Choosing the Dictionary and Penalty for IV-LASSO*

Yukun Ma
University of Rochester
yma69@ur.rochester.edu

Manu Navjeevan
Texas A&M University
mnavjeevan@tamu.edu

Bohdan Salahub
University of California, Los Angeles
bsalahub@g.ucla.edu

September 6, 2026

Abstract

Estimating the first stage of an instrumental variables (IV) model with the least absolute shrinkage and selection operator (LASSO) requires choosing a dictionary of technical instruments and a penalty level. First-order asymptotic theory offers no guidance on these choices, as any consistent implementation yields a structural parameter estimator with the same limiting distribution. In finite samples, however, these choices can have a substantial impact on the resulting structural parameter estimate. Working in a model with a single endogenous regressor and homoskedastic Gaussian errors, we use first- and second-order Stein identities to derive the approximate mean squared error (AMSE) of the instrumental-variables LASSO (IV-LASSO) estimator, which can be consistently estimated and used to rank a prespecified list of dictionary–penalty candidates. The AMSE reveals a bias–variance trade-off: more complex first-stage fits better approximate the conditional mean of the endogenous variable but are also more correlated with the structural errors, with complexity measured by the degrees of freedom of the LASSO fit. The weight on this bias rises with the endogeneity of the regressor, a quantity that neither plug-in nor cross-validation penalty rules take into account. Despite the AMSE being derived in a Gaussian model, penalty selection by minimizing the feasible AMSE criterion delivers up to a one-third lower mean squared error compared to cross-validation and plug-in penalty rules in Gaussian and non-Gaussian simulation designs calibrated to the data of Gilchrist and Sands (2016).

Keywords: Instrumental Variables, LASSO, Model Selection, Second-Order Asymptotics

JEL Codes: C26, C36, C52

1 Introduction

When a large set of technical instruments is available, LASSO makes estimation of an IV first stage feasible but leaves applied researchers to choose how heavily to penalize the fit and which dictionary of technical instruments to use. Under standard sparse-first-stage conditions, Belloni et al. (2012) show that all sufficiently

*An R package implementing IV-LASSO with the first-stage implementation guided by this paper can be found on GitHub. This package, along with the proofs of several technical lemmas in this paper, were developed with the help of Claude Fable. We take responsibility for all errors.

accurate LASSO fits yield IV estimators with the same limiting distribution. Standard asymptotic theory therefore does not rank admissible implementations, even though their finite-sample performance can differ. Figure 1 illustrates this issue in the eminent-domain application of Belloni et al. (2012). Across three nested first-stage dictionaries, the IV estimate changes materially along the penalty path. Moreover, different methods of selecting the first-stage penalty parameter can change whether the 95% confidence interval for the resulting structural parameter estimate contains zero. For example, for the Case-Shiller home price index outcome, using cross-validation (Chetverikov et al., 2021) yields a significant result on a restricted instrument dictionary while using Stein's Unbiased Risk Estimate (SURE, Tibshirani and Taylor (2012)) does not. Given this variation, how should an applied researcher choose among possible first-stage implementations?

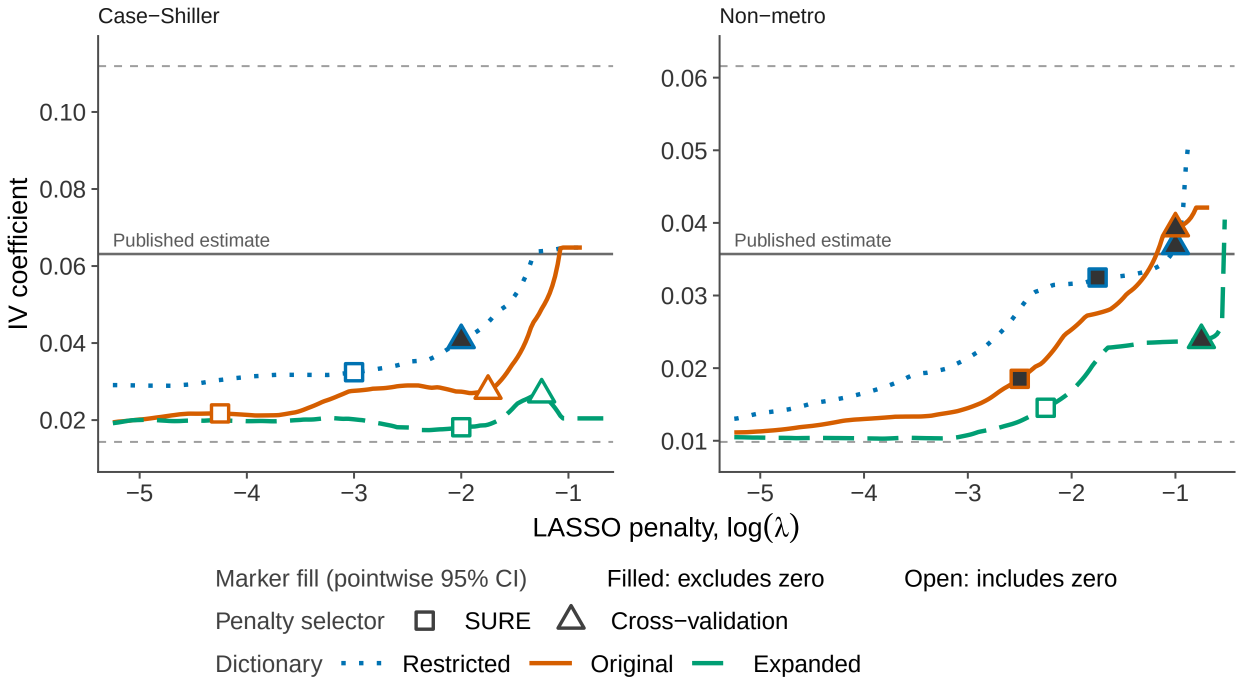

Figure 1: IV Estimates and Conventional Significance across First-Stage Implementations

Notes: Lines trace IV-LASSO estimates over LASSO penalties within three nested dictionaries. Original is the published dictionary; Restricted removes cross-products, and Expanded adds all nondegenerate second-order products. Filled markers denote pointwise 95% heteroskedasticity robust confidence intervals that exclude zero. The solid gray line and dashed bounds show the published IV-post-LASSO estimate and confidence interval. The panels report the two outcomes for which that interval excludes zero (183 Case–Shiller and 110 non-metro circuit-year observations).

Existing first-stage tuning rules are not designed to maximize the precision of resulting structural parameter estimates. The calibrated penalty rule of Belloni et al. (2012) and the bootstrap-after-cross-validation method of Chetverikov and Sørensen (2025) are designed to control first-stage estimation error, while SURE and ordinary cross-validation target first-stage prediction (Tibshirani and Taylor, 2012; Chetverikov et al., 2021). These objectives do not account for how first-stage estimation noise enters the structural estimator. A LASSO fit with a lighter penalty or richer dictionary may better approximate the true first-stage, but a more complex fit will also be more correlated with the endogenous variable. When the first-stage and structural errors are correlated, this in turn introduces bias into the structural estimate. The optimal first-stage implementation therefore depends not only on goodness of first-stage fit and instrument strength, but also on the level of endogeneity.

This paper develops a principled method for choosing the first-stage implementation by developing second-order asymptotic theory for the full-sample IV-LASSO estimator. We focus on the leading case of a single endogenous regressor and work in a homoskedastic Gaussian model. Following the higher-order tradition of Nagar (1959) and Donald and Newey (2001), we prespecify a finite list of first-stage LASSO dictionary–penalty candidates and rank them by the approximate mean squared error (AMSE) of their resulting IV-LASSO estimators. The candidate specific part of the AMSE can be decomposed into a first-stage approximation cost and a complexity cost, with complexity measured by a collinearity-robust analogue of the number of instruments selected by the LASSO. The size of this complexity cost rises with the squared covariance between the first-stage and structural errors, reflecting the discussion above. Our characterization yields a feasible criterion that can be constructed from each candidate's fitted first-stage. Minimizing this criterion selects an estimator whose AMSE is as small as that of the best candidate in the list, up to vanishing relative error.

Deriving the criterion is nonstandard. Unlike ordinary least-squares, the LASSO coefficient estimate does not yield a closed-form expression and the LASSO fit is not a fully differentiable function of the data. These features prevent standard tools from being applied to analyze higher-order terms in the asymptotic expansion. We overcome these obstacles by leveraging a geometric characterization of the LASSO-fit from Tibshirani and Taylor (2012) along with newly developed second-order Stein identities from Bellec and Zhang (2021) to examine otherwise problematic second-order terms. In addition to the distributional restrictions stated above, the result requires standard sparsity conditions to ensure convergence of the first-stage estimates as well as a partial beta-min condition which allows for consistent estimation of the complexity of the LASSO fits. This partial beta-min condition requires that a growing core of first-stage coefficients must be sufficiently strong and is notably weaker than previous beta-min conditions in the literature on model selection via the LASSO (Zhao and Yu, 2006; Wainwright, 2009; Belloni and Chernozhukov, 2013).

We evaluate the performance of the proposed criterion through a simulation study calibrated to the data of Gilchrist and Sands (2016), who use weather conditions to instrument for opening-weekend movie ticket sales. We hold the original instruments fixed and vary the strength of identification and degree of endogeneity. At low levels of endogeneity, cross-validation and the proposed criterion perform comparably, in line with the theory developed in this paper. As the level of endogeneity rises, however, cross-validation fails to account for the bias passed on to the second-stage estimator. In contrast, our proposed criterion adjusts for the endogeneity and selects first-stages with lower levels of complexity, resulting in structural parameter estimates with up to a third lower mean-squared error than their cross-validation counterparts in designs with the highest levels of endogeneity. Interestingly, the relative performance of the criterion is essentially the same in both Gaussian and non-Gaussian designs. This suggests that, although our theory is developed in a Gaussian model, the AMSE of the IV-LASSO estimator in a general model may be well approximated by its AMSE in the Gaussian model.

We contribute to a growing literature on the use of high-dimensional or "machine-learning" methods in econometric analysis (Belloni et al., 2014; Chernozhukov et al., 2018; Wager and Athey, 2018; Farrell et al., 2021) as well as to an established literature analyzing higher-order properties of econometric estimators (Sargan, 1976; Rothenberg, 1984; Newey and Smith, 2004; Hahn et al., 2004). A common feature of double/debiased machine learning estimators is that all suitably convergent nuisance parameter estimation procedures yield the same asymptotic distribution for the resulting target parameter estimator. Existing asymptotic theory therefore provides little guidance on how to choose among implementations satisfying certain high-level conditions. To our knowledge, we provide the first second-order mean squared error approximation for an econometric estimator with a high-dimensional LASSO first stage.[1] We show that the higher-order terms distinguishing alternative first-stage implementations can be consistently estimated and use them to select the first-stage dictionary and penalty. Our results thus extend the literature on selecting and regularizing instruments according to the approximate mean squared error of the structural parameter estimator (Donald and Newey, 2001; Kuersteiner and Okui, 2010; Okui, 2011; Carrasco, 2012) to high-dimensional LASSO first stages.

The rest of this paper proceeds as follows. Section 2 sets up the model, the candidate first stages, and the effective dimension. Section 3 states the conditions and the two main results. Section 4 reports the calibrated simulation study. All proofs are deferred to the appendix.

Notation. For $m \in \mathbb{N}$, let $[m] := \{1, \dots, m\}$. For an array $\{a_i\}_{i \in [n]}$, define

$$\mathbb{E}_n[a_i] := \frac{1}{n} \sum_{i \in [n]} a_i, \qquad \overline{\mathbb{E}}[a_i] := \mathbb{E}_n[\mathbb{E}[a_i]].$$

2 Model and Setup

To simplify exposition, we present the analysis assuming that any exogenous regressors have already been partialled out of both the outcome and the endogenous regressor.[2] Our analysis assumes there is only a single endogenous variable and does not cover the general case where x_i is vector-valued. Along with the first-stage, the IV model can be written as a system of simultaneous equations:

$$\begin{aligned} y_i &= \beta x_i + \varepsilon_i \\ x_i &= \Pi_i + v_i \end{aligned} \tag{1}$$

We assume the researcher observes n independent observations of the outcome $y_i \in \mathbb{R}$, the endogenous variable $x_i \in \mathbb{R}$, and instruments $z_i \in \mathbb{R}^p$ but neither the structural error $\varepsilon_i \in \mathbb{R}$ nor the first-stage errors $v_i \in \mathbb{R}$. The structural error $\varepsilon_i \in \mathbb{R}$ is assumed to be conditional-mean independent of the instruments, $\mathbb{E}[\varepsilon_i | z_i] = 0$ and we let $\Pi_i = \mathbb{E}[x_i | z_i]$ denote the conditional expectation of the endogenous variable given the instruments so that $\mathbb{E}[v_i | z_i] = 0$. We collect the random variables in the vectors

$$\begin{aligned} x &:= (x_1, \dots, x_n)' \in \mathbb{R}^n & \qquad \Pi &:= (\Pi_1, \dots, \Pi_n)' \in \mathbb{R}^n \\ \varepsilon &:= (\varepsilon_1, \dots, \varepsilon_n)' \in \mathbb{R}^n & \qquad v &:= (v_1, \dots, v_n)' \in \mathbb{R}^n \end{aligned}$$

Throughout the theoretical analysis, we condition on the realized instruments $\{z_i\}_{i=1}^n$ and treat them as fixed. All expectations and probability statements below are therefore understood as conditional on this realization, i.e, $\Pi_i = \mathbb{E}[x_i]$ and $\mathbb{E}[\varepsilon_i] = \mathbb{E}[v_i] = 0$. We will assume strong identification and conditionally homoskedastic Gaussian errors. Formally,

Assumption 1 (Conditional Gaussian Sampling)**.** *Assume*

(i) The errors $\{(\varepsilon_i, v_i)\}_{i=1}^n$ are independent of the instruments and independently and identically distributed

[1] Velez (2026) develops second-order approximations for double/debiased machine learning estimators to study the choice of the number of cross-fitting folds. His higher-order analysis requires the first-stage estimators to admit a stochastic linear expansion. Our analysis instead directly accounts for the shrinkage and variable selection induced by the LASSO.

[2] An appendix providing a formal treatment of the model with controls is available on request.

according to

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N(0, \Omega), \qquad \Omega := \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{pmatrix}, \tag{2}$$

where Ω is a fixed nonsingular matrix.

(ii) For constants $0 < \underline{H} < \overline{H} < \infty$,

$$\underline{H} \leq H := \mathbb{E}_n[\Pi_i^2] \leq \overline{H}, \qquad \sigma_{\varepsilon v} \neq 0. \tag{3}$$

Assumption 1(i) encodes the assumption of conditionally homoskedastic Gaussian errors while the first part of Assumption 1(ii) encodes the strong identification assumption. The second part of Assumption 1(ii) is a mild technical condition that assumes that there is an endogeneity problem to begin with, i.e β cannot be recovered from a simple regression of the outcome on the endogenous variable.

