when $r \to \infty$ also $(p, k) \to \infty$.

Assumption 2 *Conditional on $\boldsymbol{X}_{n,k}$, $\{\varepsilon_i\}$ is an independent sequence with $\mathrm{E}[\varepsilon_i|\boldsymbol{X}_{n,k}] = 0$, $\mathrm{E}[\varepsilon_i^2|\boldsymbol{X}_{n,k}] = \sigma^2$, $\mathrm{E}[\varepsilon_i^4|\boldsymbol{X}_{n,k}] = \mathrm{E}[\varepsilon_i^4] \leq M < \infty$.*

Assumption 2 specifies relatively standard assumptions on the error terms. The assumption that $\{\varepsilon_i\}$ is independent conditional on $\boldsymbol{X}_{n,k}$ is the same as in Chao et al. (2012) where the errors are independent conditional on the set of instruments. Notice that we assume conditional homoskedasticity. Extending the results to heteroskedastic models as in Cattaneo et al. (2018) and Anatolyev and Sølvsten (2020) would require significant extensions to the proofs.

Assumption 3 *The regressors and restrictions satisfy the following: (a) Denote by $\mu_{n,k}^{(1)}, \ldots, \mu_{n,k}^{(k)}$ the eigenvalues of $n^{-1}\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k}$ sorted in decreasing order. There exist finite positive constants m and M such that $m \leq \mu_{n,k}^{(k)} \leq \mu_{n,k}^{(1)} \leq M$ a.s.n.*

[1] An event E_n occurs *a.s.n.* if $\mathrm{P}(\exists N \colon \forall n \geq N, E_n) = 1$. Suppose $E_n = |X_n - X| < \epsilon$ for a sequence of random variables X_n. Then if, for all $\epsilon > 0$, E_n occurs *a.s.n.*, we have that $X_n \to_{a.s.} X$. However, *a.s.n.* is slightly weaker, as we can also consider events $E_n = |X_n| < M$ for some $M > 0$.

(b) The restrictions are such that $\boldsymbol{R}_{k,r}'\boldsymbol{\theta}_k - \boldsymbol{c}_r = n^{-1/2}\boldsymbol{R}_{k,r}'\boldsymbol{h}_k$.
(c) Define $\bar{\lambda}^2 = r^{-1}\sigma^{-2}\boldsymbol{h}_k'\boldsymbol{R}_{k,r}(\boldsymbol{R}_{k,r}'(n^{-1}\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k})^{-1}\boldsymbol{R}_{k,r})^{-1}\boldsymbol{R}_{k,r}'\boldsymbol{h}_k$. *Then,*

$$\bar{\lambda}^2 \to_{a.s.} \lambda^2, \quad \text{where } 0 \leq \lambda^2 < \infty. \tag{17}$$

Assumption 3 specifies the properties of the regressors and the imposed restrictions. Part (a) guarantees that the inverse of $n^{-1}\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k}$ exists almost surely for sufficiently large n. Part (b) ensures that the misspecification bias induced by the imposed restrictions is local-to-zero. This excludes the trivial case where all weight is placed on the unrestricted estimator. Part (c) ensures almost sure convergence of the noncentrality parameter of the F-test.

Finally, we impose the following convergence results.

Assumption 4 *Under Assumption 1, we have the following.*
(a) Define $\boldsymbol{S}_k = \boldsymbol{R}_{k,r}(\boldsymbol{R}_{k,r}'(\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k})^{-1}\boldsymbol{R}_{k,r})^{-1}\boldsymbol{R}_{k,r}'(\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k})^{-1}$, *and let* $\boldsymbol{h}_k$ *be as in Assumption 3. Then,*

$$\frac{1}{n^2}\sum_{i=1}^{n} |r^{-1/2}\boldsymbol{h}_k'\boldsymbol{S}_k\boldsymbol{x}_{i,k}|^4 \to_{a.s.} 0. \tag{18}$$

(b) Let $\boldsymbol{P}[\boldsymbol{A}] = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$. *Define* $\boldsymbol{P}_{G,n} = \boldsymbol{P}[\boldsymbol{X}_{n,k}(\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k})^{-1}\boldsymbol{G}_{k,p}]$, $\boldsymbol{P}_{R,n} = \boldsymbol{P}[\boldsymbol{X}_{n,k}(\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k})^{-1}\boldsymbol{R}_{k,r}]$ *and* $\boldsymbol{M}_{X_{n,k}} = \boldsymbol{I}_n - \boldsymbol{P}[\boldsymbol{X}_{n,k}]$. *Then,*

$$\begin{aligned}
\frac{1}{p}\sum_{i=1}^{n}([\boldsymbol{P}_{G,n}]_{ii} - \pi)^2 &\to_{a.s.} 0, \\
\frac{1}{p}\sum_{i=1}^{n}([\boldsymbol{P}_{R,n}]_{ii} - \rho)^2 &\to_{a.s.} 0, \\
\frac{1}{n}\sum_{i=1}^{n}([\boldsymbol{M}_{X_{n,k}}]_{ii} - (1-\kappa))^2 &\to_{a.s.} 0.
\end{aligned} \tag{19}$$

Assumption 4 provides technical conditions required for the central limit theorem we invoke. To gain intuition for part (a), it is helpful to consider the case where restrictions are imposed on all parameters in the model, i.e. $\boldsymbol{R}_{k,r} = \boldsymbol{I}_k$. In this case $\boldsymbol{S}_k = \boldsymbol{I}_k$. Moreover, suppose $\boldsymbol{h}_k$ is a vector of ones. Then, part (a) reduces to $\frac{1}{n^2}\sum_{i=1}^{n}\left|\frac{1}{\sqrt{k}}\sum_{j=1}^{k}x_{ij}\right|^4 \to_{a.s.} 0$, where $x_{ij} = [\boldsymbol{x}_{i,k}]_j$. If x_{ij} is independent across i and j with finite fourth moment,

then part (a) follows from the strong law of large numbers.

Part (b) places a condition on the convergence of the diagonal elements of projection matrices. This is similar to the assumptions used in Anderson et al. (2010) and Kunitomo (2012), who require the convergence in (19) to hold in probability. A slightly more stringent version appears in Anatolyev (2012) who assumes that for fixed regressors $\max_{i=1,\dots,n} |[\boldsymbol{P}_{G,n}]_{ii} - \pi| \to 0$. Convergence of the diagonal elements of projection matrices in high-dimensional models is discussed in depth in Anatolyev and Yaskov (2017).

3.2 Limiting distribution and power

The following theorem is the main result of this paper.

Theorem 1 *Under Assumption 1–4,*

$$D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) \Rightarrow N(0, \tau^2), \tag{20}$$

where

$$\tau^2 = 2\left(1 + \frac{\pi}{1-\kappa}\right) - 4\frac{\rho}{\pi}\frac{\lambda^2}{(\lambda^2+1)^2} - 8\frac{\rho}{1-\kappa}\frac{1}{\lambda^2+1}\left(1 - \frac{\rho}{\pi}\frac{1}{\lambda^2+1}\right), \tag{21}$$

with λ^2 defined in Assumption 3 and (κ, π, ρ) defined in Assumption 1. The asymptotic variance τ^2 is consistently estimated by (13).

The proof is provided in Appendix A.2. The key underlying result is an adaptation of Lemma A2 from Chao et al. (2012) to the current setting.

The benchmark to which we can compare τ^2 is the variance of the (rescaled) F-test on the p parameters of interest provided by Anatolyev (2012). This variance equals $\tau^2 = 2(1 + \frac{\pi}{1-\kappa})$. In a low-dimensional setting where $\pi \to 0$, this reduces to $\tau^2 = 2$. From (21), we see that averaging leads to a lower asymptotic variance by introducing two negative terms. The latter of these two terms is only present in the high-dimensional regime where $\rho \not\to 0$.

To gain insight in the effect of the negative terms, suppose that $\rho = \pi$, so that we impose restrictions on all parameters of interest, and $\lambda^2 = 1$, so that we have mild misspecification bias. In this case, $\tau^2 = 1$ and, even in a

high-dimensional set-up, the variance no longer depends on the dimensions (k, p, r).

One might expect the largest gains when the imposed restrictions are correct, but this is not the case. When $\lambda^2 = 0$, then $\tau^2 = 2(1 + \frac{\pi}{1-\kappa}) - 8\frac{\rho}{1-\kappa}(1 - \frac{\rho}{\pi})$. In this case, the lowest possible variance arises when $\rho = \frac{1}{2}\pi$, in which case $\tau^2 = 2$. We see that imposing correct restrictions will not reduce the variance in the low-dimensional case where $(\pi, \rho) \to 0$ compared to the standard F-test. However, the high-dimensional limit distribution reveals that it reduces the variance compared to the standard F-test.

Although a reduction in the variance of the limiting distribution has effects on power, the discussion in Section 2.3 shows that power is also affected by recentering the confidence region from $\hat{\boldsymbol{\beta}}_p$ to $\hat{\boldsymbol{\beta}}_p^a$. Denote by $\boldsymbol{\beta}_0$ the value of the parameter vector under the null hypothesis. We have the following result on the power of a test based on (10).

Theorem 2 *Suppose* $\boldsymbol{\beta}_p - \boldsymbol{\beta}_0 = p^{-\gamma}n^{-1/2}\boldsymbol{h}_{0,p}$ *for some* $\gamma > 0$, *and* $\boldsymbol{\beta}_p - \mathrm{E}[\tilde{\boldsymbol{\beta}}_p] = n^{-1/2}\boldsymbol{h}_p$, *with* $\boldsymbol{h}_{0,p}$ *and* $\boldsymbol{h}_p$ *satisfying Assumption 3. Suppose* γ *is such that under Assumption 1,*

$$p^{-1/2-2\gamma}\boldsymbol{h}_{0,p}'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\boldsymbol{h}_{0,p} - 2p^{-1/2-\gamma}\hat{\omega}\boldsymbol{h}_{0,p}'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\boldsymbol{h}_p \to_p \Delta < \infty \tag{22}$$

Then, with τ^2 *defined in Theorem 1,*

$$D(\hat{\boldsymbol{\beta}}^c, \boldsymbol{\beta}) \Rightarrow N(\Delta, \tau^2). \tag{23}$$

The proof is in Appendix A.3.

The first term of (22) is the noncentrality parameter from the standard F-test. It is finite when $\gamma = 1/4$. As such, local alternatives are of $O(n^{-1/2}p^{-1/4})$, which coincides with Anatolyev (2012). However, if $p^{-1}\boldsymbol{h}_{0,p}'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\boldsymbol{h}_p \to_p \Delta_2 < 0$, which occurs for example when $\boldsymbol{h}_p = b \cdot \boldsymbol{h}_{0,p}$ for some $b < 0$, then we can set $\gamma = 1/2$. As such, we have power against local alternatives of $O(n^{-1/2}p^{-1/2})$ instead of the usual $O(n^{-1/2}p^{-1/4})$.