Under conditional homoskedasticity, the conditional mean Π_i is the optimal instrument for estimating β: among IV estimators based on a nonrandom instrument, using Π_i itself minimizes the variance of the leading term.[3] Since Π_i is unknown, feasible estimation replaces it with a first-stage estimate $\widehat{\Pi}_i$. First-order asymptotics provide no guidance on this choice. Any estimator $\widehat{\beta}$ based on a sufficiently well-estimated first-stage (Section 3 states sufficient conditions) admits the same expansion,

$$\sqrt{n}(\widehat{\beta} - \beta) = \frac{\sqrt{n}\,\mathbb{E}_n[\Pi_i \varepsilon_i]}{H} + o_p(1), \tag{4}$$

whose leading term does not involve the first-stage estimate and, conditional on the instruments, is distributed exactly as $N(0, \sigma_\varepsilon^2/H)$ at every sample size. Comparing first-stage estimators thus requires analyzing terms hidden in the remainder. We provide a feasible criterion for this comparison when the first stage is estimated by the LASSO (Tibshirani, 1996), as in Belloni et al. (2012), a leading approach when the instrument set is high-dimensional or when the researcher wants to use a high-dimensional basis of technical instruments.

As will be described below, LASSO estimation of the first stage is characterized by a choice of dictionary, i.e., the choice of instrument basis, and a penalty parameter. The criterion developed in Section 3 ranks these candidate dictionary–penalty pairs by analyzing second-order terms in a mean-squared-error expansion of the resulting second-stage estimator. Our analysis is closely related to that of Donald and Newey (2001), who provide a similar criterion when the first stage is estimated by ordinary least squares (OLS) on a chosen instrument basis.

2.1 Candidate First Stages and the IV Estimator

We assume that the researcher is interested in selecting a first-stage specification c from a fixed number J of candidates. Each candidate $c \in [J]$ consists of a dictionary of technical instruments and a penalty level $\lambda_c > 0$. For each observation, candidate c's dictionary maps the observed instruments z_i into a vector of technical instruments $z_{c,i} := (z_{c,i1}, \ldots, z_{c,ip_c})' \in \mathbb{R}^{p_c}$, whose entries are fixed functions of z_i, such as powers, interactions, or indicators. The dictionary dimension p_c may exceed the sample size n. Although λ_c is

[3] For nonrandom $g \in \mathbb{R}^n$ with $\mathbb{E}_n[g_i^2]$ bounded and $\mathbb{E}_n[g_i \Pi_i]$ bounded away from zero, the IV estimator $\widehat{\beta}_g := \mathbb{E}_n[g_i y_i]/\mathbb{E}_n[g_i x_i]$ satisfies $\sqrt{n}(\widehat{\beta}_g - \beta) = \sqrt{n}\,\mathbb{E}_n[g_i \varepsilon_i]/\mathbb{E}_n[g_i \Pi_i] + o_p(1)$, and the leading term is distributed exactly as $N(0, \sigma_\varepsilon^2 \mathbb{E}_n[g_i^2]/\mathbb{E}_n[g_i \Pi_i]^2)$. By Cauchy–Schwarz this variance is at least σ_ε^2/H, with equality if and only if g is proportional to Π. When the instruments are instead modeled as random draws, $\sigma_\varepsilon^2/\mathbb{E}[\Pi_i^2]$ is the semiparametric efficiency bound for β under conditional homoskedasticity (Chamberlain, 1987; Newey, 1990); σ_ε^2/H is its fixed-design counterpart.

nonrandom, both the dictionary and the penalty level may depend on n. We collect the resulting technical instruments in the matrix

$$Z_c := (z_{c,1}, \ldots, z_{c,n})' \in \mathbb{R}^{n \times p_c}, \qquad Z_{c,B} := (Z_c)_{\cdot,B} \quad \text{for } B \subseteq [p_c],$$

Thus, the ith row of Z_c is $z'_{c,i}$, while $Z_{c,B} \in \mathbb{R}^{n \times |B|}$ contains the columns indexed by B.

For candidate c and an arbitrary response vector $w := (w_1, \ldots, w_n) \in \mathbb{R}^n$, define the LASSO coefficient estimate and fitted vector (Tibshirani, 1996) by

$$\widehat{\pi}_c(w) \in \arg\min_{\pi \in \mathbb{R}^{p_c}} \left\{ \frac{1}{2}\mathbb{E}_n[(w_i - z'_{c,i}\pi)^2] + \lambda_c \|\pi\|_1 \right\}, \qquad \mu_c(w) := Z_c \widehat{\pi}_c(w). \tag{5}$$

When the columns of Z_c are linearly dependent, the coefficient minimizer need not be unique, but the fitted vector $\mu_c(w)$ remains unique (Tibshirani and Taylor, 2012; Tibshirani, 2013). Applying this procedure to the endogenous variable x, define $\widehat{\pi}_c := \widehat{\pi}_c(x)$, $\widehat{\Pi}_c := \mu_c(x)$, and $\widehat{\Pi}_{c,i} := z'_{c,i}\widehat{\pi}_c$. The corresponding second-stage IV estimator is

$$\widehat{\beta}_c := \begin{cases} \dfrac{\mathbb{E}_n[\widehat{\Pi}_{c,i} y_i]}{\mathbb{E}_n[\widehat{\Pi}_{c,i} x_i]}, & \mathbb{E}_n[\widehat{\Pi}^2_{c,i}] > 0, \\ 0, & \mathbb{E}_n[\widehat{\Pi}^2_{c,i}] = 0. \end{cases} \tag{6}$$

A natural criterion for comparing the candidate estimators $\widehat{\beta}_c$ is their scaled mean squared error (MSE), $\mathbb{E}[n(\widehat{\beta}_c - \beta)^2]$. The exact MSE of an IV estimator, however, need not exist in finite samples. To see this, consider the infeasible estimator that uses the optimal instrument Π_i in place of $\widehat{\Pi}_{c,i}$ in (6). This estimator satisfies

$$\widehat{\beta} - \beta = \frac{\mathbb{E}_n[\Pi_i \varepsilon_i]}{\mathbb{E}_n[\Pi_i x_i]}. \tag{7}$$

Under Assumption 1, the numerator and denominator on the right-hand side are jointly Gaussian. Because Ω is nonsingular, the numerator is not an exact linear function of the denominator, and their ratio does not have a finite second moment. Thus, even the infeasible estimator based on the optimal instrument has no exact MSE. More generally, when the exact MSE exists, it may be neither tractable nor consistently estimable.

Given this restriction, we instead extend the higher-order approach of Donald and Newey (2001) to the present LASSO setting. Under suitable conditions, we show that the scaled squared error of each estimator $\widehat{\beta}_c$ admits the decomposition

$$n(\widehat{\beta}_c - \beta)^2 = C_0 + Q_c + r_c. \tag{8}$$

Here, C_0 is common to all candidates and captures the component of squared error corresponding to the leading term in (4). By contrast, Q_c captures the second-order contribution of candidate c's first-stage procedure, while r_c is asymptotically negligible in the sense that $\max_{c \in [J]} |r_c| / \mathbb{E}[Q_c] \to_p 0$.

We define $\mathbb{E}[C_0 + Q_c]$ as the approximate mean squared error (AMSE) of $\sqrt{n}(\widehat{\beta}_c - \beta)$. For the infeasible estimator based on the optimal instrument, the expectation of the candidate-specific component is zero. Hence, $\bar{Q}_c := \mathbb{E}[Q_c]$ measures the excess AMSE of candidate c relative to that benchmark. Because C_0 is common across candidates, ranking candidates by AMSE is equivalent to ranking them by $\bar{Q}_c$. In Section 3, we derive a closed-form asymptotic approximation to $\bar{Q}_c$, construct a consistent feasible counterpart, and use it to select among the first-order-equivalent candidate estimators.

2.2 The Effective Dimension

A key element of the criterion in Section 3 is a measure of the complexity, or "degrees of freedom," of each first-stage estimate. When the first stage is OLS on a linearly independent set of instruments, complexity is simply the number of instruments used. For the LASSO, however, the number of selected instruments is not always well defined. When dictionary columns are collinear, the coefficient solution need not be unique, and different solutions may have different supports, even though the fitted values and the residual are unique.

Let $u_c(w) := w - \mu_c(w)$ denote the LASSO residual and define the equicorrelation set by

$$E_c(w) := \left\{ j \in [p_c] : \left| \mathbb{E}_n[z_{c,ij} u_{c,i}(w)] \right| = \lambda_c \right\}, \tag{9}$$

This is the set of dictionary columns for which the LASSO Karush–Kuhn–Tucker (KKT) inequality binds. Because the residual is unique, $E_c(w)$ does not depend on which coefficient solution is used, and it contains the support of every solution. We define the effective dimension of the first-stage fit as

$$d_c := \operatorname{rank}(Z_{c,E_c(x)}). \tag{10}$$

The quantity d_c is observable and invariant to the choice of coefficient solution. When the LASSO solution is unique, d_c agrees with the number of selected instruments almost surely under our sampling assumption.[4] As shown by Tibshirani and Taylor (2012), the degrees of freedom of the LASSO fit are given by $\mathbb{E}[d_c]$.

3 Main Results

After stating the conditions required for our analysis, this section formally establishes the approximate mean-squared error decomposition in (8) and provides a closed-form asymptotic approximation to the expectation of its candidate-specific term, which serves as an infeasible selection criterion (Theorem 1). We then show that this infeasible criterion, and thus the expectation of the candidate-specific term itself, can be consistently estimated and use the resulting feasible criterion to select the optimal first-stage implementation (Theorem 2). Throughout, a candidate $c \in [J]$ is a dictionary–penalty pair as described in Section 2, and the number of candidates J is treated as fixed.

3.1 Conditions on the Candidate First Stages

Along with the Gaussian sampling and strong identification conditions in Assumption 1, our analysis proceeds under two additional sets of high-level assumptions. The first ensures that each candidate LASSO estimation procedure described in (5) consistently estimates the first stage and thus that each candidate IV estimator in (6) is first-order equivalent. This set of high-level assumptions is standard in the literature on ℓ_1-penalized estimation (Bickel et al., 2009; van de Geer and Bühlmann, 2009; Belloni et al., 2012). The second is a partial beta-min condition used to establish concentration of the LASSO effective dimension. It is similar in spirit to the growing-basis condition in Donald and Newey (2001) but does not require exact support recovery by the LASSO.

[4] A sufficient condition for uniqueness is that the dictionary columns are in general position, which holds with probability one when the dictionary matrix has a joint continuous distribution (Tibshirani, 2013).

To present the assumptions, we begin by introducing the notion of an approximately sparse representation. We say that the first stage Π_i admits an approximately sparse representation in the candidate dictionary Z_c if there is a nonrandom coefficient vector $\pi_c^0 \in \mathbb{R}^{p_c}$ such that

$$\Pi_i = z_{c,i}'\pi_c^0 + \xi_{c,i}, \qquad T_c := \operatorname{supp}(\pi_c^0), \qquad s_c := |T_c|, \tag{11}$$

where $\operatorname{supp}(b) := \{j : b_j \neq 0\}$ and $s_c \geq 1$. The rate restrictions on s_c and the approximation error $\xi_{c,i}$ are stated below. As in Belloni et al. (2012), approximate sparsity can be interpreted either as an assumption that only a few of the instruments in the dictionary are important for explaining variation in the endogenous variable, or as an assumption that the conditional mean Π_i can be accurately approximated using only a small number of dictionary terms.

For a vector δ and an index set T, write $\delta_T := (\delta_j)_{j \in T}$ and $\|\delta\|_0 := |\operatorname{supp}(\delta)|$. Fix a constant $c_\lambda > 1$, common to all candidates, and set

$$L_\lambda := \frac{c_\lambda + 1}{c_\lambda - 1}.$$

Following Belloni et al. (2012), define the restricted eigenvalue (Bickel et al., 2009; van de Geer and Bühlmann, 2009)

$$\kappa_c := \min_{\substack{\delta \neq 0 \\ \|\delta_{T_c^c}\|_1 \leq L_\lambda \|\delta_{T_c}\|_1}} \frac{\sqrt{s_c \mathbb{E}_n[(z_{c,i}'\delta)^2]}}{\|\delta_{T_c}\|_1}, \tag{12}$$

and the upper and lower m-sparse eigenvalues

$$\phi_{\max,c}(m) := \max_{\substack{\delta \neq 0 \\ \|\delta\|_0 \leq m}} \frac{\mathbb{E}_n[(z_{c,i}'\delta)^2]}{\|\delta\|_2^2}, \tag{13}$$

$$\phi_{\min,c}(m) := \min_{\substack{\delta \neq 0 \\ 1 \leq \|\delta\|_0 \leq m}} \frac{\mathbb{E}_n[(z_{c,i}'\delta)^2]}{\|\delta\|_2^2}. \tag{14}$$

The restricted eigenvalue bounds the design from below only over a cone of vectors whose weight is concentrated on the support of the coefficient vector π_c^0. Bounds on sparse eigenvalues require that small collections of dictionary columns be well conditioned. Conditions stated in terms of these quantities are weaker than restricting the full dictionary as larger collections of dictionary columns may still be collinear. In particular, bounds on restricted and sparse eigenvalues do not rule out cases where $p_c \gg n$.

Assumption 2 (First-Stage Convergence)**.** *There are constants $0 \leq C_a < \infty$, $\alpha > 0$, $0 < \kappa_0 \leq 1$, $0 < \underline{\phi} \leq 1 \leq \overline{\phi} < \infty$, and $M > 4L_\lambda^2\overline{\phi}/\kappa_0^2$, common to all candidates, such that the following hold for each $c \in [J]$:*

(i) The dictionary columns are normalized such that $\mathbb{E}_n[z_{c,ij}^2] = 1$ for each $j \in [p_c]$.

(ii) The approximation error satisfies

$$\mathbb{E}_n[\xi_{c,i}^2]^{1/2} \leq C_a\sqrt{\frac{s_c}{n}}.$$

(iii) The penalty satisfies

$$\lambda_c \geq c_\lambda\sqrt{2}\sigma_v\sqrt{\frac{\log(2p_c) + \alpha\log\log(p_c \vee n)}{n}}.$$

(iv) The design satisfies

$$\kappa_c \geq \kappa_0, \qquad \phi_{\max,c}(\lceil M s_c \rceil \wedge p_c) \leq \overline{\phi}, \qquad \phi_{\min,c}(\lceil (M+1) s_c \rceil \wedge p_c) \geq \underline{\phi}. \tag{15}$$

(v) The sparsity index, s_c, and penalty λ_c satisfy

$$\frac{s_c}{\sqrt{n}} \to 0, \qquad \sqrt{s_c}\lambda_c \to 0. \tag{16}$$

Versions of each condition in Assumption 2 appear in Belloni et al. (2012), who also use them to establish convergence of first-stage LASSO estimates. Assumption 2(i) is a normalization and can always be satisfied by rescaling the columns of the dictionary. Assumption 2(ii) requires that the approximately sparse representation in (11) be accurate in the sense that the approximation error is $O(\sqrt{s_c/n})$ in prediction norm. The lower bound in Assumption 2(iii) ensures that the penalty dominates the empirical score with high probability, yielding the usual LASSO prediction-error bounds. Conditions of this form motivate the penalty choices in Bickel et al. (2009), van de Geer and Bühlmann (2009), Belloni et al. (2012), and Chetverikov and Sørensen (2025). Assumption 2(iv) imposes the restricted and sparse eigenvalue conditions discussed above. Finally, part (v) imposes two rate restrictions. Together with part (ii), $\sqrt{s_c}\lambda_c \to 0$ yields first-stage prediction consistency, while $s_c/\sqrt{n} \to 0$ controls terms that arise in the higher-order expansion.

Combined with Assumption 1, these conditions ensure that (4) holds for every candidate, so the resulting IV estimators are first-order equivalent. Beyond guaranteeing suitable first-stage convergence of the candidate first-stage estimators, these conditions do little to restrict the candidate list. Penalties may vary between the bounds in parts (iii) and (v), and dictionaries may differ in both content and dimension. The criterion developed in Section 3.2 ranks these first-order-equivalent candidates.

The feasible version of the criterion, developed in Section 3.3, uses the observed effective dimension d_c in place of its unknown moments. The following condition justifies this substitution.

Assumption 3 (Partial Beta-Min)**.** *For each candidate, let*

$$\underline{b}_{n,c} := \frac{4}{\sqrt{\underline{\phi}}} \left\{ \frac{(1 + c_\lambda^{-1})\lambda_c \sqrt{s_c}}{\kappa_0} + C_a \sqrt{\frac{s_c}{n}} \right\},$$

where κ_0 and $\underline{\phi}$ are as in Assumption 2(iv). Define the set of strong coordinates by $A_c := \{j \in T_c : |\pi^0_{c,j}| \geq \underline{b}_{n,c}\}$ with $k_c := |A_c|$. There is a fixed constant $C_s \geq 1$ such that, uniformly over $c \in [J]$,

$$s_c \leq C_s k_c, \qquad k_c \to \infty, \qquad \frac{\log(p_c/k_c)}{k_c} \to 0. \tag{17}$$

Assumption 3 imposes a beta-min condition on a subset of the non-zero coefficients. In the literature on consistent model selection via the LASSO, beta-min conditions are imposed on the entire coefficient support T_c and, together with conditions on the design, imply recovery of the true support (Zhao and Yu, 2006; Wainwright, 2009). The condition here plays a different role. The strong coordinates are selected with probability approaching one. Together with the concentration bound for d_c, this yields ratio consistency of the observed effective dimension for its expectation; see Proposition 1. It does not impose a beta-min lower bound on the coefficients in $T_c \setminus A_c$, nor does it require these coefficients to be selected with probability approaching one. The requirement that $k_c \to \infty$ parallels the condition in Donald and Newey (2001) that the number of instruments used to estimate the first stage grows with the sample size, and plays a similar role in the technical analysis. Because $\underline{b}_{n,c} \to 0$, Assumption 3 allows the coefficients in the strong core to shrink with n, even though the size of that core must diverge.

3.2 The AMSE Criterion

This subsection formally establishes the decomposition in (8) and presents the infeasible criterion, a closed-form asymptotic approximation to the expectation of the candidate-specific term. Exact expressions for the terms in (8) are given in the appendix but are not needed for the characterization below.

Two population quantities enter the infeasible criterion. The first is a measure of first-stage strength,

$$h_c := \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}\Pi_i], \tag{18}$$

the population cross-moment between the fitted values and the optimal instrument. This is the population counterpart of the denominator of the IV estimator in (6). Under the conditions of Section 3.1, h_c converges to the second moment of the optimal instrument, H, and so is bounded away from zero in large samples.

The second quantity measures how well the fitted instrument approximates the optimal instrument. Because the IV estimator is unchanged when its instrument is multiplied by a nonzero constant, a fitted instrument proportional to the optimal instrument should register no error. We therefore measure the first-stage approximation error against the best rescaling of the optimal instrument,

$$\mathcal{A}_c := \min_{a \in \mathbb{R}} \overline{\mathbb{E}}\Big[(\widehat{\Pi}_{c,i} - a\Pi_i)^2\Big] = \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - \frac{h_c^2}{H}. \tag{19}$$

The minimized value is the mean squared residual from a population regression of the fitted instrument on the optimal instrument. It reflects both the error from approximating the optimal instrument with the dictionary and the shrinkage and noise introduced by LASSO estimation. The ratio $\mathcal{A}_c/h_c^2$ appearing in the infeasible criterion below is invariant to rescaling of the fitted instrument.

The closed-form approximation to the expectation of the candidate-specific term combines these two quantities with certain error moments and moments of the effective dimension of the LASSO fit defined in Section 2.2,

$$S_c := \frac{\sigma_\varepsilon^2 \mathcal{A}_c}{h_c^2} + \frac{\sigma_{\varepsilon v}^2\, \mathbb{E}[d_c^2 + d_c]}{n h_c^2}. \tag{20}$$

We refer to S_c as the infeasible criterion. Each of its components is a population quantity, so the infeasible criterion cannot be computed directly from the data. Section 3.3 constructs a consistent feasible counterpart used to select a candidate.

Theorem 1 (Infeasible AMSE Criterion)**.** *Suppose Assumptions 1, 2, and 3 hold, with $J < \infty$ fixed. Then, for all sufficiently large n, there exist random variables C_0, $\{Q_c\}_{c\in[J]}$, and $\{r_c\}_{c\in[J]}$ such that*

(i) For every candidate,

$$n(\widehat{\beta}_c - \beta)^2 = C_0 + Q_c + r_c, \tag{21}$$

where C_0 does not depend on the candidate, $\mathbb{E}[C_0] = \sigma_\varepsilon^2/H + O(n^{-1})$, and the remainder satisfies $\max_{c\in[J]} |r_c|/\mathbb{E}[Q_c] \longrightarrow_p 0$.

(ii) Uniformly over candidates, $\mathbb{E}[Q_c] > 0$ *and*

$$\mathbb{E}[Q_c] = S_c \left(1 + o(1)\right). \tag{22}$$

Theorem 1(i) establishes the validity of the decomposition in (8), while Theorem 1(ii) shows that S_c approximates the candidate-specific AMSE component $\bar{Q}_c = \mathbb{E}[Q_c]$ with vanishing relative error. Together, these results justify ranking candidates by S_c: minimizing S_c is asymptotically equivalent to minimizing $\bar{Q}_c$, and hence AMSE, over the candidate list.

The infeasible criterion consists of two terms. The first is proportional to $\mathcal{A}_c$, which measures how well the candidate LASSO approximates the optimal instrument. This term vanishes only for a candidate estimator that is always proportional to the true first-stage Π. The second term in S_c is proportional to $\mathbb{E}[d_c^2 + d_c]$, a measure of the complexity of the LASSO fitted model. A richer dictionary or a lighter penalty may reduce $\mathcal{A}_c$ while increasing the effective dimension. To examine this tradeoff, write $\rho = \mathrm{Corr}(\varepsilon_i, v_i)$, so that $\sigma_{\varepsilon v} = \rho\sigma_\varepsilon\sigma_v$, and rewrite (20) as

$$S_c = \frac{\sigma_\varepsilon^2}{h_c^2}\left\{\mathcal{A}_c + \rho^2\sigma_v^2\,\frac{\mathbb{E}[d_c^2 + d_c]}{n}\right\}. \tag{23}$$

In this form the second term is weighted by $\rho^2\sigma_v^2$, which is proportional to the squared covariance between the structural and first-stage errors. When the regressor is close to exogenous, the bias from even a complex fitted model is small and candidates are ranked mainly by the approximation term.[5] When the regressor is highly endogenous, the same complexity produces a larger bias and the criterion favors heavier regularization. This is the LASSO analogue of the many-instrument bias examined in Bekker (1994) and Donald and Newey (2001).

Cross-validation on the first stage can be thought of as selecting the candidate that minimizes $\mathcal{A}_c$ while ignoring the many-instrument bias.[6] When endogeneity is weak, the complexity term receives little weight, so the cross-validation criterion and the AMSE criterion may rank candidates similarly. As endogeneity strengthens, the AMSE criterion places greater weight on first-stage complexity, whereas cross-validation does not. In the simulations of Section 4.1, the effective dimension selected by the proposed rule falls sharply as the error correlation rises, while the cross-validation choice is unchanged because its objective uses only the first stage. The plug-in penalty of Belloni et al. (2012) is chosen to dominate the first-stage noise with high probability so that Assumption 2(iii) is satisfied, but its construction does not take into account the measure of fit $\mathcal{A}_c$ nor the level of endogeneity.

3.3 Feasible Candidate Selection

Although S_c is infeasible, we construct an observable criterion that estimates it up to a candidate-independent centering term and therefore preserves its ranking asymptotically. The only additional inputs not obtained from the candidate fits are the structural error moments σ_ε^2 and $\sigma_{\varepsilon v}$. We estimate these moments using residuals from a pilot candidate $\check{c} \in [J]$.

[5] Exact exogeneity is ruled out, since Assumption 1(ii) requires $\sigma_{\varepsilon v} \neq 0$. When $\sigma_{\varepsilon v} = 0$ the second term vanishes altogether and the remaining terms require a different expansion.

[6] Cross-validation targets a first stage estimator $\widehat{\Pi}$ that minimizes $\bar{\mathbb{E}}[(\widehat{\Pi}_i - \Pi_i)^2]$, which is an upper bound on $\mathcal{A}_c$.

For each candidate, the observed IV denominator

$$\widehat{h}_c := \mathbb{E}_n[\widehat{\Pi}_{c,i} x_i] \tag{24}$$

will serve as an estimator for first-stage cross-moment h_c. The KKT conditions for the LASSO optimization problem in (5) yield $\widehat{h}_c = \mathbb{E}_n[\widehat{\Pi}_{c,i}^2] + \lambda_c \|\widehat{\pi}_c\|_1 \geq 0$, so $\widehat{h}_c = 0$ if and only if $\widehat{\pi}_c = 0$. The structural error moments are estimated with residuals from the pilot candidate $\check{c}$. Setting $\check{\beta} := \widehat{\beta}_{\check{c}}$, let

$$\widehat{\sigma}_\varepsilon^2 := \mathbb{E}_n[(y_i - \check{\beta} x_i)^2] \qquad \text{and} \qquad \widehat{\sigma}_{\varepsilon v} := \mathbb{E}_n[(y_i - \check{\beta} x_i) x_i]. \tag{25}$$

We put together these estimates to construct the feasible criterion

$$\widehat{S}_c := \frac{\widehat{\sigma}_\varepsilon^2 \mathbb{E}_n[\widehat{\Pi}_{c,i}^2] + n^{-1} \widehat{\sigma}_{\varepsilon v}^2 (d_c^2 + d_c)}{\widehat{h}_c^2}, \tag{26}$$

with the convention $\widehat{S}_c := +\infty$ when $\widehat{h}_c = 0$. The selected candidate minimizes the feasible criterion,

$$\widehat{c} := \min \left(\arg\min_{c \in [J]} \widehat{S}_c \right), \tag{27}$$

with ties broken by the smallest index.

The feasible criterion $\widehat{S}_c$ replaces the moment $\mathbb{E}[d_c^2 + d_c]$ in the infeasible S_c with the realized value $d_c^2 + d_c$. This replacement is justified by the following proposition.

Proposition 1 (Ratio Consistency of the Effective Dimension)**.** *Suppose Assumptions 1, 2, and 3 hold, with $J < \infty$ fixed. Uniformly over candidates,*

$$\frac{d_c}{\mathbb{E}[d_c]} \longrightarrow_p 1, \tag{28}$$

$$\frac{d_c^2 + d_c}{\mathbb{E}[d_c^2 + d_c]} \longrightarrow_p 1. \tag{29}$$

The result in Proposition 1 relies on the condition from Assumption 3 that the number of nonzero coefficients in π_c^0 diverges. Under this condition, second-order Stein identities from Bellec and Zhang (2021) show that the variance of the effective dimension grows more slowly than its squared mean, yielding (28) and (29). Proposition 1 does not require the LASSO estimator to perfectly recover the support of the underlying coefficient vector π_c^0 in (11).

The numerator of $\widehat{S}_c$ contains the fitted second moment $\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]$ rather than an estimate of the approximation error $\mathcal{A}_c$. By the decomposition in (19) applied to sample moments, this substitution adds the constant $\widehat{\sigma}_\varepsilon^2/H$ relative to a criterion that directly uses an empirical analogue of $\mathcal{A}_c$. This added constant does not depend on the candidate and thus does not affect the minimizer of the feasible criterion.[7] The feasible criterion also uses the denominator $\widehat{h}_c^2$ instead of the true h_c^2. To first-order, the error from using this estimate of the denominator is $-2\widehat{\sigma}_\varepsilon^2 \mathbb{E}_n[\Pi_i v_i]/H^2$, which also does not depend on the candidate. Let B_n collect these two

[7]The added constant estimates σ_ε^2/H, which Theorem 1(i) shows is the leading expectation of the common term in (8). The feasible criterion therefore estimates the AMSE of each candidate rather than the candidate-specific term alone.

candidate-independent terms,

$$B_n := \widehat{\sigma}_\varepsilon^2 \left(\frac{1}{H} - \frac{2\mathbb{E}_n[\Pi_i v_i]}{H^2} \right). \tag{30}$$

Theorem 2 below shows that, uniformly over candidates,

$$\widehat{S}_c = B_n + \bar{Q}_c \left(1 + o_p(1)\right). \tag{31}$$

Minimizing $\widehat{S}_c$ is then asymptotically equivalent to minimizing $\bar{Q}_c$. Note that B_n itself need not be observed and may change with the sample size.

Theorem 2 (Feasible AMSE Selection)**.** *Suppose Assumptions 1, 2, and 3 hold, with $J < \infty$ fixed.*

(i) Up to the common centering B_n, the feasible criterion consistently estimates the infeasible criterion,

$$\max_{c\in[J]} \frac{|\widehat{S}_c - B_n - S_c|}{S_c} \longrightarrow_p 0, \tag{32}$$

and thus the expectation of the candidate-specific term,

$$\max_{c\in[J]} \frac{|\widehat{S}_c - B_n - \bar{Q}_c|}{\bar{Q}_c} \longrightarrow_p 0. \tag{33}$$

(ii) The selected candidate $\widehat{c}$ satisfies

$$\frac{\bar{Q}_{\widehat{c}}}{\min_{c\in[J]} \bar{Q}_c} \longrightarrow_p 1. \tag{34}$$

By Theorem 2(ii), the selected candidate attains the smallest candidate-specific AMSE component in the candidate list up to vanishing relative error.