To highlight how to select a restricted estimator, consider the following setting. Denote by $\boldsymbol{e}_{p,1} = (1, 0, \ldots, 0)'$. Suppose $\boldsymbol{\beta}_0 = \mathbf{0}_p$ and the unknown parameter vector of interest $\boldsymbol{\beta}_p = n^{-1/2}\sigma_{11} \cdot d \cdot \boldsymbol{e}_{p,1}$, where $\sigma_{11}^2 = [\boldsymbol{\Sigma}_{\beta,p}]_{1,1}$. When $\gamma = 1/2$, $\boldsymbol{h}_{0,p} = p^{1/2}\sigma_{11} \cdot d \cdot \boldsymbol{e}_{p,1}$. Note that in this set-up, a t-test on

the nonzero coefficient would converge to $N(d,1)$.

Suppose a researcher has a (correct) prior belief which coefficient is nonzero and what the sign of this coefficient is. Denote by $\tilde{\boldsymbol{\beta}}_p$ the estimator that imposes the exclusion restrictions, and define the restricted estimator

$$\bar{\boldsymbol{\beta}}_p = \tilde{\boldsymbol{\beta}}_p + \left(\frac{p}{n}\right)^{1/2} s_{11} \cdot \bar{d} \cdot \boldsymbol{e}_{p,1}, \tag{24}$$

where $s_{11}^2 = [\hat{\boldsymbol{\Sigma}}_{\beta,p}]_{1,1}$ and we assume that $s_{11}^2 \to_p \sigma_{11}^2$. Since $\mathrm{E}[\tilde{\boldsymbol{\beta}}_p] = \boldsymbol{\beta}_p$, we have $\boldsymbol{h}_p = -p^{1/2} s_{11} \cdot \bar{d} \cdot \boldsymbol{e}_{p,1}$. Also, the averaging weight $\hat{\omega} \to_p (\bar{d}^2 + 1)^{-1}$.

Substituting the foregoing expressions into (22), the non-centrality parameter $\Delta = 2d(\bar{d}^2 + 1)^{-1}\bar{d}$. This is maximized for $\bar{d} = 1$, in which case $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) \Rightarrow N(d,1)$. So if the researcher has correct prior beliefs, the asymptotic distribution coincides with that of a t-test on the nonzero coefficient in the unrestricted model. The F-test however has noncentrality parameter $p^{-1/2}d \to 0$ and only has trivial power.

What if the researcher is wrong about the sign of the coefficient? In this case, $\bar{d} = -1$, and $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) \Rightarrow N(-d,1)$. If we think of the test as an F-test, then we only reject when $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$ is large and positive, i.e. we perform a one-sided test. In this case, choosing the wrong sign implies a loss of power. However, the asymptotic results suggest that this can be circumvented by carrying out a two-sided test. Numerically, we find that for small p, most rejections stem from positive values of $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$, and a two-sided test is conservative. However, for large values of p, the two-sided test maintains power even when the researcher imposes the wrong sign.

It is also of interest to consider when we lose power compared to the standard F-test. Suppose that $\gamma = 1/4$, so that the standard F-test indeed has power. A power loss occurs when the second term of (22) (partly) cancels against the first. For this to happen, we need to have $\boldsymbol{h}_p = b \cdot p^{-1/4}\boldsymbol{h}_{0,p}$ for some $b > 0$. However, in this case, $\hat{\omega} \to_p 1$. Therefore, as a practical recommendation, the regular F test can be used when the averaging weight is fully placed on the restricted model.

4 Numerical analysis

We consider the linear regression model (1) where all parameters are of interest, so $k = p$ and $\boldsymbol{G}_{k,p} = \boldsymbol{I}_p$. We vary the sample size $n = \{100, 400, 1600\}$ and the number of parameters of interest as $p = \{6, 12, 24, n/4, n/2\}$.

The parameters of interest $\boldsymbol{\beta}$ are set to reflect the setting that there are a few large parameters, while the remaining are close to zero, i.e. for $j = 1, \dots, p$,

$$\beta_j = \frac{c}{\sqrt{n}} \frac{j^{-1}}{\left(\sum_{i=1}^{p} i^{-2}\right)^{1/2}}. \tag{25}$$

The parameter c governs the magnitude of the coefficients and is varied as $c = \{-6, \dots, 6\}$.

We set $\boldsymbol{x}_{i,p} = \boldsymbol{L}_p \boldsymbol{z}_{i,p}$ and generate $\boldsymbol{z}_{i,p}$ and ε_i independently (1) from a standard normal distribution, or (2) as standardized squared $t(10)$ random variables to reflect a setting where regressors and errors only have lower order moments. The matrix $\boldsymbol{L}_p$ is obtained from a Cholesky decomposition of $\boldsymbol{\Sigma}_p = \mathrm{E}[\boldsymbol{x}_{i,p}\boldsymbol{x}_{i,p}']$ with elements $[\boldsymbol{\Sigma}_p]_{ij} = \rho^{|i-j|}$ where we set $\rho = 0.4$. All results are averaged over 10,000 draws of the set $\{\boldsymbol{x}_i, \varepsilon_i\}$. Estimation always includes an intercept.

Confidence regions are constructed as described in Section 2.2. The unrestricted estimator is obtained as in (7). Following the discussion below Theorem 2, the unrestricted estimator is averaged with the restricted estimator corresponding to (24), with $\bar{\boldsymbol{\beta}}$ the restricted estimator from (7) imposing that $\beta_i = 0$ for $i > 1$. The case where all but the first three coefficients are set to zero is presented in Appendix B.1.

We consider a researcher that beliefs the true effect of the first coefficient is positive. We therefore set $\bar{d} = 1$ in (24) corresponding to the optimal choice if there is only a single nonzero coefficient, which in the setting here is only approximately correct. Notice that the researcher is correct when $c > 0$ in (25). We perform both a one-sided test that rejects when $D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p) > \Phi^{-1}(1-\alpha)\hat{\tau}$ as well as a two-sided test that rejects when $|D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p)| > \Phi^{-1}(1-\alpha/2)\hat{\tau}$ with $\hat{\tau}$ as in (13) and $\alpha = 0.05$.

In Table 1, we show the coverage rate for the proposed confidence regions. In the upper panel, we consider the one-sided test. If the regressors and the errors are drawn from a normal distribution, the coverage rate is

Table 1: Coverage rate.

		n	$p=6$	$p=12$	$p=24$	$p=\frac{n}{4}$	$p=\frac{n}{2}$
One-sided	N	100	0.959	0.955	0.948	0.946	0.931
		400	0.959	0.954	0.952	0.949	0.947
		1600	0.961	0.958	0.950	0.957	0.956
	$t(10)^2$	100	0.913	0.898	0.891	0.896	0.876
		400	0.928	0.919	0.911	0.918	0.919
		1600	0.939	0.934	0.929	0.936	0.934
Two-sided	N	100	0.970	0.970	0.967	0.965	0.954
		400	0.973	0.968	0.969	0.965	0.961
		1600	0.975	0.974	0.971	0.967	0.964
	$t(10)^2$	100	0.928	0.916	0.911	0.918	0.900
		400	0.942	0.936	0.931	0.940	0.936
		1600	0.952	0.949	0.947	0.951	0.946

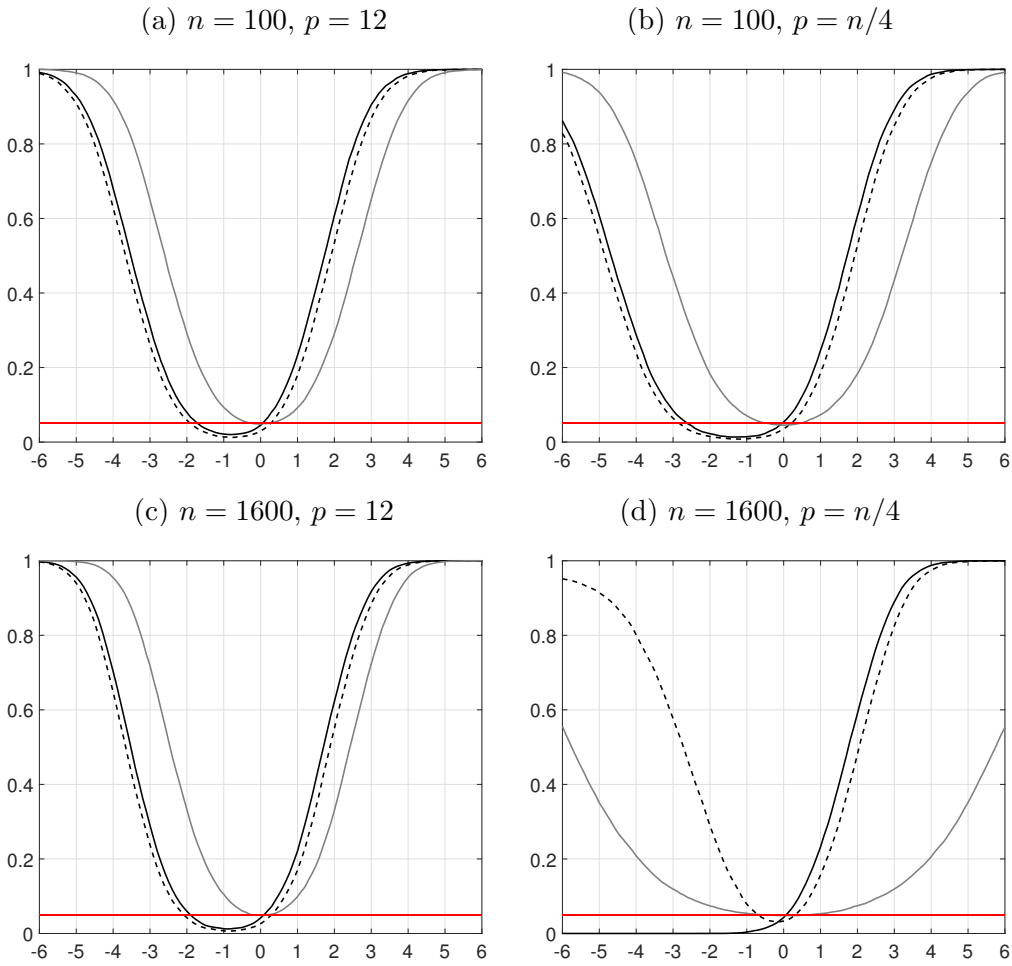
Note: coverage rate at $\boldsymbol{\beta} = \mathbf{0}$, sample size n, number of parameters of interest p. The unrestricted estimator from (7) is averaged with (24) where $\bar{d} = 1$. N indicates regressors and errors drawn from a normal distribution, $t(10)^2$ from a standardized squared $t(10)$ distribution. Nominal coverage equals 0.95.

close to the nominal level. When the regressors and errors are standardized $t(10)^2$ random variables, coverage drops for small values of n and this drop is worse for large values of p. As n increases, the coverage rate approaches the nominal rate. For large n, the coverage rate does not depend on the number of parameters of interest p, showing the robustness of the confidence regions to the inclusion of many parameters.

In the lower panel of Table 1 we consider the two-sided test that is motivated by the discussion following Theorem 2. For normally distributed regressors and errors, we see that the test is conservative for small values of p. Here, $D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p)$ behaves like a regular F-test, and nearly all rejections occur for large positive values. However, as p increases, we see coverage approaching the nominal rate. For standardized $t(10)^2$ random variables, there is again some undercoverage for $n = 100$, but the test has close to nominal coverage for larger values of n.

Figure 2 shows the power of the one-sided test (solid black line), the two-sided test (dashed black line) and the standard F-test (solid gray line)

Figure 2: Linear regression model: power.



Note: the figure shows power of a test based on (10) against $H_0\colon \boldsymbol{\beta} = \mathbf{0}$ at sample size n with number of parameters of interest p. The unrestricted estimator (7) is averaged with the restricted estimator in (24). Black solid lines correspond to a one-sided test based on Theorem 1, black dashed lines to the two-sided variant. Gray solid lines correspond to the usual F-test. The nominal level of the test is $\alpha = 0.05$.

against the null that $\beta_p = 0$. In the upper left panel, the sample size ($n = 100$) and the number of parameters of interest ($p = 12$) are small. We observe modest power gains when the researcher is right on the sign of the first coefficient (positive values on the x-axis), while a loss is observed when the researcher is wrong (negative values on the x-axis). These differences become stronger when moving to the upper right panel, where we increase the number of parameters of interest to $p = n/4$.

In the lower left panel, we increase the sample size ($n = 1600$), but keep

the number of parameters of interest fixed to $p = 12$. We observe almost no difference with the power curve in the upper left panel. However, if we move to the high-dimensional case where $p = n/4$ in the right lower panel, we see a strong loss of power of the standard F-test. We also clearly observe the effects described in the discussion following Theorem 2. The one-sided test now really is one-sided, while the two-sided test maintains power even when the wrong sign is imposed (dashed line). As the theory indicates, the two-sided test is now close to dominating the standard F-test.

Additional results The case where $s = 3$, that is when the restricted estimator allows the first three coefficients to be different from zero is reported in Appendix B.1. There are no substantial differences, although for $s = 3$, the power for negative values of the true coefficients is higher. Appendix B.2 compares the developed confidence regions with the procedures by Samworth (2005) and Casella and Hwang (1983), which we find to be conservative when the number of parameters under the test increases.

5 Conclusion

To investigate whether averaging estimators can be used to improve inference, we construct confidence regions centered at James-Stein averaging estimators. These regions are valid when the number of restrictions on the parameters of interest increases, possibly proportionally, with the sample size. When used for hypothesis testing, the imposed model restrictions can lead to a power enhancement over standard F-tests.

Important future extensions include the high-dimensional linear regression model under heteroskedasticity, as considered in Cattaneo et al. (2018) and recently in Anatolyev and Sølvsten (2020), or a joint limit theory for more general models estimated e.g. by maximum likelihood. This requires significant extensions to the methodology proposed here.

References

Anatolyev, S. (2012). Inference in regression models with many regressors. *Journal of Econometrics*, 170(2):368–382.

Anatolyev, S. and Sølvsten, M. (2020). Testing many restrictions under heteroskedasticity.

Anatolyev, S. and Yaskov, P. (2017). Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory*, 33(3):717–738.

Anderson, T., Kunitomo, N., and Matsushita, Y. (2010). On the asymptotic optimality of the liml estimator with possibly many instruments. *Journal of Econometrics*, 157(2):191–204.

Beran, R. (1995). Stein confidence sets and the bootstrap. *Statistica Sinica*, 5:109–127.

Casella, G. and Hwang, J. G. (2012). Shrinkage confidence procedures. *Statistical Science*, 27(1):51–60.

Casella, G. and Hwang, J. T. (1982). Limit expressions for the risk of James-Stein estimators. *Canadian Journal of Statistics*, 10(4):305–309.

Casella, G. and Hwang, J. T. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *Journal of the American Statistical Association*, 78(383):688–698.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361.

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., and Woutersen, T. (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory*, 28(1):42–86.

Cheng, X., Liao, Z., and Shi, R. (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics*, 10(3):931–979.

Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging. *Cambridge Books*.

DiTraglia, F. J. (2016). Using invalid instruments on purpose: focused moment selection and averaging for GMM. *Journal of Econometrics*, 195(2):187–208.

Efron, B. (2006). Minimum volume confidence regions for a multivariate normal mean vector. *Journal of the Royal Statistical Society: Series B*, 68(4):655–670.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 5(3):495–530.

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

Hansen, B. E. (2017). A Stein-like 2SLS estimator. *Econometric Reviews*, 36(6-9):840–852.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Hwang, J. G. and Ullah, A. (1994). Confidence sets centered at James-Stein estimators: A surprise concerning the unknown-variance case. *Journal of Econometrics*, 60(1-2):145–156.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Kunitomo, N. (2012). An optimal modification of the liml estimation for many instruments and persistent heteroscedasticity. *Annals of the Institute of Statistical Mathematics*, 64(5):881–910.

Leeb, H. and Kabaila, P. (2017). Admissibility of the usual confidence set for the mean of a univariate or bivariate normal population: the

unknown variance case. *Journal of the Royal Statistical Society: Series B*, 79(3):801–813.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186(1):142–159.

Liu, C.-A. and Kuo, B.-S. (2016). Model averaging in predictive regressions. *The Econometrics Journal*, 19(2):203–231.

Samworth, R. (2005). Small confidence sets for the mean of a spherically symmetric distribution. *Journal of the Royal Statistical Society: Series B*, 67(3):343–361.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151.

Ullah, A. (2004). *Finite sample econometrics*. Oxford University Press.

Zhang, X. and Liu, C.-A. (2019). Inference after model averaging in linear regression models. *Econometric Theory*, 35(4):816–841.

Appendix A Mathematical details

A.1 Finite sample results under exact normality and known error variance

In this section, we assume that $\varepsilon_i|\boldsymbol{X}_{n,k} \overset{i.i.d}{\sim} N(0,\sigma^2)$ and σ^2 is known. We calculate an unbiased risk estimator and the variance of $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$ defined in (10). The former is a standard calculation, while the latter appears new. The risk estimator is used to recenter the loss in (10), while the variance motivates a finite sample correction in (13).

We use the following notation in addition to that defined in Section 2. The estimators for the covariance matrix, conditional on $\boldsymbol{X}_{n,k}$, of the restricted estimators $\tilde{\boldsymbol{\theta}}_k$ from (5) and $\tilde{\boldsymbol{\beta}}_p$ from (7) are

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_{\theta,k} &= \hat{\boldsymbol{\Sigma}}_{\theta,k} - \hat{\boldsymbol{\Sigma}}_{\theta,k}\boldsymbol{R}_{k,r}(\boldsymbol{R}'_{k,r}\hat{\boldsymbol{\Sigma}}_{\theta,k}\boldsymbol{R}_{k,r})^{-1}\boldsymbol{R}'_{k,r}\hat{\boldsymbol{\Sigma}}_{\theta,k}, \\ \tilde{\boldsymbol{\Sigma}}_{\beta,p} &= \hat{\sigma}^2\boldsymbol{G}'_{k,p}\tilde{\boldsymbol{\Sigma}}_{\theta,k}\boldsymbol{G}_{k,p}.\end{aligned} \tag{A.1}$$

The difference between the unrestricted estimator and the restricted estimator is denoted by

$$\hat{\boldsymbol{\delta}}_p = \hat{\boldsymbol{\beta}}_p - \tilde{\boldsymbol{\beta}}_p, \quad \mathrm{E}[\hat{\boldsymbol{\delta}}_p] = \boldsymbol{\delta}_p. \tag{A.2}$$

The estimator for the covariance matrix of $\hat{\boldsymbol{\delta}}_p$ conditional on $\boldsymbol{X}_{n,k}$ is

$$\hat{\boldsymbol{\Sigma}}_{\delta,p} = \hat{\boldsymbol{\Sigma}}_{\beta,p}\boldsymbol{B}_{p,r}(\boldsymbol{B}'_{p,r}\hat{\boldsymbol{\Sigma}}_{\beta,p}\boldsymbol{B}_{p,r})^{-1}\boldsymbol{B}'_{p,r}\hat{\boldsymbol{\Sigma}}_{\beta,p}, \tag{A.3}$$

with $\boldsymbol{B}_{p,r}$ as in (3) and $\hat{\boldsymbol{\Sigma}}_{\beta,p}$ as in (8).

A.1.1 Unbiased risk estimator

We will show that the risk estimator $\hat{\rho}(\hat{\boldsymbol{\beta}}^a_p, \boldsymbol{\beta}_p)$ defined in (11) satisfies

$$\mathrm{E}\left[\hat{\rho}\left(\hat{\boldsymbol{\beta}}^a_p, \boldsymbol{\beta}_p\right)\right] = \rho\left(\hat{\boldsymbol{\beta}}^a_p, \boldsymbol{\beta}_p\right). \tag{A.4}$$

In the derivations below, we define the conditional risk of any estimator $\bar{\boldsymbol{\beta}}_p$ as $\rho(\bar{\boldsymbol{\beta}}_p, \boldsymbol{\beta}_p|\boldsymbol{X}_{n,k}) = \mathrm{E}[\ell(\bar{\boldsymbol{\beta}}_p, \boldsymbol{\beta}_p)|\boldsymbol{X}_{n,k}]$.

Standard calculations show that

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p \\ \tilde{\boldsymbol{\beta}}_p - \mathrm{E}[\tilde{\boldsymbol{\beta}}_p] \\ \hat{\boldsymbol{\delta}}_p - \boldsymbol{\delta}_p \end{pmatrix} \sim N(\boldsymbol{0}, \boldsymbol{V}_{3p}), \quad \boldsymbol{V}_{3p} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{\beta,p} & \hat{\boldsymbol{\Sigma}}_{\beta,p} - \boldsymbol{A}_p & \boldsymbol{A}_p \\ \hat{\boldsymbol{\Sigma}}_{\beta,p} - \boldsymbol{A}_n & \hat{\boldsymbol{\Sigma}}_{\beta,p} - \boldsymbol{A}_p & \boldsymbol{O} \\ \boldsymbol{A}_p & \boldsymbol{O} & \boldsymbol{A}_p \end{pmatrix} \tag{A.5}$$

where $\boldsymbol{A}_p = \hat{\boldsymbol{\Sigma}}_{\beta,p}\boldsymbol{B}_{p,r}(\boldsymbol{B}'_{p,r}\hat{\boldsymbol{\Sigma}}_{\beta,p}\boldsymbol{B}_{p,r})^{-1}\boldsymbol{B}'_{p,r}\hat{\boldsymbol{\Sigma}}_{\beta,p}$, with $\boldsymbol{B}_{p,r}$ as in (3).