The conditional Gaussianity and homoskedasticity restrictions in Assumption 1 are used to derive the AMSE ranking. If the errors are non-Gaussian or heteroskedastic, Theorem 2 no longer guarantees that the selected candidate is AMSE-optimal. If, under suitable alternative conditions, all candidates continue to admit the common first-order expansion in (4) uniformly over the fixed candidate list, then the selected estimator retains the corresponding asymptotic distribution and first-order inference remains valid.

4 Simulation Study

In this section, we examine the finite-sample performance of the proposed criterion and compare to first-stage selection via cross-validation and the plug-in penalty of Belloni et al. (2012). The simulations are calibrated to the data of Gilchrist and Sands (2016), who study the effect of social spillovers on movie consumption. The authors originally instrument for ticket sales on a given opening weekend day with 48 linearly independent instruments that measure national weather conditions. We hold these initial instrumnts as fixed and consider two main error regimes. We first consider Gaussian errors, which matches the setting of Section 3, and then repeat the exercise with heavier-tailed Laplace errors to asses the sensitivity of our proposed criterion based selection to non-Gaussian errors. Appendix H provides implementation details and complete results.

4.1 Calibrated Gaussian Designs

We hold the observed weather instruments fixed and generate the outcome and endogenous variable from the IV model described in (1) with $\beta = 1$. The true first stage Π is proportional to the fitted values from a least-squares regression of opening-weekend sales on the original 48 linearly independent instruments. We rescale these fitted values to vary identification strength, targeting expected first-stage F statistics of 3.8, 6.6, 12, or 25 at the original sample size of $n = 1{,}671$. We also evaluate the performance of our proposed criterion on a smaller, fixed subsample of $n = 800$. Holding the instruments and first-stage direction fixed lets us examine how each rule's performance changes with identification strength and endogeneity in an empirically plausible setting.

The errors are generated homoskedastic Gaussian, with $\varepsilon_i = \rho v_i + \sqrt{1-\rho^2}\, e_i$ for $v_i, e_i \overset{\text{iid}}{\sim} N(0,1)$. The parameter ρ controls the degree of endogeneity. We consider a value calibrated to the data, $\rho \approx -0.29$, along with alternate values $\rho = -0.40, -0.50, \ldots, -0.90$ to examine how the performance of various criterion changes with the level of endogeneity. Each of the 56 designs is simulated 1,000 times. We report the average of $n(\widehat{\beta} - \beta)^2$, which we refer to as either the risk or the mean squared error.

Each candidate first stage implementation uses one of three nested dictionaries: either 34 temperature instruments (Temperature), the original 48 weather instruments (Original), or 524 instruments including temperature–weather interactions (Expanded). For each dictionary, we consider 13 multiples of the plug-in penalty of Belloni et al. (2012) ("BCCH"), giving 39 candidates in total. This plug-in penalty scale uses the known first-stage error variance in place of an estimate. The proposed criterion selects among these candidate dictionary-penalty pairs by minimizing $\widehat{S}_c$ in (26), with estimated error moments coming from a pilot estimator of β. Cross-validation chooses from the same candidates using ten-fold first-stage prediction error, keeping observations from the same opening weekend in the same fold. BCCH uses its plug-in penalty on the Original dictionary. When it selects no instruments, we use the pilot estimate instead.

Figure 2 shows how the rules respond to endogeneity at the two intermediate levels of identification strength, $F = 6.6$ and $F = 12$. Cross-validation performs well when the regressor is only mildly endogenous, but its relative risk increases as endogeneity rises. In contrast, the BCCH rules seems to impose a high degree of regularization and, as such, performs poorly when endogeneity is low but similarly to the criterion based rule at the highest levels of endogeneity. Both rules do not account for the level of endogeneity and so their mean effective dimension remains the same across all values of $|\rho|$.

By comparasion, the proposed criterion adjusts the first-stage complexity based on the level of endogeneity, consistent with the increasing complexity cost seen in (23). At both $F = 6.6$ and $F = 12$, the mean criterion effective dimension falls by nearly four-fold as $|\rho|$ ranges from 0.29 to 0.90. As a result of this adjustment, the criterion is able to achieve nearly optimal performance across the entire range of regimes considered in Figure 2. The improvements in relative risk from using this criterion can be sizeable. At low levels of endogeneity, the criterion's risk is about 40% lower than that of the BCCH plug-in rule while at the highest level of endogeneity, the criterion's risk is nearly 35% lower than that of cross-validation.

Table 1 summarizes performance across all designs. The "Known Moments" column uses the proposed criterion with the error moments set to their population values, $\sigma_\varepsilon^2 = 1$ and $\sigma_{\varepsilon v} = \rho$. The "Infeasible Criterion" column selects the candidate minimizing the infeasible criterion S_c, with its population moments computed from 1,000 independent simulation replications. Compared to the Cross-Validation and BCCH selection procedures, the feasible criterion has the smallest average and maximum relative risk at both

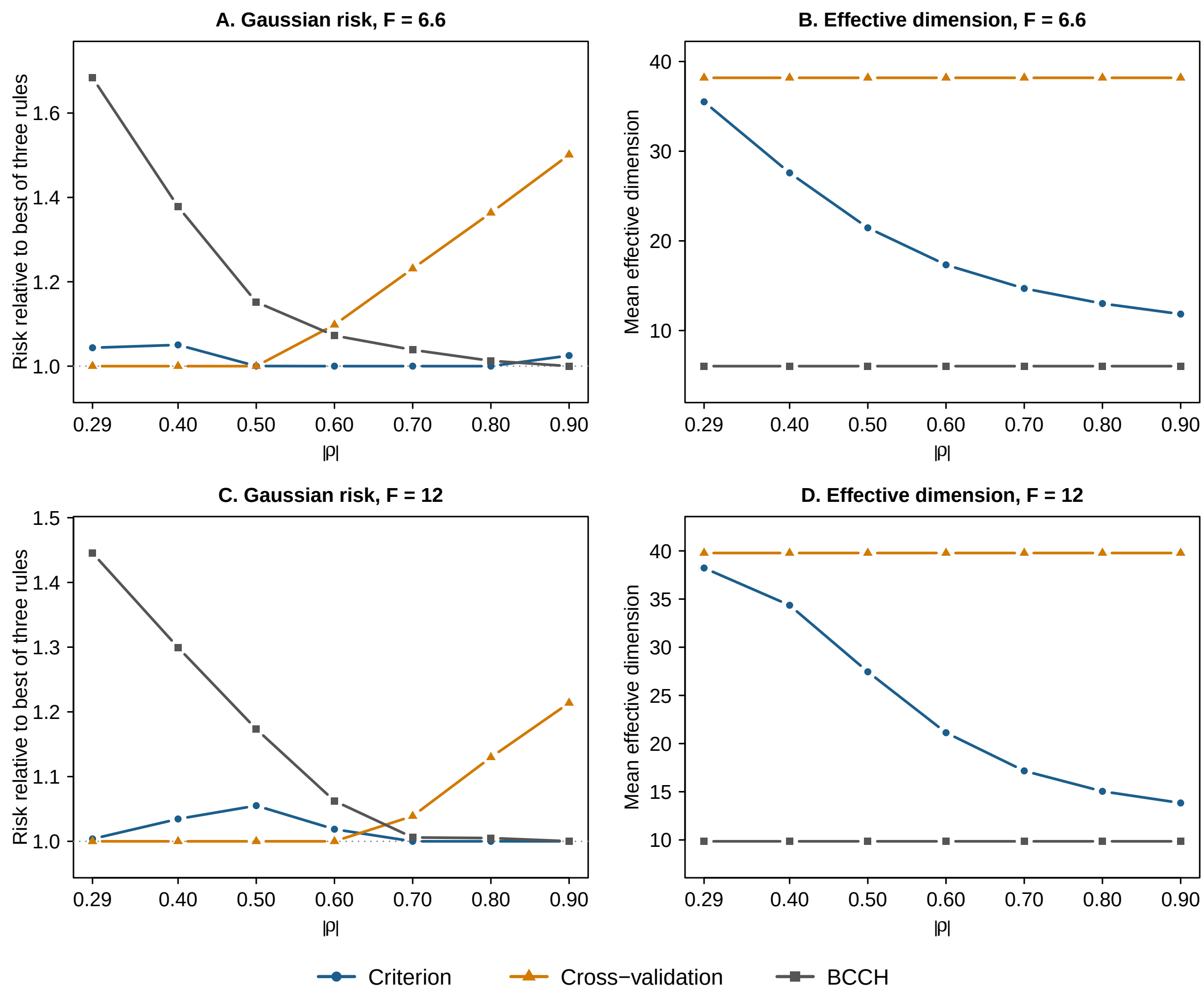

Figure 2: Relative risk and effective dimension under Gaussian errors.

Notes: Results use n = 1,671 and 1,000 replications. Panels A and C report risk relative to the lowest risk among the three rules; Panels B and D report mean effective dimension. Panels A–B use $F = 6.6$ and Panels C–D use $F = 12$. The horizontal axis is the absolute error correlation $|\rho|$.

sample sizes, over essentially all identification strength regimes, and over the entire grid of considered DGPs. At the original sample size, its largest relative risk is about 13% above that of the best rule, compared with 55% for cross-validation and 105% for BCCH. As identification strengthens, the criterion and cross-validation perform more similarly. At n = 1,671 and $F = 25$, cross-validation has slightly lower average relative risk. This is consistent with first-stage estimation error having less effect on the structural estimate as identification strengthens. Intuitively, the behavior of the structural parameter estimate is driven more by the first-stage "signal" than first-stage estimation noise under stronger indentification.

Panel C also shows the effect of estimating the error moments used in the criterion. Using their population values reduces average risk relative to the infeasible criterion from 1.164 to 1.116, removing roughly 30% of the average excess risk and suggesting that the cost of estimating the error-variances is somewhat substantial at the Gilchrist and Sands (2016) sample size. Even with known error moments, the feasible criterion still uses estimated first-stage quantities to rank candidates, and its risk remains above that of the infeasible

Table 1: Relative risk, Gaussian errors.

Design	Summary Measure	Feasible Criterion	Known Moments	Infeasible Criterion	Cross-validation	BCCH
Panel A: All F and ρ, by sample size						
$n = 800$	Average	1.045	–	–	1.135	1.189
	Maximum	1.266	–	–	1.665	1.833
$n = 1{,}671$	Average	1.026	–	–	1.106	1.194
	Maximum	1.128	–	–	1.551	2.045
Panel B: $n = 1{,}671$, by first-stage strength						
$F = 3.8$	Average	1.040	–	–	1.174	1.297
	Maximum	1.128	–	–	1.551	2.045
$F = 6.6$	Average	1.017	–	–	1.171	1.191
	Maximum	1.050	–	–	1.501	1.684
$F = 12$	Average	1.016	–	–	1.055	1.142
	Maximum	1.055	–	–	1.214	1.445
$F = 25$	Average	1.030	–	–	1.023	1.146
	Maximum	1.072	–	–	1.117	1.379
Panel C: All cells, relative to the oracle						
Full grid	Average	1.164	1.116	1.000	1.268	1.326
	Maximum	1.440	1.289	1.000	1.811	2.010

Notes: Risk is the mean of $n(\widehat{\beta}-\beta)^2$ over 1,000 replications. Panels A and B present risk relative to the best of the three feasible rules in each design. Panel A averages or maximizes over all 28 designs at each n. Panel B averages over the seven values of ρ at each F at $n = 1671$. Panel C reports risk relative to the candidate minimizing the infeasible S_c. Known moments evaluates the feasible criterion at population error moments. Details of these infeasible benchmarks are given in Appendix H.

criterion in this comparison. The

4.2 Robustness to Non-Gaussian Errors

Since the theory in Section 3 is derived under Gaussian errors, a reasonable question might be whether the proposed criterion is still useful in models with non-Gaussian errors. In other words, is the characterization of IV-LASSO AMSE in the Gaussian model informative about the AMSE of IV-LASSO in more general models. To assess this, we repeat the simulations of Section 4.1 with variance-one Laplace errors. We retain the same first stages and candidate list and rerun the simulation experiment by drawing first-stage and structural errors from the heavier-tailed Laplace distribution.

Figure 3 shows a similar response to endogeneity as under Gaussian errors. At $n = 1{,}671$ and $F = 6.6$, cross-validation a slightly lower risk than the proposed criterion at the three smallest values of $|\rho|$, while the proposed criterion has the lowest risk from $|\rho| = 0.60$ through 0.90. At the highest level of endogeneity, the criterion's risk is about 28% below that of cross-validation. The degrees of freedom of the criterion selected first-stage falls from 35.6 to 11.7, closely matching the decline under Gaussian errors. The similarity in selected dimensions suggests that use of the non-Gaussian errors has little effect on how heavily first-stage complexity is punished.

This similarity between results in the Gaussian and non-Gaussian regimes is echoed in Table 2 which shows that the feasible criterion again has the smallest average and maximum relative risk at both sample sizes.