From (A.5), we conclude that $\tilde{\boldsymbol{\beta}}_p$ and $\hat{\boldsymbol{\delta}}_p$ are independent. As the weight $\hat{\omega}$ only depends on $\hat{\boldsymbol{\delta}}_p$, the asymptotic risk of the averaging estimator con-

sists of two terms,

$$\rho(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) = \mathrm{E}\left[n(\tilde{\boldsymbol{\beta}}_p - (\boldsymbol{\beta}_p - \boldsymbol{\delta}_p))'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\tilde{\boldsymbol{\beta}}_p - (\boldsymbol{\beta}_p - \boldsymbol{\delta}_p))\right] + \rho(\hat{\boldsymbol{\delta}}_p^a, \boldsymbol{\delta}_p). \tag{A.6}$$

The first term equals $\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\beta,p}\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}) = p - r$, and the second term is

$$\rho(\hat{\boldsymbol{\delta}}_p^a, \boldsymbol{\delta}_p) = \mathrm{E}\left[(\hat{\boldsymbol{\delta}}_p^a - \boldsymbol{\delta}_p)'\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\hat{\boldsymbol{\delta}}_p^a - \boldsymbol{\delta}_p)\right], \quad \hat{\boldsymbol{\delta}}_p^a = (1-\hat{\omega})\hat{\boldsymbol{\delta}}_p. \tag{A.7}$$

The following quantities are helpful in the derivations below.

$$\begin{aligned} \hat{\omega} &= \frac{r-2}{n\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n}, \\ \boldsymbol{g}(\boldsymbol{y}_n) &= -\frac{r-2}{n\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n}\boldsymbol{y}_n, \\ \boldsymbol{h}(\boldsymbol{y}_n) &= -\frac{r-2}{n\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n}\boldsymbol{P}_{R,n}\boldsymbol{y}_n, \end{aligned} \tag{A.8}$$

with $\boldsymbol{P}_{R,n}$ as in Assumption 4.

In terms of the quantities in (A.8), we have

$$\begin{aligned} \rho(\hat{\boldsymbol{\delta}}_p^a, \boldsymbol{\delta}_p|\boldsymbol{X}_{n,k}) &= \mathrm{E}\left[\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n + 2\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{\varepsilon}_n + \boldsymbol{g}(\boldsymbol{y}_n)'\boldsymbol{h}(\boldsymbol{y}_n)|\,\boldsymbol{X}_{n,k}\right] \\ &= \mathrm{tr}(\boldsymbol{P}_{R,n}) + 2\mathrm{E}[\boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n)|\,\boldsymbol{X}_{n,k}] + \mathrm{E}[\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{g}(\boldsymbol{y}_n)|\,\boldsymbol{X}_{n,k}], \end{aligned} \tag{A.9}$$

where the second term in the last line is obtained by applying Stein's lemma to the second term on the first line.

The second term of (A.9) can be further written out using that

$$\frac{\partial h_i(\boldsymbol{y}_n)}{\partial y_k} = -(r-2)\left[\frac{[\boldsymbol{P}_{R,n}]_{ik}}{\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n} - 2\frac{\sum_{l,n}[\boldsymbol{P}_{R,n}]_{il}y_l[\boldsymbol{P}_{R,n}]_{km}y_n}{(\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n)^2}\right], \tag{A.10}$$

such that

$$\begin{aligned} \boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n) &= -(r-2)\left[\frac{\mathrm{tr}(\boldsymbol{P}_{R,n})}{\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n} - 2\frac{\boldsymbol{y}_n'\boldsymbol{P}_{R,n}^2\boldsymbol{y}_n}{(\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n)^2}\right] \\ &= -(r-2)^2\frac{1}{\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n}. \end{aligned} \tag{A.11}$$

The conditional risk of the averaging estimator is then found to be

$$\rho(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p | \boldsymbol{X}_{n,k}) = p - (r-2)\mathrm{E}[\hat{\omega} | \, \boldsymbol{X}_{n,k}]. \tag{A.12}$$

It follows that the risk is $\rho(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) = p - (r-2)\mathrm{E}[\hat{\omega}]$ and the unbiased risk estimator is $\hat{\rho}(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) = p - (r-2)\hat{\omega}$.

A.1.2 Variance of $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$

To calculate the variance of $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$ we split the statistic $D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)$ into two zero mean components,

$$D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) = A_{rr} + A_{\delta\delta}, \tag{A.13}$$

where

$$\begin{aligned}
A_{rr} &= p^{-\frac{1}{2}} \left[n(\tilde{\boldsymbol{\beta}}_p - (\boldsymbol{\beta}_p - \boldsymbol{\delta}_p))' \hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1} (\tilde{\boldsymbol{\beta}}_p - (\boldsymbol{\beta}_p - \boldsymbol{\delta}_p)) - (p-r) \right], \\
A_{\delta\delta} &= p^{-\frac{1}{2}} \left[n(\hat{\boldsymbol{\delta}}_p - \boldsymbol{\delta}_p)' \hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1} (\hat{\boldsymbol{\delta}}_p - \boldsymbol{\delta}_p) - r - 2\hat{\omega} \left(n \hat{\boldsymbol{\delta}}_p' \hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1} (\hat{\boldsymbol{\delta}}_p - \boldsymbol{\delta}_p) - (r-2) \right) \right].
\end{aligned}$$

Since $\tilde{\boldsymbol{\beta}}_p$ and $\hat{\boldsymbol{\delta}}_p$ are independent, $\mathrm{cov}(A_{rr}, A_{\delta\delta}) = 0$. It is therefore sufficient to determine the variance of the individual terms. The variance of A_{rr} follows from standard results on quadratic forms in normal vectors.

$$\mathrm{E}[A_{rr}^2] = \frac{2}{p} \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1} \tilde{\boldsymbol{\Sigma}}_{\beta,p} \hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1} \tilde{\boldsymbol{\Sigma}}_{\beta,p}) = 2\frac{p-r}{p}. \tag{A.14}$$

For the variance of $A_{\delta\delta}$, we use definitions (A.8) to write

$$\begin{aligned}
\mathrm{E}[A_{\delta\delta}^2] &= \frac{1}{p} \mathrm{E} \left\{ \left[\boldsymbol{\varepsilon}_n' \boldsymbol{P}_{R,n} \boldsymbol{\varepsilon}_n - r + 2 \left(\boldsymbol{h}(\boldsymbol{y}_n)' \boldsymbol{\varepsilon}_n - \boldsymbol{\nabla}' \boldsymbol{h}(\boldsymbol{y}_n) \right) \right]^2 \right\} \\
&= \frac{1}{p} \mathrm{E} \left\{ \left[\boldsymbol{\varepsilon}_n' \boldsymbol{P}_{R,n} \boldsymbol{\varepsilon}_n - r \right]^2 + 4 \left(\boldsymbol{h}(\boldsymbol{y}_n)' \boldsymbol{\varepsilon}_n - \boldsymbol{\nabla}' \boldsymbol{h}(\boldsymbol{y}_n) \right)^2 \right. \\
&\qquad \left. + 4 \left(\boldsymbol{h}(\boldsymbol{y}_n)' \boldsymbol{\varepsilon}_n - \boldsymbol{\nabla}' \boldsymbol{h}(\boldsymbol{y}_n) \right) \left[\boldsymbol{\varepsilon}' \boldsymbol{P}_{R,n} \boldsymbol{\varepsilon}_n - r \right] \right\} \\
&= 2\frac{r}{p} + \frac{4}{p} \mathrm{E} \left\{ \left(\boldsymbol{h}(\boldsymbol{y}_n)' \boldsymbol{\varepsilon}_n \right)^2 + \left(\boldsymbol{\nabla}' \boldsymbol{h}(\boldsymbol{y}_n) \right)^2 - 2\boldsymbol{\varepsilon}_n' \boldsymbol{h}(\boldsymbol{y}_n) \boldsymbol{\nabla}' \boldsymbol{h}(\boldsymbol{y}_n) \right. \\
&\qquad \left. + \boldsymbol{h}(\boldsymbol{y}_n)' \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' \boldsymbol{P}_{R,n} \boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n' \boldsymbol{P}_{R,n} \boldsymbol{\varepsilon}_n \boldsymbol{\nabla}' \boldsymbol{h}(\boldsymbol{y}_n) \right\}.
\end{aligned}$$

To proceed, we use the following result derived in Theorem 3 of Stein (1981)

by repeatedly applying Stein's lemma.

$$\begin{aligned}\mathrm{E}\left[(\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{\varepsilon}_n)^2\right] &= \mathrm{E}\Big[\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{h}(\boldsymbol{y}_n) + (\boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n))^2 \\ &\quad + \mathrm{tr}[(\boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{y}_n)')^2] + 2\sum_{i=1}^{p}\sum_{j=1}^{p} h_i(\boldsymbol{y}_n)\boldsymbol{\nabla}_j\boldsymbol{\nabla}_i h_j(\boldsymbol{y}_n)\Big], \end{aligned} \tag{A.15}$$

$$\mathrm{E}\left[\boldsymbol{\varepsilon}_n'\boldsymbol{h}(\boldsymbol{y}_n)\boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n)\right] = \mathrm{E}\left[(\boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n))^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} h_i(\boldsymbol{y}_n)\boldsymbol{\nabla}_j\boldsymbol{\nabla}_i h_j(\boldsymbol{y}_n)\right].$$

The final two terms of (A.15) require an extension to the results presented by Stein (1981). Applying Stein's lemma twice, we have

$$\begin{aligned}\mathrm{E}[\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{\varepsilon}_n] &= \mathrm{E}\left[(\boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n)\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n + 2\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n\right] \\ &= \mathrm{E}\left[\boldsymbol{\nabla}'\boldsymbol{h}(\boldsymbol{y}_n)\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n + 2\boldsymbol{\nabla}'\boldsymbol{P}_{R,n}\boldsymbol{h}(\boldsymbol{y}_n)\right]. \end{aligned} \tag{A.16}$$

In total, we now have

$$\mathrm{E}[A_{\delta\delta}^2] = \frac{2r}{p} + \frac{4}{p}\mathrm{E}\left[\boldsymbol{h}(\boldsymbol{y}_n)'\boldsymbol{h}(\boldsymbol{y}_n) + \mathrm{tr}\left[(\boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{y}_n)')^2\right] + 2\boldsymbol{\nabla}'\boldsymbol{P}_{R,n}\boldsymbol{h}(\boldsymbol{y}_n)\right]. \tag{A.17}$$

We can work out the final two terms explicitly,

$$\begin{aligned}\mathrm{tr}\left[(\boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{y}_n)')^2\right] &= (r-2)^2\left[\frac{r}{(\boldsymbol{y}_n\boldsymbol{P}_{R,n}\boldsymbol{y}_n)^2}\right], \\ \boldsymbol{\nabla}'\boldsymbol{P}_{R,n}\boldsymbol{h}(\boldsymbol{y}_n) &= -\frac{(r-2)^2}{\boldsymbol{y}_n'\boldsymbol{P}_{R,n}\boldsymbol{y}_n}. \end{aligned} \tag{A.18}$$

Substituting this into (A.17) gives

$$\mathrm{E}[A_{\delta\delta}^2] = 2\frac{r}{p} - \frac{4(r-2)^2}{p}\mathrm{E}\left[\frac{1}{n\hat{\boldsymbol{\delta}}_p\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\hat{\boldsymbol{\delta}}_p} - \frac{r}{(n\hat{\boldsymbol{\delta}}_p\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}\hat{\boldsymbol{\delta}}_p)^2}\right]. \tag{A.19}$$

Adding the variances of A_{rr} and $A_{\delta\delta}$, we obtain

$$\mathrm{V}[D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)] = 2 - \frac{4(r-2)^2}{pr}\mathrm{E}\left[\frac{\hat{F}-1}{\hat{F}^2}\right]. \tag{A.20}$$

From this expression it is straightforward to obtain an unbiased estima-

tor for the variance by removing the expectation. However, this estimator has the drawback that in finite samples potentially $\hat{F} - 1 < 0$, while asymptotically this quantity is always nonnegative. We therefore define $\hat{\lambda}^2 = \max(0, \hat{F} - 1)$ and estimate the variance by

$$\mathrm{V}[D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p)] = 2 - \frac{4(r-2)^2}{pr} \frac{\hat{\lambda}^2}{(\hat{\lambda}^2 + 1)^2}. \tag{A.21}$$

Compared to (13), this expression misses the terms that appear in the high-dimensional case where $(r, p, k)/n \not\to 0$. The reason is that these additional term arise from estimation uncertainty in σ^2, while this section supposes σ^2 to be known. The main contribution of (A.21) is the finite sample correction appearing in the term $(r-2)^2$, which we found to be effective for small r in numerical computations.