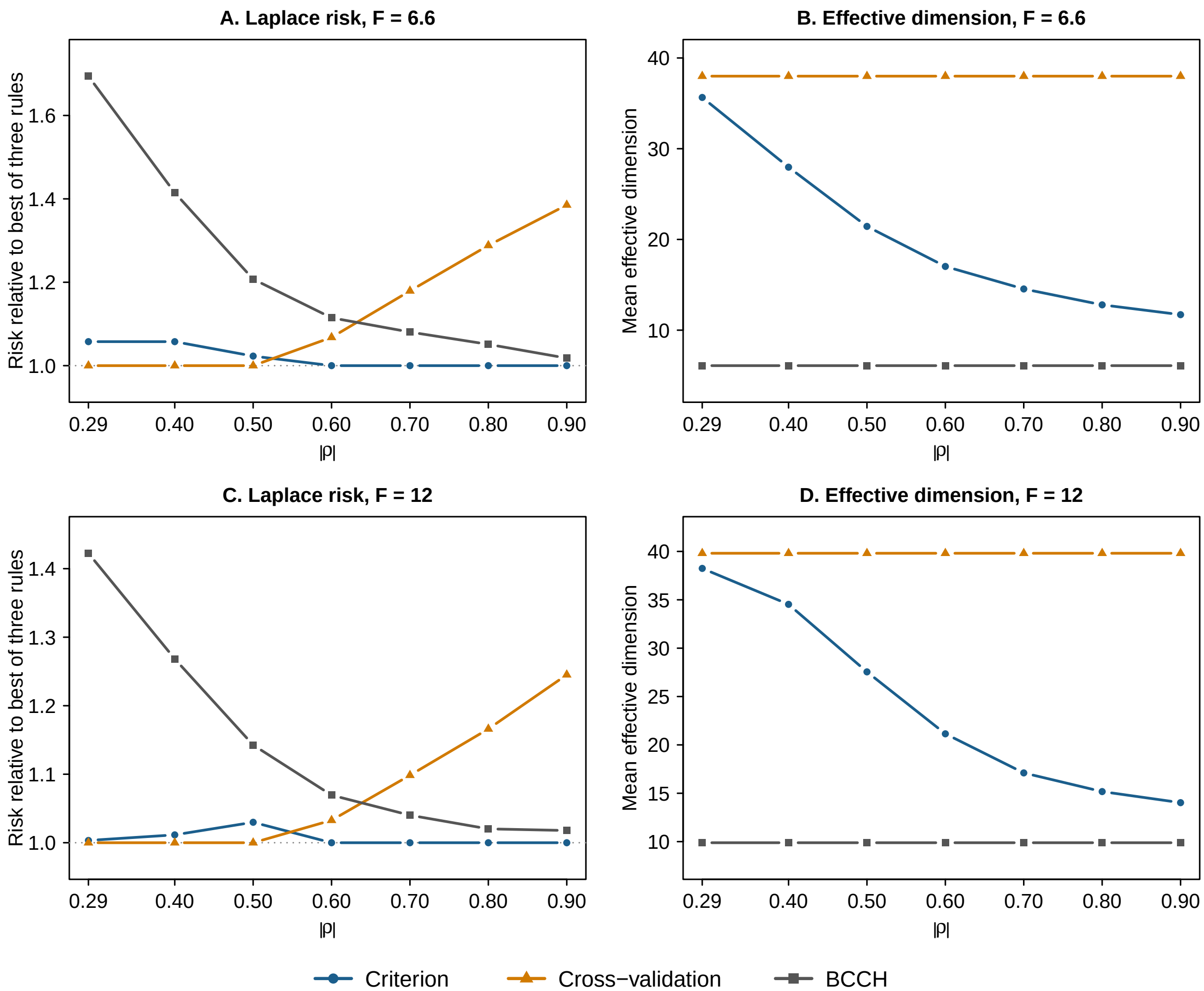

Figure 3: Relative risk and effective dimension under Laplace errors.

Notes: Results use $n = 1{,}671$ and 1,000 replications. Panels A and C report risk relative to the lowest risk among the three rules; Panels B and D report mean effective dimension. Panels A–B use $F = 6.6$ and Panels C–D use $F = 12$. The horizontal axis is the absolute error correlation $|\rho|$.

At $n = 1{,}671$, its risk is on average 2.9% above the lowest risk in each design, compared with 2.6% under Gaussian errors. The relative performance of the competing rules does change in some designs. At $F = 3.8$ and the original sample size, BCCH has lower average relative risk than cross-validation under Laplace errors, while the reverse holds under Gaussian errors. The proposed criterion has the lowest average relative risk in both cases. At the strongest level of identification, cross-validation again performs slightly better on average.

The comparison with the infeasible criterion is also similar across error distributions. The criterion's average risk relative to the infeasible criterion is 1.162 under Laplace errors, compared with 1.164 under Gaussian errors. As in the Gaussian design, using the population error moments in place of their estimates removes roughly 30% of average excess risk . Together, these results suggest that the AMSE of the IV-LASSO estimator in a more general model may be well approximated by its AMSE in the Gaussian model, in which case the ranking of candidates derived in Section 3 remains informative.

Table 2: Relative risk, Laplace errors.

Design	Summary Measure	Feasible Criterion	Known Moments	Infeasible Criterion	Cross-validation	BCCH
Panel A: All F and ρ, by sample size						
$n = 800$	Average	1.044	–	–	1.130	1.205
	Maximum	1.261	–	–	1.631	2.010
$n = 1{,}671$	Average	1.029	–	–	1.111	1.179
	Maximum	1.134	–	–	1.627	1.694
Panel B: $n = 1{,}671$, by first-stage strength						
$F = 3.8$	Average	1.061	–	–	1.224	1.160
	Maximum	1.134	–	–	1.627	1.683
$F = 6.6$	Average	1.020	–	–	1.132	1.226
	Maximum	1.058	–	–	1.385	1.694
$F = 12$	Average	1.006	–	–	1.078	1.140
	Maximum	1.030	–	–	1.245	1.423
$F = 25$	Average	1.028	–	–	1.011	1.191
	Maximum	1.049	–	–	1.079	1.435
Panel C: All cells, relative to the oracle						
Full grid	Average	1.162	1.115	1.000	1.263	1.325
	Maximum	1.432	1.288	1.000	1.784	1.860

Notes: Results are computed as in Table 1, using 1,000 replications under Laplace errors. Risk is the mean of $n(\widehat{\beta}-\beta)^2$ over 1,000 replications. Panels A and B present risk relative to the best of the three feasible rules in each design. Panel A averages or maximizes over all 28 designs at each n. Panel B averages over the seven values of ρ at each F at $n = 1671$. Panel C reports risk relative to the candidate minimizing the infeasible S_c. Known moments evaluates the feasible criterion at population error moments. Details of these infeasible benchmarks are given in Appendix H.

References

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*(3), 657–681.

Bellec, P. C. and C.-H. Zhang (2021). Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *The Annals of Statistics 49*(4), 1864–1903.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

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

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bogachev, V. I. (1998). *Gaussian Measures*, Volume 62 of *Mathematical Surveys and Monographs*. Providence, RI: American Mathematical Society.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics 170*(2), 383–398.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated Lasso in high dimensions. *The Annals of Statistics 49*(3), 1300–1317.

Chetverikov, D. and J. R.-V. Sørensen (2025). Selecting penalty parameters of high-dimensional M-estimators using bootstrapping after cross validation. *Journal of Political Economy 133*(10), 3208–3248.

Donald, S. G. and W. K. Newey (2001). Choosing the number of instruments. *Econometrica 69*(5), 1161–1191.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Gilchrist, D. S. and E. G. Sands (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy 124*(5), 1339–1382.

Hahn, J., J. Hausman, and G. Kuersteiner (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal 7*(1), 272–306.

Heinonen, J. (2005). Lectures on Lipschitz analysis. Report 100, University of Jyväskylä, Jyväskylä, Finland.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika 12*(1–2), 134–139.

Kuersteiner, G. and R. Okui (2010). Constructing optimal instruments by first-stage prediction averaging. *Econometrica 78*(2), 697–718.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics 28*(5), 1302–1338.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica 27*(4), 575–595.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*(4), 809–837.

Newey, W. K. and R. J. Smith (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics 165*(1), 70–86.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume 2, Chapter 15, pp. 881–935. Elsevier.

Sargan, J. D. (1976). Econometric estimators and the Edgeworth approximation. *Econometrica 44*(3), 421–448.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B 58*(1), 267–288.

Tibshirani, R. J. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

Tibshirani, R. J. and J. Taylor (2012). Degrees of freedom in Lasso problems. *The Annals of Statistics 40*(2), 1198–1232.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Velez, A. (2026). Debiased machine learning with many cross-fitting folds. Working paper, Cornell University. August 2, 2026.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research 7*, 2541–2563.

A Notation and Preliminary Reductions

Here we collect the notation used throughout the appendix and construct the terms C_0, Q_c, and r_c of the decomposition (8).

LASSO Notation. Lemma E.1, below, recalls certain facts about the LASSO fit from Tibshirani and Taylor (2012). In particular, it notes that the LASSO fitted-value map $\mu_c : \mathbb{R}^n \to \mathbb{R}^n$ is one-Lipschitz, and hence Lebesgue-almost-everywhere differentiable by Rademacher's theorem. At a differentiability point $w \in \mathbb{R}^n$, write

$$\dot{\mu}_c(w) := D\mu_c(w), \qquad \operatorname{div}\mu_c(w) := \operatorname{tr}\{\dot{\mu}_c(w)\}.$$

Under Assumption 1(i) the vector $\Pi + v$ has a Lebesgue density, so (E.5) holds almost surely at $w = \Pi + v$. Almost surely, then, $\dot{\mu}_c(\Pi + v)$ is the orthogonal projection onto $\operatorname{col}(Z_{c,E_c(x)})$ and hence

$$\dot{\mu}_c(\Pi+v) = \dot{\mu}_c(\Pi+v)' = \dot{\mu}_c(\Pi+v)^2, \qquad \operatorname{div}\mu_c(\Pi+v) = d_c, \qquad \dot{\mu}_c(\Pi+v)\widehat{\Pi}_c = \widehat{\Pi}_c, \tag{A.1}$$

where the first property follows from symmetry and idempotence of projection onto $\operatorname{col}(Z_{c,E_c(x)})$, the second property follows because the trace of a projection matrix is equal to its rank, and the third property comes from the fact that $\widehat{\Pi}_c$ is supported on $E_c(x)$. In addition, by nature of being projections, both $\dot{\mu}_c(\Pi+v)$ and $I - \dot{\mu}_c(\Pi+v)$ are contractions.

With d_c defined in (10), we will save notation later in the proofs and write

$$\bar{d}_c := \mathbb{E}[d_c], \qquad R_{d,c} := \mathbb{E}[d_c^2 + d_c] = \bar{d}_c^2 + \operatorname{Var}(d_c) + \bar{d}_c. \tag{A.2}$$

For each candidate, define

$$\begin{gathered} \widetilde{\Pi}_c(a) := \mu_c(\Pi + a) - \frac{h_c}{H}\Pi, \qquad a \in \mathbb{R}^n, \\ \widetilde{\Pi}_c := \widetilde{\Pi}_c(v), \qquad \widetilde{\Pi}_{c,i} := \widehat{\Pi}_{c,i} - \frac{h_c}{H}\Pi_i. \end{gathered} \tag{A.3}$$

We refer to the vector $\tilde{\Pi}_c$ as the "centered" LASSO fit. By (18), $\mathbb{E}_n[\Pi_i^2] = H$, and (19),

$$\begin{gathered} \overline{\mathbb{E}}[\Pi_i\widetilde{\Pi}_{c,i}] = \overline{\mathbb{E}}[\Pi_i\widehat{\Pi}_{c,i}] - h_c = 0, \\ \overline{\mathbb{E}}[\widetilde{\Pi}_{c,i}^2] = \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - \frac{h_c^2}{H} = \mathcal{A}_c, \\ \mathbb{E}[\Pi'\widetilde{\Pi}_c] = 0, \qquad \mathbb{E}[\|\widetilde{\Pi}_c\|_2^2] = n\mathcal{A}_c. \end{gathered} \tag{A.4}$$

Using the fact that $\mathcal{A}_c$ is minimized at $a = \frac{h_c}{H}$, we can write

$$\mathcal{A}_c = \overline{\mathbb{E}}\left[\left(\widehat{\Pi}_{c,i} - \frac{h_c}{H}\Pi_i\right)^2\right]. \tag{A.5}$$

Gaussian error decomposition. Under Assumption 1(i) set $e := \varepsilon - \frac{\sigma_{\varepsilon v}}{\sigma_v^2}v$. Then e is Gaussian with mean zero and $\operatorname{Cov}(e_i, v_i) = 0$, hence independent of v:

$$\varepsilon = \frac{\sigma_{\varepsilon v}}{\sigma_v^2}v + e, \qquad e \perp v, \qquad \operatorname{Var}(e_i) = \sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2}. \tag{A.6}$$

This decomposition will be useful in characterizing moments of the IV-LASSO fit, particularly in deriving expressions for the correlation between the structural error ε and the LASSO fitted values $\widehat{\Pi}_c$. For example, consider the expectation

$$\mathbb{E}[\varepsilon'\widehat{\Pi}_c] = \mathbb{E}\left[\left(\frac{\sigma_{\varepsilon v}}{\sigma_v^2}v + e\right)'\widehat{\Pi}_c\right] = \frac{\sigma_{\varepsilon v}}{\sigma_v^2}\mathbb{E}[v'\widehat{\Pi}_c] + \mathbb{E}[e'\widehat{\Pi}_c].$$

Stein's identity for Gaussian random variables (Stein, 1981) yields that $\mathbb{E}[v'\widehat{\Pi}_c] = \sigma_v^2\bar{d}_c$ since, conditional on the instruments, $\widehat{\Pi}_c$ is an almost-everywhere differentiable function of v whose divergence is d_c by (A.1). The second term is zero since e is independent of v, and thus independent of $\widehat{\Pi}_c$, and has mean zero. Put together, we then have that $\mathbb{E}[\varepsilon'\widehat{\Pi}_c] = \sigma_{\epsilon v}\bar{d}_c$. This expression appears in the second term of the infeasible criterion S_c, defined at (19).

Positive Denominator Event. With $\widehat{h}_c$ defined by (24), Lemma G.2 shows that there is a fixed constant $\underline{h} > 0$ such that for $\mathcal{H}_n$ defined below, $\Pr(\mathcal{H}_n) \to 1$ as $n \to \infty$.

$$\delta_{h,c} := \widehat{h}_c - h_c, \qquad \mathcal{H}_n := \left\{\min_{c\in[J]} \widehat{h}_c \geq \frac{\underline{h}}{2}\right\}. \tag{A.7}$$

Since analyzing the properties of the IV-LASSO difficult is difficult when the denominator is close to zero, our proofs will often condition on the event $\mathcal{H}_c$ in order to show various stochastic bounds.

Squared-error decomposition. We now present the exact squared error decomposition that we use throughout the appendix. Under Assumptions 1 and 2, (G.14) of Lemma G.2 gives $\min_c h_c \geq \underline{h} > 0$ for all sufficiently large n. Without loss of generality, assume all n satisfy this condition and define

$$L_n := \frac{\sqrt{n}\,\mathbb{E}_n[\varepsilon_i\Pi_i]}{H}, \tag{A.8}$$

$$\Gamma_c := \frac{\sqrt{n}\,\mathbb{E}_n[\varepsilon_i\widehat{\Pi}_{c,i}]}{h_c}, \tag{A.9}$$

$$\Lambda_c := \Gamma_c + \left(1 - \frac{\widehat{h}_c}{h_c}\right)L_n, \tag{A.10}$$

$$q := \frac{\mathbb{E}_n[\Pi_i v_i]}{H}. \tag{A.11}$$

Here L_n is the first-order term in (4). For the infeasible estimator that uses the true optimal instrument Π, it can be shown that $\Lambda = (1-q)L_n$. Using the above, we can define the terms in (8):

$$C_0 := (1-q)^2L_n^2, \qquad Q_c := \Lambda_c^2 - C_0, \qquad r_c := n(\widehat{\beta}_c - \beta)^2 - \Lambda_c^2. \tag{A.12}$$

These definitions make (8) an identity, and since q and L_n are candidate-invariant, so is C_0. Write

$$C_{0n} := \frac{\sigma_\varepsilon^2}{H} + \frac{\sigma_\varepsilon^2\sigma_v^2 + 2\sigma_{\varepsilon v}^2}{nH^2}, \qquad \bar{Q}_c := \mathbb{E}[\Lambda_c^2] - C_{0n}. \tag{A.13}$$

Note that $Q_c \geq -C_0$ and, by Lemma F.4, $\mathbb{E}[C_0] = C_{0n} < \infty$. Hence $\mathbb{E}[Q_c]$ is well defined in $(-\infty, \infty]$ and equals $\mathbb{E}[\Lambda_c^2] - C_{0n} = \bar{Q}_c$, a quantity that Lemma C.3 shows is finite for all sufficiently large n.

B Proofs of the Main Results

B.1 A Linear Expansion

The proof of Theorem 1 relies on the following linear expansion of the IV-LASSO estimator.

Lemma B.1 (Linear Expansion). *Suppose Assumptions 1, 2, and 3 hold, with $J < \infty$ fixed. Then*

$$\max_{c\in[J]} \frac{|\sqrt{n}(\widehat{\beta}_c - \beta) - \Lambda_c|}{\bar{Q}_c} \to_p 0, \tag{B.1}$$

and

$$\max_{c\in[J]} |\Lambda_c| = O_p(1), \qquad \max_{c\in[J]} |\sqrt{n}(\widehat{\beta}_c - \beta)| = O_p(1). \tag{B.2}$$

Proof. Work on $\mathcal{H}_n$, whose probability tends to one by Lemma C.1. Lemma C.2, Proposition G.1, and Markov's inequality give

$$\Gamma_c = O_p(1), \qquad \Gamma_c - L_n = O_p(\sqrt{S_c}), \qquad L_n = O_p(1),$$

where the last bound also follows from $\mathbb{E}[L_n^2] = \sigma_\varepsilon^2/H$. On $\mathcal{H}_n$, $\widehat{h}_c \geq \underline{h}/2$ and

$$\sqrt{n}(\widehat{\beta}_c - \beta) = \Gamma_c \frac{h_c}{\widehat{h}_c}. \tag{B.3}$$

By Lemmas C.1, C.3, and G.2, $\underline{h} \leq h_c \leq C$, $\delta_{h,c} = O_p(n^{-1/2})$, $S_c/\bar{Q}_c \to 1$, and $\min_c n\bar{Q}_c \to \infty$. Hence

$$\begin{aligned} \frac{|\sqrt{n}(\widehat{\beta}_c - \beta) - L_n|}{\sqrt{\bar{Q}_c}} &\leq \frac{h_c}{\widehat{h}_c} \frac{|\Gamma_c - L_n|}{\sqrt{\bar{Q}_c}} + \frac{|\delta_{h,c}|}{\widehat{h}_c} \frac{|L_n|}{\sqrt{\bar{Q}_c}} \\ &= O_p(1) + O_p\{(n\bar{Q}_c)^{-1/2}\} = O_p(1). \end{aligned} \tag{B.4}$$

Since $\widehat{h}_c = h_c + \delta_{h,c}$, (B.3) gives $\Gamma_c = (1 + \delta_{h,c}/h_c)\sqrt{n}(\widehat{\beta}_c - \beta)$. Substituting this into $\Lambda_c = \Gamma_c - (\delta_{h,c}/h_c)L_n$ from (A.10) and rearranging yields

$$\sqrt{n}(\widehat{\beta}_c - \beta) - \Lambda_c = -\frac{\delta_{h,c}}{h_c}\{\sqrt{n}(\widehat{\beta}_c - \beta) - L_n\}. \tag{B.5}$$

Together with (B.4), $\delta_{h,c} = O_p(n^{-1/2})$, and $n\bar{Q}_c \to \infty$ this gives

$$\frac{|\sqrt{n}(\widehat{\beta}_c - \beta) - \Lambda_c|}{\bar{Q}_c} = O_p\{(n\bar{Q}_c)^{-1/2}\} = o_p(1).$$

The boundedness conclusions follow from (B.3) and $\Lambda_c = \Gamma_c - (\delta_{h,c}/h_c)L_n$. Since J is fixed and $\mathbb{P}(\mathcal{H}_n) \to 1$, these bounds prove (B.1) and (B.2). $\square$

B.2 Proof of Theorem 1

Proof. Step 1: Decomposition of $n(\widehat{\beta}_c - \beta)^2$.

By (G.14) of Lemma G.2, $\min_c h_c \geq \underline{h} > 0$ for all sufficiently large n, so Γ_c and Λ_c in (A.9) and (A.10) are well posed. For (21), the construction in (A.12) makes $n(\widehat{\beta}_c - \beta)^2 = C_0 + Q_c + r_c$ an identity with C_0 candidate invariant. Apply Lemma F.4 to see that

$$\mathbb{E}[C_0] = C_{0n} = \frac{\sigma_\varepsilon^2}{H} + O(n^{-1}),$$

while $\mathbb{E}[Q_c] = \bar{Q}_c$ by the definitions in (A.12) and (A.13). Lemma C.3 gives $\bar{Q}_c = S_c\{1 + o(1)\} > 0$ for all sufficiently large n, which is Theorem 1(ii). It remains to show that

$$\max_{c\in[J]} \frac{|r_c|}{\mathbb{E}[Q_c]} \to_p 0.$$

Step 2: Bound for $n(\widehat{\beta}_c - \beta)^2 - \Lambda_c^2$.

To this end, note that

$$\begin{aligned}\max_{c\in[J]} \frac{|n(\widehat{\beta}_c - \beta)^2 - \Lambda_c^2|}{\bar{Q}_c} &\leq \max_{c\in[J]} \frac{|\sqrt{n}(\widehat{\beta}_c - \beta) - \Lambda_c|}{\bar{Q}_c} \cdot \max_{c\in[J]} |\sqrt{n}(\widehat{\beta}_c - \beta) + \Lambda_c| \\ &\to_p 0,\end{aligned}$$

where the first factor tends to zero by (B.1), and the second is $O_p(1)$ by the triangle inequality and (B.2). Since $r_c = n(\widehat{\beta}_c - \beta)^2 - \Lambda_c^2$ by (A.12) and $\mathbb{E}[Q_c] = \bar{Q}_c$, the above display establishes $\max_{c\in[J]} |r_c|/\mathbb{E}[Q_c] \to_p 0$, completing Theorem 1(i). □

B.3 Proof of Proposition 1

Proof. The first conclusion (28) is part of (G.16) of Lemma G.2. For (29), write

$$\frac{d_c^2 + d_c}{R_{d,c}} = \frac{(d_c/\bar{d}_c)^2 + (d_c/\bar{d}_c)/\bar{d}_c}{R_{d,c}/\bar{d}_c^2}. \tag{B.6}$$

By (G.16), $d_c/\bar{d}_c \to_p 1$, $\bar{d}_c \to \infty$, and $R_{d,c}/\bar{d}_c^2 \to 1$. Since J is fixed and $R_{d,c} = \mathbb{E}[d_c^2 + d_c]$, (B.6) proves (29) uniformly over candidates. □

B.4 Proof of Theorem 2

Proof. By Lemma C.1, $\mathbb{P}(\mathcal{H}_n) \to 1$, and on $\mathcal{H}_n$ the definition in (A.7) gives $\min_{c\in[J]} \widehat{h}_c \geq \underline{h}/2 > 0$. Hence

$$\mathbb{P}\left(\min_{c\in[J]} \widehat{h}_c > 0\right) \to 1. \tag{B.7}$$

Step 1: The variance term.

Work on $\mathcal{H}_n$ through Step 2; since $\mathbb{P}(\mathcal{H}_n) \to 1$, the stochastic orders obtained there hold unconditionally. Since $\widehat{\Pi}_c = (h_c/H)\Pi + \widetilde{\Pi}_c$, $x = \Pi + v$, $\delta_{h,c} = \widehat{h}_c - h_c$, and $\mathbb{E}_n[\Pi_i^2] = H$, we can write

$$\mathbb{E}_n[\widehat{\Pi}_{c,i}^2] = \frac{h_c^2}{H} + \frac{2h_c}{H}\mathbb{E}_n[\Pi_i\widetilde{\Pi}_{c,i}] + \mathbb{E}_n[\widetilde{\Pi}_{c,i}^2],$$
$$\delta_{h,c} = h_c q + \mathbb{E}_n[\Pi_i\widetilde{\Pi}_{c,i}] + \mathbb{E}_n[\widetilde{\Pi}_{c,i}v_i].$$

Using the second equality to substitute for $\mathbb{E}_n[\Pi_i\widetilde{\Pi}_{c,i}]$ in the first gives

$$\begin{aligned}\mathbb{E}_n[\widehat{\Pi}_{c,i}^2] &= \frac{h_c^2 + 2h_c\delta_{h,c}}{H} - \frac{2h_c^2 q}{H} + \mathbb{E}_n[\widetilde{\Pi}_{c,i}^2] - \frac{2h_c}{H}\mathbb{E}_n[\widetilde{\Pi}_{c,i}v_i] \\ &= \frac{\widehat{h}_c^2}{H} - \frac{2h_c^2 q}{H} + \mathbb{E}_n[\widetilde{\Pi}_{c,i}^2] - \frac{2h_c}{H}\mathbb{E}_n[\widetilde{\Pi}_{c,i}v_i] - \frac{\delta_{h,c}^2}{H},\end{aligned} \tag{B.8}$$

where the second equality uses $\widehat{h}_c^2 = (h_c + \delta_{h,c})^2$. Next, write

$$\mathbb{E}_n[\widetilde{\Pi}_{c,i}^2] = \mathcal{A}_c + \big\{\mathbb{E}_n[\widetilde{\Pi}_{c,i}^2] - \mathcal{A}_c\big\}. \tag{B.9}$$

Factoring h_c^2 from $-2h_c^2 q/H + \mathcal{A}_c$ and substituting $h_c^2 = \widehat{h}_c^2 - (2h_c\delta_{h,c} + \delta_{h,c}^2)$, we obtain

$$\begin{aligned}-\frac{2h_c^2 q}{H} + \mathcal{A}_c &= \widehat{h}_c^2\left(-\frac{2q}{H} + \frac{\mathcal{A}_c}{h_c^2}\right) \\ &\quad + \left(\frac{2q}{H} - \frac{\mathcal{A}_c}{h_c^2}\right)(2h_c\delta_{h,c} + \delta_{h,c}^2).\end{aligned} \tag{B.10}$$

Substituting (B.9) and (B.10) into (B.8), dividing by $\widehat{h}_c^2$, and recalling $q = \mathbb{E}_n[\Pi_i v_i]/H$, gives

$$\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\widehat{h}_c^2} = \frac{1}{H} - \frac{2\mathbb{E}_n[\Pi_i v_i]}{H^2} + \frac{\mathcal{A}_c}{h_c^2} + R_{n,c}^{\text{fit}}, \tag{B.11}$$

where we define

$$\begin{aligned}R_{n,c}^{\text{fit}} := \frac{1}{\widehat{h}_c^2}\Big\{&\mathbb{E}_n[\widetilde{\Pi}_{c,i}^2] - \mathcal{A}_c - \frac{2h_c}{H}\mathbb{E}_n[\widetilde{\Pi}_{c,i}v_i] \\ &+ \left(\frac{2q}{H} - \frac{\mathcal{A}_c}{h_c^2}\right)(2h_c\delta_{h,c} + \delta_{h,c}^2) - \frac{\delta_{h,c}^2}{H}\Big\}.\end{aligned}$$

With $R_{n,c}^{\text{fit}} := 0$ on $\mathcal{H}_n^c$, Lemma D.2 gives $\max_{c\in[J]} |R_{n,c}^{\text{fit}}|/\bar{Q}_c \to_p 0$. Multiply (B.11) by $\widehat{\sigma}_\varepsilon^2$ and use the definition of B_n in (30) to obtain the exact identity

$$\frac{\widehat{\sigma}_\varepsilon^2\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\widehat{h}_c^2} = B_n + \frac{\sigma_\varepsilon^2\mathcal{A}_c}{h_c^2} + (\widehat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)\frac{\mathcal{A}_c}{h_c^2} + \widehat{\sigma}_\varepsilon^2 R_{n,c}^{\text{fit}}. \tag{B.12}$$

By (C.7) and (G.14), $\mathcal{A}_c/(h_c^2\bar{Q}_c) \le C$. Thus (D.5) and (D.9) give

$$\begin{aligned}\frac{\left|(\widehat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)\mathcal{A}_c/h_c^2 + \widehat{\sigma}_\varepsilon^2 R_{n,c}^{\text{fit}}\right|}{\bar{Q}_c} &\le C|\widehat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2| + |\widehat{\sigma}_\varepsilon^2|\frac{|R_{n,c}^{\text{fit}}|}{\bar{Q}_c} \\ &= o_p(1).\end{aligned}$$

Substituting this bound into (B.12) leaves

$$\frac{\widehat{\sigma}_\varepsilon^2\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\widehat{h}_c^2} = B_n + \frac{\sigma_\varepsilon^2\mathcal{A}_c}{h_c^2} + o_p(\bar{Q}_c). \tag{B.13}$$

Step 2: The rank term and the expansion of $\widehat{S}_c$*.*

Write the rank term as

$$\widehat{\sigma}_{\varepsilon v}^2\frac{d_c^2 + d_c}{n\widehat{h}_c^2} = \frac{\sigma_{\varepsilon v}^2 R_{d,c}}{nh_c^2}\cdot\frac{\widehat{\sigma}_{\varepsilon v}^2}{\sigma_{\varepsilon v}^2}\cdot\frac{d_c^2 + d_c}{R_{d,c}}\cdot\frac{h_c^2}{\widehat{h}_c^2}, \tag{B.14}$$

where, by Proposition 1, Lemmas C.1, D.3, and G.2, and the fixed $\sigma_{\varepsilon v} \ne 0$ in Assumption 1,

$$\begin{aligned}\frac{\widehat{\sigma}_{\varepsilon v}^2}{\sigma_{\varepsilon v}^2} &= 1 + O_p(n^{-1/2}), \\ \frac{d_c^2 + d_c}{R_{d,c}} &= 1 + o_p(1), \\ \frac{h_c^2}{\widehat{h}_c^2} &= 1 + O_p(n^{-1/2}).\end{aligned}$$

Since $\sigma_{\varepsilon v}^2 R_{d,c}/(nh_c^2) \le S_c$ by (20), (B.14) gives

$$\widehat{\sigma}_{\varepsilon v}^2\frac{d_c^2 + d_c}{n\widehat{h}_c^2} = \sigma_{\varepsilon v}^2\frac{R_{d,c}}{nh_c^2} + o_p(S_c). \tag{B.15}$$

Since $\bar{Q}_c/S_c \to 1$ by (C.6), the $o_p(\bar{Q}_c)$ remainder in (B.13) is equivalently $o_p(S_c)$. Substitute (B.13) and (B.15) into (26) and collect the two leading terms into S_c using (20); the remainders are $o_p(S_c)$, so

$$\frac{|\widehat{S}_c - B_n - S_c|}{S_c} \to_p 0. \tag{B.16}$$

Since J is fixed, this proves (32).