A.2 Proof of Theorem 1

A.2.1 Preliminaries

Throughout, M denotes a positive, finite constant that can differ between occurrences. We first prove that a key result from Chao et al. (2012), henceforth CSHNW, holds under a set of conditions that is adapted to our case. We will follow the notation of CSHNW as close as possible.

Lemma A.1 (Adaptation of CSHNW, Lemma A2) *Suppose that, conditional on $\boldsymbol{X}_{n,k}$, the following conditions hold a.s.*

(i). The matrix $\boldsymbol{A}_n \in \mathbb{R}^{n,n}$ is symmetric and satisfies:

- *(a) $\boldsymbol{e}_i' \boldsymbol{A}_n^j \boldsymbol{e}_i \leq M$ for $j = 1, 2$,*
- *(b) $\mathrm{tr}(\boldsymbol{A}_n^j) \leq Mp$ for $j = 2, 3, 4$.*

(ii). $\{W_i, \varepsilon_i\}$ is an independent sequence, with $D = \sum_{i=1}^n \mathrm{E}[W_i^2 | \boldsymbol{X}_{n,k}] \leq M$ a.s.n.

(iii). $\mathrm{E}[W_i | \boldsymbol{X}_{n,k}] = 0$, $\mathrm{E}[\varepsilon_i | \boldsymbol{X}_{n,k}] = 0$, $\mathrm{E}[\varepsilon_i^2 | \boldsymbol{X}_{n,k}] = \mathrm{E}[\varepsilon_i^2] = \sigma^2$, $\mathrm{E}[\varepsilon_i^4 | \boldsymbol{X}_{n,k}] \leq M$.

(iv). $\sum_{i=1}^n \mathrm{E}[|W_i|^4 | \boldsymbol{X}_{n,k}] \to_{a.s.} 0$.

(v). $p \to \infty$ *as* $n \to \infty$.

Then for

$$\bar{\Sigma} = \frac{2\sigma^4}{p} \sum_{i \neq j} A_{ij}^2, \tag{A.22}$$

and $\Xi = D + \bar{\Sigma} > M$ *a.s.n., it follows that*

$$Y = \Xi^{-1/2} \left[\sum_{i=1}^{n} W_i + \frac{1}{\sqrt{p}} \sum_{i \neq j} \varepsilon_i \varepsilon_j A_{ij} \right] \Rightarrow N(0,1), \quad a.s. \tag{A.23}$$

i.e. $\mathrm{P}(Y \leq y | \boldsymbol{X}_{n,k}) \to \Phi(y)$ *a.s. for all* y.

Proof: The proof of Lemma A2 of CSHNW only relies on $\boldsymbol{A}_n$ (in their notation $\boldsymbol{P}$) through their Lemma B1 to B4 and the requirement that $\Xi > M$ *a.s.n.* We first show that Lemma B1 to B4 continue to hold under Assumption (i) instead of assuming $\boldsymbol{A}_n$ to be idempotent. Throughout, we denote the i,j-th element of $\boldsymbol{A}_n$ as A_{ij}.

Lemma A.1.1 (Adaptation of CSHNW, Lemma B1) *Under Assumption (i) of Lemma A.1 and for any subset* I_2 *of the set* $\{(i,j)_{i,j=1}^n\}$ *and any subset* I_3 *of* $\{(i,j,k)_{i,j,k=1}^n\}$,

(i). $\sum_{I_2} A_{ij}^4 \leq Mp$.

(ii). $\sum_{I_3} A_{ij}^2 A_{jk}^2 \leq Mp$.

(iii). $\sum_{I_3} |A_{ij}^2 A_{ik} A_{jk}| \leq Mp$.

Proof: Part (i)

$$\begin{aligned} \sum_{I_2} A_{ij}^4 &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^4 + \sum_{i=1}^{n} \sum_{j \neq j'} A_{ij}^2 A_{ij'}^2 \\ &= \sum_{i=1}^{n} (\boldsymbol{e}_i' \boldsymbol{A}_n^2 \boldsymbol{e}_i)^2 \\ &\leq \sum_{i=1}^{n} \boldsymbol{e}_i' \boldsymbol{A}_n^4 \boldsymbol{e}_i \\ &= \mathrm{tr}(\boldsymbol{A}_n^4) \leq Mp, \end{aligned} \tag{A.24}$$

where the second inequality uses the fact that in general $\boldsymbol{v}'\boldsymbol{B}\boldsymbol{v} \leq \lambda_{\max}(\boldsymbol{B})\boldsymbol{v}'\boldsymbol{v}$, where $\lambda_{\max}(\boldsymbol{B})$ denotes the maximum eigenvalue of $\boldsymbol{B}$. We apply this with

$\boldsymbol{B} = \boldsymbol{e}_i \boldsymbol{e}_i'$, so that $\lambda_{\max}(\boldsymbol{B}) = 1$. The last equality follows by Assumption (i) part (b).

Continuing with part (ii), we have

$$\begin{aligned} \sum_{I_3} A_{ij}^2 A_{jk}^2 &\leq \sum_{j=1}^{n} \left(\sum_{i=1}^{n} A_{ij}^2 \right) \left(\sum_{k=1}^{n} A_{kj}^2 \right) \\ &= \sum_{j=1}^{n} (\boldsymbol{e}_j' \boldsymbol{A}_n^2 \boldsymbol{e}_j)^2 \\ &\leq Mp, \end{aligned} \tag{A.25}$$

with the last inequality derived as in part (i).

Finally, for part (iii)

$$\begin{aligned} \sum_{I_3} |A_{ij}^2 A_{ik} A_{jk}| &\leq \sum_{i,j} A_{ij}^2 \sum_{k} |A_{ik} A_{)} jk| \\ &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 \sqrt{\sum_{k=1}^{n} A_{ik}^2 \sum_{k=1}^{n} A_{jk}^2} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 \sqrt{\boldsymbol{e}_j' \boldsymbol{A}_n^2 \boldsymbol{e}_j \boldsymbol{e}_i' \boldsymbol{A}_n^2 \boldsymbol{e}_i} \\ &\leq M \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 \\ &= M \mathrm{tr}(\boldsymbol{A}_n^2) \leq Mp, \end{aligned} \tag{A.26}$$

where on the fourth line we use Assumption (i) part (a) and on the final line Assumption (i) part (b). ■

In the following lemma, we momentarily suspend the subscript to indicate the dimensions off the involved matrices, so $\boldsymbol{A}_n = \boldsymbol{A}$ and $\boldsymbol{D}_n = \boldsymbol{D}$.

Lemma A.1.2 (Adaptation of CSHNW, Lemma B2) *Suppose Assumption (i) of Lemma A.1 holds for $\boldsymbol{A}$. Then a.s.n.*

(i). $\mathrm{tr}[(\boldsymbol{A} - \boldsymbol{D})^4] \leq Mp$ *where* $[\boldsymbol{D}]_{ii} = [\boldsymbol{A}]_{ii}$.

(ii). $|\sum_{i<j<k<l} A_{ik} A_{jk} A_{il} A_{jl}| \leq Mp$.

(iii). $|S_n| \leq Mp$,

where $S_n = \sum_{i<j<k<l}(P_{ik}P_{jk}P_{il}P_{jl} + P_{ij}P_{jk}P_{il}P_{kl} + P_{ij}P_{ik}P_{jl}P_{kl})$.

Proof: Parts (ii) and (iii) only rely on part (i) and Lemma A.1.1. We therefore only prove part (i).

$$\begin{aligned}(\boldsymbol{A}-\boldsymbol{D})^4 = \boldsymbol{A}^4 + \boldsymbol{D}^4 + \boldsymbol{A}\boldsymbol{D}^2\boldsymbol{A} + \boldsymbol{D}\boldsymbol{A}^2\boldsymbol{D} \\ - \boldsymbol{A}^2\boldsymbol{D}\boldsymbol{A} - \boldsymbol{A}^3\boldsymbol{D} + \boldsymbol{A}^2\boldsymbol{D}^2 \\ - \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^2 - \boldsymbol{D}\boldsymbol{A}^3 + \boldsymbol{D}^2\boldsymbol{A}^2 \\ - \boldsymbol{D}^2\boldsymbol{A}\boldsymbol{D} - \boldsymbol{A}\boldsymbol{D}^3 + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}\boldsymbol{D} \\ - \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^2 - \boldsymbol{D}^3\boldsymbol{A} + \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}.\end{aligned} \tag{A.27}$$

Taking the trace, we find

$$\mathrm{tr}(\boldsymbol{A}-\boldsymbol{D})^4 = \mathrm{tr}(\boldsymbol{A}^4)+\mathrm{tr}(\boldsymbol{D}^4)+4\mathrm{tr}(\boldsymbol{A}^2\boldsymbol{D}^2)-4\mathrm{tr}(\boldsymbol{D}\boldsymbol{A}^3)-4\mathrm{tr}(\boldsymbol{A}\boldsymbol{D}^3)+2\mathrm{tr}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}\boldsymbol{D}). \tag{A.28}$$

Now $\mathrm{tr}(\boldsymbol{A}^4) \leq Mp$ by Assumption (i) part (b). For the second term, using repeatedly that $\boldsymbol{v}'\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{v} \leq \boldsymbol{v}'\boldsymbol{v}$, we have

$$\mathrm{tr}(\boldsymbol{D}^4) = \sum_{i=1}^{n}(\boldsymbol{e}_i'\boldsymbol{A}\boldsymbol{e}_i)^4 \leq \sum_{i=1}\boldsymbol{e}_i'\boldsymbol{A}^4\boldsymbol{e}_i = \mathrm{tr}(\boldsymbol{A}^4) \leq Mp. \tag{A.29}$$