Step 3: Conversion to $\bar{Q}_c$*.*

Adding and subtracting S_c and applying the triangle inequality gives

$$\frac{|\widehat{S}_c - B_n - \bar{Q}_c|}{\bar{Q}_c} \le \frac{|\widehat{S}_c - B_n - S_c|}{S_c}\cdot\frac{S_c}{\bar{Q}_c} + \frac{|S_c - \bar{Q}_c|}{\bar{Q}_c}.$$

By (B.16) and (C.6), both terms on the right-hand side are $o_p(1)$; since J is fixed, this proves (33) and hence Theorem 2(i).

Step 4: The oracle inequality.

Define

$$\delta_n := \max_c \frac{|\widehat{S}_c - B_n - \bar{Q}_c|}{\bar{Q}_c}.$$

By Lemma C.3, $\min_{c\in[J]} \bar{Q}_c > 0$ for all sufficiently large n, so δ_n is well defined. Let

$$c_n^* := \min \arg\min_{c\in[J]} \bar{Q}_c.$$

Work on the event $\{\min_{c\in[J]} \widehat{h}_c > 0,\ \delta_n < 1\}$. By the definition of δ_n and optimality of $\widehat{c}$,

$$B_n + (1-\delta_n)\bar{Q}_{\widehat{c}} \le \widehat{S}_{\widehat{c}} \le \widehat{S}_{c_n^*} \le B_n + (1+\delta_n)\bar{Q}_{c_n^*}.$$

After cancelling B_n and using $\bar{Q}_{\widehat{c}} \ge \bar{Q}_{c_n^*}$,

$$0 \le \frac{\bar{Q}_{\widehat{c}}}{\bar{Q}_{c_n^*}} - 1 \le \frac{1+\delta_n}{1-\delta_n} - 1 = \frac{2\delta_n}{1-\delta_n}. \tag{B.17}$$

By (33) and (B.7),

$$\mathbb{P}\left(\min_{c\in[J]} \widehat{h}_c > 0,\ \delta_n < 1\right) \to 1, \qquad \delta_n \to_p 0.$$

Equation (B.17) therefore proves (34), and hence Theorem 2(ii). ☐

C Supporting Results for Theorem 1

Lemma C.1 (Denominator Moments and Rate)**.** *(i) Under Assumption 1, for each candidate,*

$$\mathbb{E}[\widehat{h}_c] = h_c + \frac{\sigma_v^2 \bar{d}_c}{n}, \tag{C.1}$$

$$\mathrm{Var}(\widehat{h}_c)^{1/2} \le 2\sigma_v \sqrt{\frac{H+\sigma_v^2}{n}}. \tag{C.2}$$

(ii) If Assumptions 2 and 3 also hold, with $J < \infty$ fixed, then

$$\max_{c\in[J]} |\widehat{h}_c - h_c| = O_p(n^{-1/2}). \tag{C.3}$$

(iii) Under the hypotheses of (ii), with $\mathcal{H}_n$ defined by (A.7), $\mathbb{P}(\mathcal{H}_n) \to 1$, and on $\mathcal{H}_n$ every candidate satisfies $\mathbb{E}_n[\widehat{\Pi}_{c,i}^2] > 0$ and $\widehat{\beta}_c = \mathbb{E}_n[\widehat{\Pi}_{c,i} y_i]/\widehat{h}_c$.

Proof. Step 1: Mean of $\widehat{h}_c$.

By the definition of $\widehat{h}_c$ in (24) and $x = \Pi + v$,

$$\begin{aligned}
\mathbb{E}[\widehat{h}_c] &= \overline{\mathbb{E}}[\widehat{\Pi}_{c,i} x_i] \\
&= \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}\Pi_i] + \overline{\mathbb{E}}[\widehat{\Pi}_{c,i} v_i] \\
&= h_c + \frac{\mathbb{E}[v'\widehat{\Pi}_c]}{n} = h_c + \frac{\sigma_v^2 \bar{d}_c}{n},
\end{aligned}$$

where the third equality uses the definition of h_c in (18) and the last uses (F.17) with $\bar{d}_c := \mathbb{E}[d_c]$ as in (A.2).

Step 2: Variance bound for $\widehat{h}_c$.

Note that $n\widehat{h}_c = x'\mu_c(x)$. At $b = 0$ and $c_0 = 0$, the offset variance bounds in (F.13) of Lemma F.5 give

$$\mathrm{Var}\{x'\mu_c(x)\} \leq 2\sigma_v^2 \left\{\mathbb{E}[\|\widehat{\Pi}_c\|_2^2] + \mathbb{E}[\|x\|_2^2]\right\} \leq 4\sigma_v^2 n(H + \sigma_v^2),$$

where the second inequality uses $\|\widehat{\Pi}_c\|_2 \leq \|x\|_2$ from (E.4) and $\overline{\mathbb{E}}[x_i^2] = \mathbb{E}_n[\Pi_i^2] + \sigma_v^2 = H + \sigma_v^2$. Dividing by n^2 and taking square roots proves (C.2).

Step 3: Root-n rate for $\widehat{h}_c - h_c$ and the event $\mathcal{H}_n$.

Equation (G.16) gives $\bar{d}_c = o(\sqrt{n})$, so by (C.1),

$$\sqrt{n}|\mathbb{E}[\widehat{h}_c] - h_c| = \frac{\sigma_v^2 \bar{d}_c}{\sqrt{n}} \to 0.$$

Chebyshev's inequality and (C.2) give $\sqrt{n}|\widehat{h}_c - \mathbb{E}[\widehat{h}_c]| = O_p(1)$, because $n\,\mathrm{Var}(\widehat{h}_c) \leq 4\sigma_v^2(H + \sigma_v^2)$, which is bounded by Assumption 1. The triangle inequality then yields

$$\begin{aligned}\sqrt{n}|\widehat{h}_c - h_c| &\leq \sqrt{n}|\widehat{h}_c - \mathbb{E}[\widehat{h}_c]| + \sqrt{n}|\mathbb{E}[\widehat{h}_c] - h_c| \\ &= O_p(1) + o(1) = O_p(1).\end{aligned}$$

Since $[J]$ is finite, (C.3) follows; together with $\min_c h_c \geq \underline{h}$ from (G.14) it gives $\mathbb{P}(\mathcal{H}_n) \to 1$. On $\mathcal{H}_n$, by Cauchy–Schwarz, for every $c \in [J]$,

$$0 < \frac{\underline{h}}{2} \leq \widehat{h}_c = \mathbb{E}_n[\widehat{\Pi}_{c,i} x_i] \leq \mathbb{E}_n[\widehat{\Pi}_{c,i}^2]^{1/2} \mathbb{E}_n[x_i^2]^{1/2},$$

so $\mathbb{E}_n[\widehat{\Pi}_{c,i}^2] > 0$ and $\widehat{\beta}_c$ equals the ratio in (6). □

Lemma C.2 (Moments of Γ_c)**.** *Suppose Assumptions 1 and 2 hold, with $J < \infty$ fixed. Then, for each candidate and all sufficiently large n,*

$$\mathbb{E}[\Gamma_c^2] = \frac{\sigma_\varepsilon^2}{H} + S_c, \tag{C.4}$$

and

$$\mathbb{E}[(\Gamma_c - L_n)^2] \leq 2S_c. \tag{C.5}$$

Proof. Suppress the candidate subscript. By (G.14) of Lemma G.2, $h \geq \underline{h} > 0$ for all sufficiently large n, so Γ is well defined eventually. For $\mathcal{A}$, R_d, and Γ defined by (19), (A.2), and (A.9), the LASSO moments derived in (F.15) of Lemma F.6 give

$$\mathbb{E}[\Gamma^2] = \frac{\mathbb{E}[(\varepsilon'\widehat{\Pi})^2]}{nh^2} = \frac{\sigma_\varepsilon^2 \overline{\mathbb{E}}[\widehat{\Pi}_i^2]}{h^2} + \frac{\sigma_{\varepsilon v}^2 R_d}{nh^2} = \frac{\sigma_\varepsilon^2}{H} + S,$$

where the last equality substitutes $\overline{\mathbb{E}}[\widehat{\Pi}_i^2] = \mathcal{A} + h^2/H$ from (19) and the definition of S in (20).

For $\widetilde{\zeta} := v'\widetilde{\Pi} - \sigma_v^2 d$, the decomposition (A.6) gives $\varepsilon'\widetilde{\Pi} - \sigma_{\varepsilon v} d = (\sigma_{\varepsilon v}/\sigma_v^2)\widetilde{\zeta} + e'\widetilde{\Pi}$. Since e is independent of

v with mean zero, (A.4) and (D.3) give

$$\begin{aligned}\mathbb{E}[(\varepsilon'\widetilde{\Pi} - \sigma_{\varepsilon v} d)^2] &= \left(\sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2}\right) n\mathcal{A} + \frac{\sigma_{\varepsilon v}^2}{\sigma_v^4}\mathbb{E}[\widetilde{\zeta}^2] \\ &= \sigma_\varepsilon^2 n\mathcal{A} + \sigma_{\varepsilon v}^2 \bar{d}.\end{aligned}$$

By (A.3), $\Gamma - L_n = \varepsilon'\widetilde{\Pi}/(\sqrt{n}\,h)$, so $R_d = \mathbb{E}[d^2] + \bar{d}$ and (20) yield

$$\begin{aligned}\mathbb{E}[(\Gamma - L_n)^2] &= \frac{1}{nh^2}\mathbb{E}[(\varepsilon'\widetilde{\Pi})^2] \\ &= \frac{1}{nh^2}\mathbb{E}[(\varepsilon'\widetilde{\Pi} - \sigma_{\epsilon v} d + \sigma_{\epsilon v} d)^2] \\ &\le \frac{2}{nh^2}\left\{\mathbb{E}[(\varepsilon'\widetilde{\Pi} - \sigma_{\varepsilon v} d)^2] + \sigma_{\varepsilon v}^2 \mathbb{E}[d^2]\right\} \\ &= 2S.\end{aligned}$$

□

Lemma C.3 (AMSE Equivalence and Rates)**.** *Suppose Assumptions 1, 2, and 3 hold, with $J < \infty$ fixed. Then, uniformly over candidates,*

$$\frac{\bar{Q}_c}{S_c} \to 1. \tag{C.6}$$

Moreover, uniformly over candidates,

$$\bar{Q}_c \asymp \mathcal{A}_c + \frac{\bar{d}_c^2}{n}, \qquad \max_{c\in[J]} \bar{Q}_c \to 0, \qquad \min_{c\in[J]} n\bar{Q}_c \to \infty. \tag{C.7}$$

For every candidate, $\bar{Q}_c > 0$ for all sufficiently large n. The bounds in (G.24) continue to hold with $\bar{Q}_c$ in place of S_c.

Proof. Step 1: Decomposition $\Lambda = \Psi - \varrho L_n$.

Suppress the candidate subscript throughout the proof. Let $\widetilde{\Pi}(\cdot)$ be the centered-fit map of (A.3) and $\widetilde{\Pi} := \widetilde{\Pi}(v)$. By (G.14), $\underline{h} \le h \le C$ for all sufficiently large n; constants below absorb h and $1/h$. Define

$$\begin{aligned}\varrho &:= \frac{(\Pi + v)'\widetilde{\Pi}}{nh}, \\ U &:= \Gamma - L_n = \frac{\varepsilon'\widetilde{\Pi}}{\sqrt{n}\,h}, \\ \Psi &:= (1-q)L_n + U.\end{aligned} \tag{C.8}$$

Substituting $\widehat{\Pi} = (h/H)\Pi + \widetilde{\Pi}$ and $q = \Pi'v/(nH)$ into (24) and (A.10) gives

$$\begin{aligned}\widehat{h} &= \frac{1}{n}(\Pi + v)'\left(\frac{h}{H}\Pi + \widetilde{\Pi}\right) = h + hq + h\varrho, \\ \frac{\widehat{h} - h}{h} &= q + \varrho, \\ \Lambda &= \Gamma - (q+\varrho)L_n = \Psi - \varrho L_n.\end{aligned} \tag{C.9}$$

Step 2: Moments of ϱ and $\dot{\varrho}$.

Write $P := \dot{\mu}(\Pi + v)$, noting the properties of P from (A.1), and define

$$u_n := \frac{\Pi' v}{\sigma_v \sqrt{nH}}, \qquad v_\perp := \left(I - \frac{\Pi\Pi'}{nH}\right) v,$$

So that we can rewrite

$$v = \frac{\sigma_v u_n}{\sqrt{nH}}\Pi + v_\perp, \qquad q = \frac{\sigma_v u_n}{\sqrt{nH}}. \tag{C.10}$$

The pair $(u_n, v_\perp)$ is jointly Gaussian, and since $\mathrm{Var}(v) = \sigma_v^2 I$ and $\Pi'\Pi = nH$,

$$\mathrm{Cov}(u_n, v_\perp) = \frac{\sigma_v}{\sqrt{nH}}\Pi'\left(I - \frac{\Pi\Pi'}{nH}\right) = \frac{\sigma_v}{\sqrt{nH}}(\Pi' - \Pi') = 0.$$

Hence $u_n \sim N(0,1)$ independent of $v_\perp$. For locally Lipschitz $g : \mathbb{R}^n \to \mathbb{R}$, view $g(v) = g(\sigma_v u_n \Pi/\sqrt{nH} + v_\perp)$ as a function of $(u_n, v_\perp)$. Holding $v_\perp$ fixed,

$$\dot{g} := \frac{\partial g}{\partial u_n} = \frac{\sigma_v}{\sqrt{nH}}\Pi'\nabla g(v) \qquad \text{almost everywhere.} \tag{C.11}$$

Set $f(v) := (\Pi + v)'\widetilde{\Pi}$, so that we can rewrite ϱ as $\varrho = f(v)/(nh)$. The map $\widetilde{\Pi}(\cdot)$ is one-Lipschitz by Lemma E.1. Rademacher's theorem and the chain and product rules give, almost everywhere,

$$\begin{aligned}
D\widetilde{\Pi}(v) &= D\mu(\Pi + v) = P, \\
\nabla f(v) &= \widetilde{\Pi} + D\widetilde{\Pi}(v)'(\Pi + v) \\
&= \widetilde{\Pi} + P'(\Pi + v) \\
&= \widetilde{\Pi} + P(\Pi + v).
\end{aligned} \tag{C.12}$$

where the last equality follows from $P = P'$ in (A.1). Taking expectations in the definition of f,

$$\mathbb{E}[f(v)] = \mathbb{E}[\Pi'\widetilde{\Pi}] + \mathbb{E}[v'\widetilde{\Pi}] = 0 + \sigma_v^2\mathbb{E}[d] = \sigma_v^2\bar{d}, \tag{C.13}$$

where $\mathbb{E}[\Pi'\widetilde{\Pi}] = 0$ by (A.4) and, since $\mathbb{E}[v'\Pi] = 0$, $\mathbb{E}[v'\widetilde{\Pi}] = \mathbb{E}[v'\widehat{\Pi}] = \sigma_v^2\mathbb{E}[d]$ by (F.17).