For the third term, we first use Assumption (i) part (a) and then Assumption (i) part (b) to get

$$\mathrm{tr}(\boldsymbol{A}^2\boldsymbol{D}^2) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{D}^2\boldsymbol{A}) = \sum_{i=1}^{n}\boldsymbol{e}_i'\boldsymbol{A}\boldsymbol{D}^2\boldsymbol{A}\boldsymbol{e}_i \leq M\sum_{i=1}^{n}\boldsymbol{e}_i'\boldsymbol{A}^2\boldsymbol{e}_i \leq Mp. \tag{A.30}$$

The same argument can be applied to the fourth and the fifth term. For the final term, notice that it equals

$$\mathrm{tr}(\boldsymbol{D}^{1/2}\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^{1/2}) \leq M\mathrm{tr}(\boldsymbol{D}^{1/2}\boldsymbol{A}^2\boldsymbol{D}^{1/2}) = M\mathrm{tr}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}) \leq M\mathrm{tr}(\boldsymbol{A}^2) \leq Mp. \tag{A.31}$$

This completes the proof of part (i). ■

Continuing with Lemma B3 of CSHNW, we note that it only relies on Lemma B1 and B2, so it holds in our case as well. Lemma B4 of CSHNW uses the fact that $A_{ii} < 1$ in equation B.8. This can be replaced by our

Assumption (i) since in the second to last line of that equation, we have

$$\frac{M}{p^2}\sum_{i=1}^{n}(\boldsymbol{e}_i'\boldsymbol{A}_n^2\boldsymbol{e}_i)^2 \leq \frac{M}{p^2}\sum_{i=1}^{n}\boldsymbol{e}_i'\boldsymbol{A}_n^2 e_i \leq \frac{M}{p}. \tag{A.32}$$

We conclude that Lemma B1-B4 also hold under Assumption (i) and Lemma A2 of CHSNW holds under the stated assumptions of Lemma A.1. ■

A.2.2 Consistency of error variance and weights

For the proof of Theorem 1 we need the following result on the consistency of the estimator of the error variance $\hat{\sigma}^2$ and the weights $\hat{\omega}$.

Lemma A.2 *Define $\hat{\sigma}^2$ as in (6) and $\hat{\omega}$ as in (8). Under Assumption 1–4,*
(a) $\hat{\sigma}^2 - \sigma^2 = O_p(n^{-1/2})$,
(b) $\hat{\omega} - \omega = o_p(1)$ with $\omega = (\lambda^2+1)^{-1}$ and λ^2 defined in Assumption 3.

Proof: Part (a). We have

$$\begin{aligned}
\mathrm{E}[\hat{\sigma}^2] &= \sigma^2, \\
\mathrm{Var}(\hat{\sigma}^2|\boldsymbol{X}_{n,k}) &= \frac{1}{(n-k)^2}\left[\mathrm{E}[(\boldsymbol{\varepsilon}_n'\boldsymbol{M}_{X_{n,k}}\boldsymbol{\varepsilon}_n)^2|\boldsymbol{X}_{n,k}] - \mathrm{E}[\boldsymbol{\varepsilon}_n'\boldsymbol{M}_{X_{n,k}}\boldsymbol{\varepsilon}_n|\boldsymbol{X}_{n,k}]^2\right] \\
&= \frac{\sigma^4}{(n-k)^2}\left[(\mathrm{E}[\varepsilon_i^4|\boldsymbol{X}_{n,k}]/\sigma^4 - 3)\sum_{i=1}^{n}[\boldsymbol{M}_{X_{n,k}}]_{ii}^2 + 2(n-k)\right] \\
&\leq M\frac{\sigma^4}{n-k}(1+o_{a.s.}(1)),
\end{aligned}$$

where the last line uses Assumption 2 to bound the fourth moment and Assumption 4 to write

$$\begin{aligned}
\frac{n}{(n-k)^2}\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{M}_{X_{n,k}}]_{ii}^2 &= \frac{n}{(n-k)^2}\left\{(1-\kappa)^2 + \frac{1}{n}\sum_{i=1}^{n}([\boldsymbol{M}_{X_{n,k}}]_{ii} - (1-\kappa))^2\right\} \\
&= \frac{1}{n-k}(1+o_{a.s.}(1)).
\end{aligned} \tag{A.33}$$

Define now $Y_n = \mathrm{P}(|\hat{\sigma}^2 - \sigma^2| > \epsilon|\boldsymbol{X}_{n,k})$, then $Y_n \to_{a.s.} 0$ and $Y_n \leq 1$. By the dominated convergence theorem, $\mathrm{P}(|\hat{\sigma}^2 - \sigma^2| > \epsilon) = \mathrm{E}[Y_n] \to 0$.

Part (b). Recall that $\hat{\omega} = \frac{r}{r-2}\frac{1}{\hat{F}}$. Rescaling the F-statistic from (8) with $\hat{\sigma}^2$, we have

$$\hat{\sigma}^2\hat{F} = r^{-1}\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n + \sigma^2\bar{\lambda}^2 + 2r^{-1}n^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n. \tag{A.34}$$

The second term converges almost surely to $\sigma^2\lambda^2$ by Assumption 3. For the first term, note that $\text{tr}(\boldsymbol{P}_{R,n}) = r$. Then,

$$\begin{aligned} r^{-1}\text{E}[\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n] &= r^{-1}\text{E}[\text{E}[\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n|\boldsymbol{X}_{n,k}]] \\ &= r^{-1}\sigma^2\text{E}[\text{tr}(\boldsymbol{P}_{R,n})] \\ &= \sigma^2. \end{aligned} \tag{A.35}$$

For the variance, by Ullah (2004) (Appendix A5), and using that conditional on $\boldsymbol{X}_{n,k}$, ε_i has bounded fourth moment,

$$\begin{aligned} \text{var}(r^{-1}\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n) &= r^{-2}\text{E}\left[\text{E}[(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n)^2|\boldsymbol{X}_{n,k}]\right] - \sigma^4 \\ &\leq r^{-2}M\text{E}[\text{tr}(\boldsymbol{P}_{R,n}(\boldsymbol{I}_n \odot \boldsymbol{P}_{R,n}))] + 2\sigma^4 r^{-1} \\ &= r^{-2}M\text{E}[\text{tr}((\boldsymbol{I}_n \odot \boldsymbol{P}_{R,n})^{1/2}\boldsymbol{P}_{R,n}(\boldsymbol{I}_n \odot \boldsymbol{P}_{R,n})^{1/2})] + 2\sigma^4 r^{-1} \\ &\leq r^{-2}M\text{E}[\text{tr}(\boldsymbol{P}_{R,n})] + 2\sigma^4 r^{-1} \\ &= O(r^{-1}), \end{aligned} \tag{A.36}$$

where the inequality on the fourth line uses that $\boldsymbol{P}_{R,n} \preceq \boldsymbol{I}_n$, since $\boldsymbol{P}_{R,n}$ is idempotent. By Chebyshev's inequality, as $(r, n \to \infty)$, $r^{-1}\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{R,n}\boldsymbol{\varepsilon}_n \to_p \sigma^2$.

The final term of (A.34) has expected value equal to zero, and

$$\begin{aligned} &\text{var}(r^{-1}n^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n|\boldsymbol{X}_{n,k}) \\ &= \sigma^2 \cdot r^{-1} \cdot \left[r^{-1}\boldsymbol{h}'\boldsymbol{R}_{k,r}(\boldsymbol{R}_{k,r}'(n^{-1}\boldsymbol{X}_{n,k}'\boldsymbol{X}_{n,k})^{-1}\boldsymbol{R}_{k,r})^{-1}\boldsymbol{R}_{k,r}'\boldsymbol{h}\right] \\ &= \sigma^2 r^{-1}(\lambda^2 + o_{a.s}(1)). \end{aligned} \tag{A.37}$$

Using again the dominated convergence theorem, this shows that the final term of (A.34) converges to zero in probability.

We have now established that, as $(r, n \to \infty)$, $\hat{\sigma}^2\hat{F} \to_p \sigma^2(\lambda^2 + 1)$ and

by the consistency of $\hat{\sigma}^2$,

$$\hat{\omega} = \frac{r-2}{r}\frac{1}{\hat{F}} \to_p \frac{1}{\lambda^2+1}. \tag{A.38}$$

This concludes the proof of Lemma A.2. ■

A.2.3 Main proof of Theorem 1

After some rearrangements, we can write the object of interest (10) as

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) &= p^{-1/2}\left[(\hat{\boldsymbol{\beta}}_p^a - \boldsymbol{\beta}_p)\hat{\boldsymbol{\Sigma}}_{\beta,p}^{-1}(\hat{\boldsymbol{\beta}}_p^a - \boldsymbol{\beta}_p) - p + (r-2)\hat{\omega}\right] \\ &= p^{-1/2}\left[\frac{\sigma^2}{\hat{\sigma}^2}\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n} - 2\hat{\omega}\boldsymbol{P}_{R,n}\right)\boldsymbol{\varepsilon}_n - p + 2\hat{\omega}r\right] \\ &\quad - 2\hat{\omega}\hat{\sigma}^{-2}(pn)^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n + 4p^{-1/2}\hat{\omega} \\ &= p^{-1/2}\left[\frac{\sigma^2}{\hat{\sigma}^2}\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n} - 2\omega\boldsymbol{P}_{R,n}\right)\boldsymbol{\varepsilon}_n - p + 2\omega r\right] \\ &\quad - 2\omega\hat{\sigma}^{-2}(pn)^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n + o_p(1), \end{aligned} \tag{A.39}$$

where the third equality uses that $\hat{\omega} \to_p \omega \leq M$.

The estimation uncertainty in $\hat{\sigma}^2$ will have an important effect in the high-dimensional setting. We follow Anatolyev (2012) and rewrite

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}_p^a, \boldsymbol{\beta}_p) &= p^{-1/2}\left[\left(1 + \frac{\hat{\sigma}^2}{\sigma^2} - 1\right)^{-1}\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n} - 2\omega\boldsymbol{P}_{R,n}\right)\boldsymbol{\varepsilon}_n - p + 2\omega r\right] \\ &\quad - 2\omega\sigma^{-2}(pn)^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n + o_p(1) \\ &= p^{-1/2}\left[\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n} - 2\omega\boldsymbol{P}_{R,n}\right)\boldsymbol{\varepsilon}_n - p + 2\omega r\right] \\ &\quad - p^{-1/2}(p - 2\omega r)\left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) - 2\omega\sigma^{-2}(pn)^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n \\ &\quad - p^{-1/2}\left[\left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right)\left(\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n} - 2\hat{\omega}\boldsymbol{P}_{R,n}\right)\boldsymbol{\varepsilon}_n - p + 2\hat{\omega}r\right)\right] + o_p(1) \\ &= p^{-1/2}\left[\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n} - 2\omega\boldsymbol{P}_{R,n} - \frac{p-2\omega r}{n-k}\boldsymbol{M}_{X_{n,k}}\right)\boldsymbol{\varepsilon}_n\right] \\ &\quad - 2\omega\sigma^{-2}(pn)^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{X}_{n,k}'\boldsymbol{\varepsilon}_n + o_p(1). \end{aligned} \tag{A.40}$$

In the above, the second equality follows from a Taylor expansion that uses

$\left(1+\frac{\hat{\sigma}^2}{\sigma^2}-1\right)^{-1} = 1-\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)+O_p(n^{-1/2})$, which holds by Lemma A.2. To obtain the final line, we use the definition of $\hat{\sigma}^2$ and the fact that

$$p^{-1/2}\left[\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)\left(\frac{1}{\sigma^2}\boldsymbol{\varepsilon}_n'\left(\boldsymbol{P}_{G,n}-2\omega\boldsymbol{P}_{R,n}\right)\boldsymbol{\varepsilon}_n-p+2\omega r\right)\right]=o_p(1), \tag{A.41}$$

since the second term in round brackets multiplied by $p^{-1/2}$ is $O_p(1)$ by (A.36), while the first term in round brackets is $O_p(n^{-1/2})$ by Lemma A.2.

Define now

$$\begin{aligned}\boldsymbol{A}_n &= \boldsymbol{P}_{G,n}-2\omega\boldsymbol{P}_{R,n}-\frac{p-2\omega r}{n-k}\boldsymbol{M}_{X_{n,k}}\\ W_i &= -2\omega\sigma^{-2}(pn)^{-1/2}\boldsymbol{h}'\boldsymbol{S}_k\boldsymbol{x}_{i,k}\varepsilon_i.\end{aligned} \tag{A.42}$$

We can then write

$$\begin{aligned}D(\hat{\boldsymbol{\beta}}_p^a,\boldsymbol{\beta}_p) &= \sum_{i=1}^n W_i+\frac{1}{p^{1/2}}\frac{1}{\sigma^2}\sum_{i\neq j}\varepsilon_i\varepsilon_j A_{ij}+\frac{1}{p^{1/2}}\frac{1}{\sigma^2}\sum_{i=1}^n A_{ii}\varepsilon_i^2\\ &= \sum_{i=1}^n W_i+\frac{1}{p^{1/2}}\frac{1}{\sigma^2}\sum_{i\neq j}\varepsilon_i\varepsilon_j A_{ij}+o_p(1).\end{aligned} \tag{A.43}$$

where we used that the final term on the first line has expectation 0 and variance $\frac{\mathrm{E}[\varepsilon_i^4]}{\sigma^4}\frac{1}{p}\sum_{i=1}^n A_{ii}^2 = o_{a.s.}(1)$ by Assumption 3. The dominated convergence theorem then again shows that the final term is $o_p(1)$.

To apply Lemma A.1, we verify the underlying assumptions. For Assumption (i), notice that $\boldsymbol{P}_{G,n}$, $\boldsymbol{P}_{R,n}$, $\boldsymbol{M}_{X_{n,k}}$ are idempotent and satisfy $\boldsymbol{P}_{G,n}\boldsymbol{P}_{R,n}=\boldsymbol{P}_{R,n}$ and $\boldsymbol{P}_{G,n}\boldsymbol{M}_{X_{n,k}}=\boldsymbol{O}_n$, $\boldsymbol{P}_{R,n}\boldsymbol{M}_{X_{n,k}}=\boldsymbol{O}_n$. We therefore have

$$\begin{aligned}\boldsymbol{A}_n^2 &= \boldsymbol{P}_{G,n}+4\omega(\omega-1)\boldsymbol{P}_{R,n}+\left(\frac{p-2\omega r}{n-k}\right)^2\boldsymbol{M}_{X_{n,k}},\\ \boldsymbol{A}_n^3 &= \boldsymbol{P}_{G,n}-2\omega(4\omega^2-6\omega+3)\boldsymbol{P}_{R,n}-\left(\frac{p-2\omega r}{n-k}\right)^3\boldsymbol{M}_{X_{n,k}}\\ \boldsymbol{A}_n^4 &= \boldsymbol{P}_{G,n}+8\omega(2\omega^3-4\omega^2+3\omega-1)\boldsymbol{P}_{R,n}+\left(\frac{p-2\omega r}{n-k}\right)^4\boldsymbol{M}_{X_{n,k}}.\end{aligned} \tag{A.44}$$

where $0\leq\omega\leq 1$. We can now verify Assumption (i). For instance, under

Assumption 1, and using that the diagonal elements of a projection matrix are ≤ 1,

$$\begin{aligned} \boldsymbol{e}_i' \boldsymbol{A}_n^2 \boldsymbol{e}_i &\leq 1 + 3\omega(\omega-1) + \left(\frac{p-2\omega r}{n-k}\right)^2 \leq M, \\ \operatorname{tr}(\boldsymbol{A}_n^2) &= p + 4\omega(\omega-1)r + \left(\frac{p-2\omega r}{n-k}\right)(p-2\omega r) \leq Mp. \end{aligned} \tag{A.45}$$

The remainder of Assumption (i) follows analogously.

The first part of Assumption (ii) of Lemma A.1 follows from Assumption 2. We also have by Assumption 3, $D = \sum_{i=1}^n \mathrm{E}[W_i^2 | \boldsymbol{X}_{n,k}] = 4\omega^2 \frac{r}{p} \bar{\lambda}^2 \to_{a.s.} 4\omega^2 \lambda^2 \frac{\rho}{\pi} \leq M$ so that the second part of Assumption (ii) also holds. Assumption (iii) follows from Assumption 2. Assumption (iv) follows from Assumption 3. Assumption (v) follows from Assumption 1.

Finally, we need to verify that Ξ is bounded from below $a.s.n.$ Notice that

$$\begin{aligned} \bar{\Sigma} &= \frac{2\sigma^4}{p}\left(\operatorname{tr}(\boldsymbol{A}_n^2) - \sum_{i=1}^n A_{ii}^2\right) \\ &= \frac{2\sigma^4}{p}\operatorname{tr}(\boldsymbol{A}_n^2) + o_{a.s.}(1) \\ &= 2\left[1 + 4\omega(\omega-1)\frac{r}{p} + \frac{p-2\omega r}{n-k}\left(1 - 2\omega\frac{r}{p}\right)\right] + o_{a.s.}(1), \\ D &= 4\omega^2\lambda^2\frac{r}{p} + o_{a.s.}(1). \end{aligned} \tag{A.46}$$

Using that $\omega = (\lambda^2+1)^{-1}$, i.e. $\lambda^2 = \omega - 1$, the minimal value of Ξ equals

$$\begin{aligned} \Xi_{\min} &= \frac{(2p-r)(n-k)+2p^2}{p(n-k+2r)} + o_{a.s.}(1) \\ &\geq \frac{(2p-p)(n-k)+2p^2}{p(n-k+2p)} + o_{a.s.}(1) \\ &= 1 + o_{a.s.}(1). \end{aligned} \tag{A.47}$$

Hence, Ξ is bounded from below $a.s.n.$

We now conclude from Lemma A.1 that

$$\Xi^{-1/2} D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p) \Rightarrow N(0,1), \quad a.s. \tag{A.48}$$

Following the argument at the top of p. 81 of Chao et al. (2012), the unconditional probability $\mathrm{P}(\Xi^{-1/2}D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p) \leq y) = \mathrm{E}[\mathrm{P}(\Xi^{-1/2}D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p) \leq y|\boldsymbol{X}_{n,k})]$. Since for some $\epsilon > 0$, $\sup_n \mathrm{E}[|\mathrm{P}(\Xi^{-1/2}D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p) \leq y)|^{1+\epsilon}] < \infty$, the convergence holds unconditionally, i.e. $\mathrm{P}(\Xi^{-1/2}D(\boldsymbol{\beta}_p^a, \boldsymbol{\beta}_p) \leq y) \to \Phi(y)$.

Moreover, using (A.46) and the fact that and $\omega = (\lambda^2+1)^{-1}$, we have

$$\Xi = D + \bar{\Sigma} = 2 - 4\frac{c}{\lambda^2+1}\frac{\rho}{\pi} + 2\frac{\pi}{1-\kappa} - 8\frac{\rho}{1-\kappa}\frac{1}{\lambda^2+1}\left(1 - \frac{\rho}{\pi}\frac{1}{\lambda^2+1}\right) + o_p(1). \tag{A.49}$$

This implies that $\Xi \to_p \tau^2 = 2 - 4\frac{\lambda^2}{\lambda^2+1}\frac{\rho}{\pi} + 2\frac{\pi}{1-\kappa} - 8\frac{\rho}{1-\kappa}\frac{1}{\lambda^2+1}\left(1 - \frac{\rho}{\pi}\frac{1}{\lambda^2+1}\right)$. Hence, as $(r, n \to \infty)$, $N \Rightarrow N(0, \tau^2)$. ■

A.3 Power

Denote by $T(\boldsymbol{\beta}_p^0)$ the test statistic under the parameter vector $\boldsymbol{\beta}_p^0$. Then,

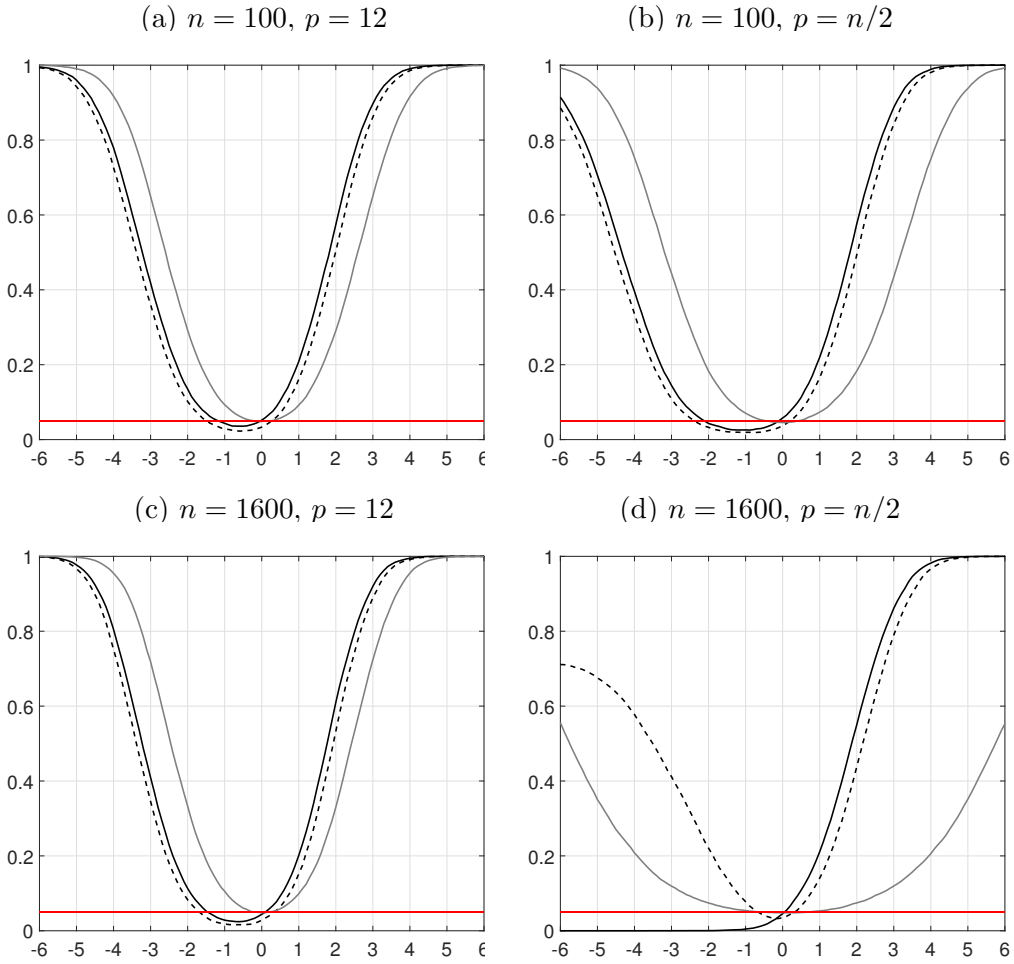
$$\begin{aligned}
T(\boldsymbol{\beta}_p^0) &= T(\boldsymbol{\beta}_p) + p^{-1/2}n\left[(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^0)'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^0) + 2(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^0)'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}(\hat{\boldsymbol{\beta}}_p^a - \boldsymbol{\beta}_p)\right] \\
&= T(\boldsymbol{\beta}_p) + p^{-1/2}n\left[(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^0)'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^0)\right. \\
&\qquad \left. + 2(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^0)'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}(\tilde{\boldsymbol{\beta}}_p - \mathrm{E}[\tilde{\boldsymbol{\beta}}_p]) + (1-\hat{\omega})(\hat{\boldsymbol{\delta}}_p - \boldsymbol{\delta}_p) - \hat{\omega}\boldsymbol{\delta}_p)\right] \\
&= T(\boldsymbol{\beta}_p) + p^{-1/2-2\gamma}\left[\boldsymbol{h}_{0,n}'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}\boldsymbol{h}_{0,n} - 2p^{\gamma-1/2}\hat{\omega}\boldsymbol{h}_{0,n}'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}\boldsymbol{h}_p\right] \\
&\qquad + 2p^{-1/2-\gamma}\boldsymbol{h}_{0,n}\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}\sqrt{n}(\tilde{\boldsymbol{\beta}}_p - \mathrm{E}[\tilde{\boldsymbol{\beta}}_p] + (1-\hat{\omega})(\hat{\boldsymbol{\delta}}_p - \boldsymbol{\delta}_p)).
\end{aligned} \tag{A.50}$$

The second line is asymptotically mean zero with finite asymptotic variance when $\gamma = 0$ since $\boldsymbol{h}_{0,n}$ satisfies Assumption 3. Since $\gamma > 0$, Chebyshev's theorem then gives that

$$T(\boldsymbol{\beta}_p^0) = T(\boldsymbol{\beta}_p) + p^{-1/2-\gamma}\left[\boldsymbol{h}_{0,n}'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}\boldsymbol{h}_{0,n} - 2p^{\gamma-1/2}\hat{\omega}\boldsymbol{h}_{0,n}'\hat{\boldsymbol{\Sigma}}_{u,n}^{-1}\boldsymbol{h}_p\right] + o_p(1). \tag{A.51}$$

This completes the proof. ■

Figure 3: Linear regression model: power, restricted estimator with $s = 3$.



Note: see the note following Figure 2 with the change that $[\boldsymbol{b}]_i = \sqrt{p/3n}$ when $i = 1, 2, 3$ and 0 otherwise.

Appendix B Additional simulations results

B.1 Imposing less restrictions

The restricted estimator in Section 4 sets all but the first coefficient equal to zero. Here we consider an estimator that sets all but the first three coefficients equal to zero. In this case, we add the fixed vector $\boldsymbol{b}$ with $[\boldsymbol{b}]_i = \sqrt{p/(3 \cdot n)}$ to the restricted estimator $\bar{\boldsymbol{\beta}}$. The division by 3 is motivated by reconsidering the example below Theorem 2 and assuming that the first three coefficients are nonzero and equal.

The power curves shown in Figure 3 show that for positive values of

the coefficients, the power difference is quite small. For negative values, we find that the power increases somewhat when allowing the first three coefficients, instead of only the first one, to be different from zero.

B.2 Benchmark confidence regions

Consider the standard definition of a spherical confidence region.

Definition 1 *For any estimator $\bar{\boldsymbol{\beta}}_p$ of the parameter vector of interest $\boldsymbol{\beta}_p$, and critical values $\hat{b}$, the confidence region is defined in terms of the loss (9) as*

$$C(\bar{\boldsymbol{\beta}}_p, \hat{b}) = \left\{ \boldsymbol{t}_p : \ell(\bar{\boldsymbol{\beta}}_p, \boldsymbol{t}_p) \leq \hat{b}^2 \right\}. \tag{B.1}$$

The confidence regions by Casella and Hwang (1983) are given by Definition 1 centered at the averaging estimator with weights (8) restricted to be less or equal then 1. Defining $R(x) = 1 - \frac{p-2}{x}$, the critical values are given in their equation 4.7 as

$$\hat{b}^2_{CH} = \begin{cases} R(b^2_\chi)\left[b^2_\chi - p\log(R(b^2_\chi))\right] & \text{if } n\hat{q}_n \leq b^2_\chi, \\ R(n\hat{q}_n)[b^2_\chi - p\log(R(n\hat{q}_n))] & \text{if } n\hat{q}_n > b^2_\chi, \end{cases} \tag{B.2}$$

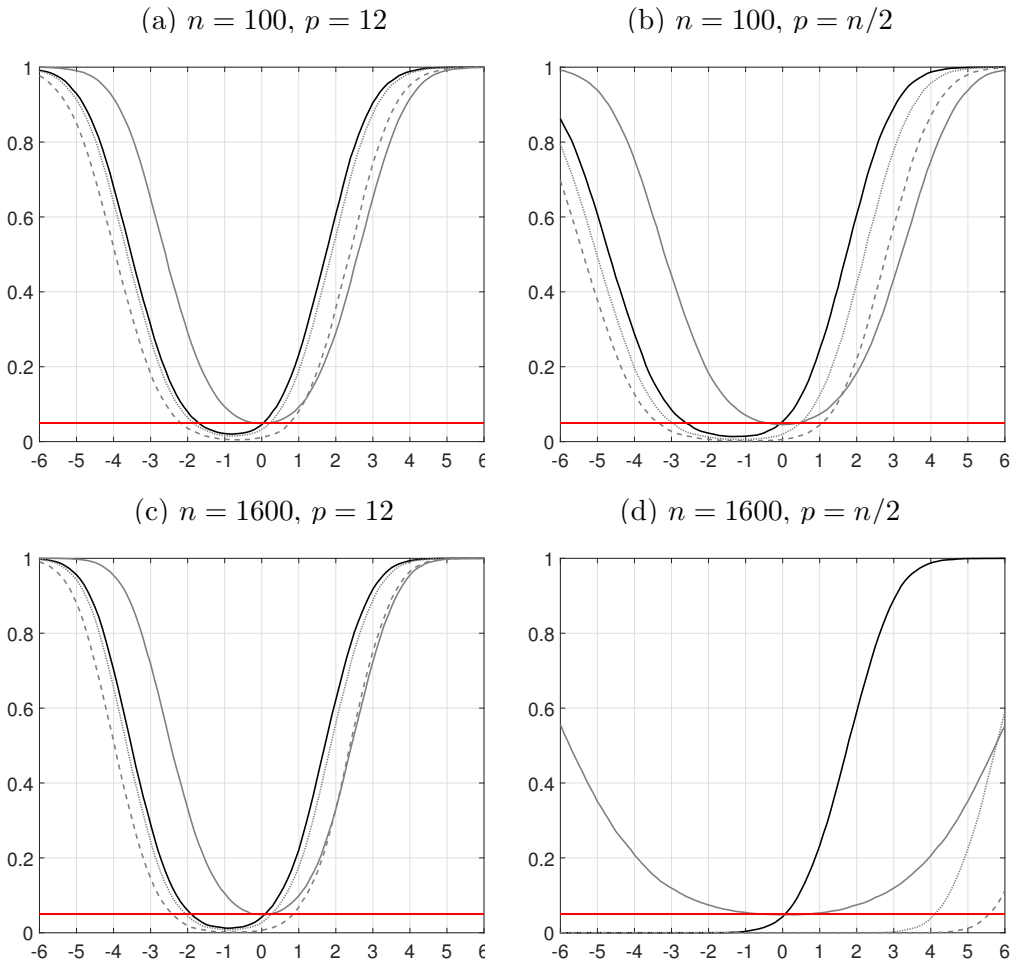
where b^2_χ indicates the critical values from a $\chi^2(p)$ distribution.

Samworth (2005) Taylor expands $(\hat{\boldsymbol{\beta}}^a_p - \boldsymbol{\beta}_p)'\boldsymbol{\Sigma}^{-1}_{\beta,p}(\hat{\boldsymbol{\beta}}^a_p - \boldsymbol{\beta}_p)$ around $\boldsymbol{\beta}_p = \mathbf{0}_p$ to get

$$\begin{aligned} \hat{b}^2_S &= \min\left\{ w_\alpha(0) + \frac{1}{2}w''_\alpha(0)n\hat{q}_n, b^2_\chi \right\}, \quad w_\alpha(0) = \left(b_\chi - \frac{p-2}{b_\chi}\right)^2, \\ w_\alpha(0)'' &= \frac{2}{k}\left(1 - \frac{(p-2)}{b^2_\chi}\right)\left[\frac{(p-2)(p-1)}{b^2_\chi + p - 2} - \frac{2(p-2)b^2_\chi}{(b^2_\chi + p - 2)^2} + \frac{(p-2)^2}{b^2_\chi + p - 2}\right] \\ &\quad + \frac{2(p-2)(p-1)}{b^2_\chi \cdot p}. \end{aligned} \tag{B.3}$$

In Figure 4, we compare the power of the confidence regions based on (10) (one-sided) to those by Casella and Hwang (1983) and Samworth (2005). We see that for a small number of parameters of interest ($p = 12$), the difference between the regions is small, although the procedure by Samworth (2005) is somewhat conservative. When the sample size increases

Figure 4: Linear regression model: power compared to alternatives.



Note: see the note follow Figure 2. Gray dashed lines correspond to the regions proposed by Samworth (2005), gray dotted lines to the regions proposed by Casella and Hwang (1983).

and we consider a high-dimensional setting where $p = n/4$, the alternative procedures are very conservative, and show substantial loss of power.