Since $\varrho = f(v)/(nh)$, (A.4), (C.13), and (F.13) at $b = (h/H)\Pi$ and $c_0 = 0$ give

$$\mathbb{E}[\varrho^2] \le \frac{2\sigma_v^2}{nh^2}(\mathcal{A} + H + \sigma_v^2) + \frac{\sigma_v^4\bar{d}^2}{n^2h^2} \le \frac{C}{n}, \tag{C.14}$$

where the final inequality follows from $h \ge \underline{h}$, $\mathcal{A} + H \le C$, and $\bar{d} = o(\sqrt{n})$ by Lemma G.2.

Consider now $\dot{\varrho} = \partial\varrho/\partial u_n$. For $m := \Pi/(\sigma_v\sqrt{nH})$ and an arbitrary locally Lipschitz $g : \mathbb{R}^n \to \mathbb{R}$ as above,

$$Dm = 0, \qquad \operatorname{div} m = 0, \qquad v'm = u_n, \qquad \sigma_v^2 m'\nabla g = \dot{g}.$$

Also, by (C.11) and Lemma F.3,

$$\mathbb{E}[u_n g(v)] = \mathbb{E}[\dot{g}(v)], \tag{C.15}$$

for every locally Lipschitz g satisfying the growth condition of Lemma F.3. Here $g(v)$ is viewed as the function $u_n \mapsto g(\sigma_v u_n \Pi/\sqrt{nH} + v_\perp)$, with $v_\perp$ held fixed, and $\dot{g}(v)$ is its derivative with respect to u_n. Every $g(v)$ used below is a bounded-degree polynomial of u_n, ϱ, and $\Pi'\widetilde{\Pi}$. The one-Lipschitz property of μ and (E.4) of Lemma E.1, with $\|\Pi\|_2^2 = nH$, give

$$\begin{aligned}
\|\widetilde{\Pi}\|_2 &\leq C(\sqrt{n} + \|v\|_2), \\
|\varrho(v)| &\leq C\left(1 + \frac{\|v\|_2^2}{n}\right), \\
\|\nabla\varrho(v)\|_2 &\leq C\left(\frac{1}{\sqrt{n}} + \frac{\|v\|_2}{n}\right) \quad \text{almost everywhere,}
\end{aligned}$$

where the last inequality uses (C.12). As functions of v, u_n is linear and ϱ and $\Pi'\widetilde{\Pi}$ are locally Lipschitz. Thus every polynomial in $(u_n, \varrho, \Pi'\widetilde{\Pi})$ of finite degree g satisfies into the polynomial-growth condition of Lemma F.3.

Combine (C.11) and (C.12) with $\varrho = f(v)/(nh)$ and $P' = P$ to see that

$$\dot{\varrho} = \frac{\sigma_v}{nh\sqrt{nH}}\Pi'\{\widetilde{\Pi} + P(\Pi + v)\} = \frac{\sigma_v}{nh\sqrt{nH}}\left\{\Pi'\widetilde{\Pi} + (\Pi + v)'P\Pi\right\}. \tag{C.16}$$

Cauchy–Schwarz, $(a+b)^2 \leq 2a^2 + 2b^2$, contraction of P, $\|\Pi\|_2^2 = nH$, and (A.4) give

$$\begin{aligned}
\mathbb{E}[\dot{\varrho}^2] &\leq \frac{2\sigma_v^2\|\Pi\|_2^2}{n^3h^2H}\left\{\mathbb{E}[\|\widetilde{\Pi}\|_2^2] + \mathbb{E}[\|P(\Pi + v)\|_2^2]\right\} \\
&\leq \frac{2\sigma_v^2}{n^2h^2}\left\{\mathbb{E}[\|\widetilde{\Pi}\|_2^2] + \mathbb{E}[\|\Pi + v\|_2^2]\right\} \\
&= \frac{2\sigma_v^2}{nh^2}(\mathcal{A} + H + \sigma_v^2) \leq \frac{C}{n}.
\end{aligned} \tag{C.17}$$

The growth bounds above give $\mathbb{E}[u_n^2\varrho^2] < \infty$ and justify applying (C.15) to $u_n\varrho^2$. Thus

$$\begin{aligned}
\mathbb{E}[u_n^2\varrho^2] &= \mathbb{E}[\varrho^2] + 2\mathbb{E}[u_n\varrho\dot{\varrho}] \\
&\leq \mathbb{E}[\varrho^2] + \tfrac{1}{2}\mathbb{E}[u_n^2\varrho^2] + 2\mathbb{E}[\dot{\varrho}^2],
\end{aligned}$$

where the inequality uses Young's inequality $2ab \leq a^2/2 + 2b^2$ with $a = u_n\varrho$ and $b = \dot{\varrho}$. After moving the $\frac{1}{2}\mathbb{E}[u_n^2\varrho^2]$ term to the left-hand side and multiplying by 2, we get

$$\mathbb{E}[u_n^2\varrho^2] \leq 2\mathbb{E}[\varrho^2] + 4\mathbb{E}[\dot{\varrho}^2]$$

This, together with (C.14) and (C.17), yields

$$\mathbb{E}[\varrho^2(1 + u_n^2)] \leq 3\mathbb{E}[\varrho^2] + 4\mathbb{E}[\dot{\varrho}^2] \leq \frac{C}{n}. \tag{C.18}$$

Step 3: Moments of $\mathbb{E}[\varrho(1-q)]$ *and* $\mathbb{E}[\varrho(1-q)u_n^2]$.

Applying (C.15) to ϱ, $u_n\varrho$, and $u_n^2\varrho$ gives

$$\begin{aligned} \mathbb{E}[u_n\varrho] &= \mathbb{E}[\dot{\varrho}], \\ \mathbb{E}[u_n^2\varrho] &= \mathbb{E}[\varrho] + \mathbb{E}[u_n\dot{\varrho}], \\ \mathbb{E}[u_n^3\varrho] &= 2\mathbb{E}[\dot{\varrho}] + \mathbb{E}[u_n^2\dot{\varrho}]. \end{aligned} \tag{C.19}$$

By (C.17), $\mathbb{E}[u_n^4] = 3$, and Cauchy–Schwarz,

$$\begin{aligned} |\mathbb{E}[\dot{\varrho}]| &\le \mathbb{E}[\dot{\varrho}^2]^{1/2} \le \frac{C}{\sqrt{n}}, \\ |\mathbb{E}[u_n^2\dot{\varrho}]| &\le \mathbb{E}[u_n^4]^{1/2}\mathbb{E}[\dot{\varrho}^2]^{1/2} \le \frac{C}{\sqrt{n}}. \end{aligned} \tag{C.20}$$

For $\mathbb{E}[u_n\dot{\varrho}]$, begin by noting that $P\mu(\Pi+v) = \mu(\Pi+v)$ almost surely by (A.1). Substitute $\mu(\Pi+v) = (h/H)\Pi + \widetilde{\Pi}$ into both sides of this identity to get

$$\begin{aligned} (I-P)\Pi &= -\frac{H}{h}(I-P)\widetilde{\Pi}, \\ \implies P\Pi &= \Pi + \frac{H}{h}(I-P)\widetilde{\Pi}, \\ \implies (\Pi+v)'P\Pi &= (\Pi+v)'\Pi + \frac{H}{h}(\Pi+v)'(I-P)\widetilde{\Pi}. \end{aligned} \tag{C.21}$$

Since $D\widetilde{\Pi}(v) = P$ almost everywhere by (C.12) and $P^2 = P$ by (A.1), we have both $\nabla(\Pi'\widetilde{\Pi}) = P\Pi$, and $0 \le (P\Pi)'(P\Pi) = \Pi'P\Pi \le nH$. Thus (C.15) along with (C.11) applied to the function $\Pi'\widetilde{\Pi}$ gives

$$|\mathbb{E}[u_n\Pi'\widetilde{\Pi}]| = \frac{\sigma_v}{\sqrt{nH}}\mathbb{E}[\Pi'P\Pi] \le \sigma_v\sqrt{nH}. \tag{C.22}$$

Meanwhile, since $\mathbb{E}[u_n] = 0$ and $\Pi'v = \sigma_v\sqrt{nH}\,u_n$,

$$\mathbb{E}[u_n(\Pi+v)'\Pi] = \mathbb{E}[u_n\Pi'v] = \sigma_v\sqrt{nH}\,\mathbb{E}[u_n^2] = \sigma_v\sqrt{nH}. \tag{C.23}$$

Because $I-P$ is a contraction by (A.1), $\|(I-P)\widetilde{\Pi}\|_2 \le \|\widetilde{\Pi}\|_2$ almost surely. Taking expectations and applying the last identity in (A.4) gives

$$\mathbb{E}[\|(I-P)\widetilde{\Pi}\|_2^2] \le \mathbb{E}[\|\widetilde{\Pi}\|_2^2] = n\mathcal{A}.$$

The independent Gaussian components u_n and $v_\perp$ in (C.10) satisfy $\Pi'v_\perp = 0$ and

$$\mathbb{E}[\|v_\perp\|_2^2] = \sigma_v^2\operatorname{tr}\left(I - \frac{\Pi\Pi'}{nH}\right) = \sigma_v^2(n-1),$$

where the last equality uses linearity of trace, $\operatorname{tr}(I) = n$, and $\operatorname{tr}(\Pi\Pi') = \Pi'\Pi = nH$. By (C.10),

$$\|\Pi+v\|_2^2 = \left\|\left(1 + \frac{\sigma_v u_n}{\sqrt{nH}}\right)\Pi + v_\perp\right\|_2^2$$

$$
\begin{aligned}
&= \left(1+\frac{\sigma_v u_n}{\sqrt{nH}}\right)^2 nH + \|v_\perp\|_2^2 \\
&= (\sqrt{nH}+\sigma_v u_n)^2 + \|v_\perp\|_2^2,
\end{aligned}
$$

where the second equality uses $\|\Pi\|_2^2 = nH$ and $\Pi' v_\perp = 0$, which makes the cross term zero. Thus, by independence of u_n and $v_\perp$,

$$
\begin{aligned}
\mathbb{E}[u_n^2\|\Pi+v\|_2^2] &= \mathbb{E}[u_n^2(\sqrt{nH}+\sigma_v u_n)^2] + \mathbb{E}[u_n^2]\mathbb{E}[\|v_\perp\|_2^2] \\
&= nH\mathbb{E}[u_n^2] + 2\sigma_v\sqrt{nH}\,\mathbb{E}[u_n^3] + \sigma_v^2\mathbb{E}[u_n^4] + \sigma_v^2(n-1) \\
&= nH + 3\sigma_v^2 + \sigma_v^2(n-1) = nH + \sigma_v^2(n+2) \le Cn.
\end{aligned}
$$

The second equality uses the derived expression for $\mathbb{E}[\|v_\perp\|_2^2]$ from above and $\mathbb{E}[u_n^2]=1$. The third uses the fact that $u_n \sim N(0,1)$ so that $\mathbb{E}[u_n^2]=1$, $\mathbb{E}[u_n^3]=0$, and $\mathbb{E}[u_n^4]=3$. The final inequality uses $H \le \overline{H}$ and $n+2 \le 3n$. Cauchy–Schwarz then gives

$$
\begin{aligned}
\left|\mathbb{E}\Big[u_n(\Pi+v)'(I-P)\widetilde{\Pi}\Big]\right| &\le \mathbb{E}[u_n^2\|\Pi+v\|_2^2]^{1/2}\mathbb{E}[\|(I-P)\widetilde{\Pi}\|_2^2]^{1/2} \\
&\le Cn\sqrt{\mathcal{A}}. \qquad \text{(C.24)}
\end{aligned}
$$

Combining the above bounds yields

$$
\begin{aligned}
|\mathbb{E}[u_n\dot{\varrho}]| &= \frac{\sigma_v}{nh\sqrt{nH}}\left|\mathbb{E}[u_n\Pi'\widetilde{\Pi}] + \mathbb{E}[u_n(\Pi+v)'P\Pi]\right| \\
&\le \frac{\sigma_v}{nh\sqrt{nH}}\left\{|\mathbb{E}[u_n\Pi'\widetilde{\Pi}]| + |\mathbb{E}[u_n(\Pi+v)'\Pi]| + \frac{H}{h}\left|\mathbb{E}[u_n(\Pi+v)'(I-P)\widetilde{\Pi}]\right|\right\} \\
&\le \frac{\sigma_v}{nh\sqrt{nH}}\left\{2\sigma_v\sqrt{nH} + \frac{H}{h}Cn\sqrt{\mathcal{A}}\right\} \\
&\le C\left\{\sqrt{\frac{\mathcal{A}}{n}} + \frac{1}{n}\right\}, \qquad \text{(C.25)}
\end{aligned}
$$

where the first equality follows from (C.16), the first inequality from (C.21) and the triangle inequality, the second from (C.22)–(C.24), and the last from (G.14) and Assumption 1.

Since $\varrho = f(v)/(nh)$, (C.13) gives $\mathbb{E}[\varrho] = \mathbb{E}[f(v)]/(nh) = \sigma_v^2\bar{d}/(nh)$. With $q = \sigma_v u_n/\sqrt{nH}$, (C.19) gives

$$
\begin{aligned}
\mathbb{E}[\varrho(1-q)] &= \mathbb{E}[\varrho] - \frac{\sigma_v}{\sqrt{nH}}\mathbb{E}[\dot{\varrho}], \\
\mathbb{E}[\varrho(1-q)u_n^2] &= \mathbb{E}[\varrho] + \mathbb{E}[u_n\dot{\varrho}] \\
&\quad - \frac{\sigma_v}{\sqrt{nH}}\left\{2\mathbb{E}[\dot{\varrho}] + \mathbb{E}[u_n^2\dot{\varrho}]\right\}.
\end{aligned}
$$

By (C.20) and (C.25), respectively,

$$
|\mathbb{E}[\dot{\varrho}]| + |\mathbb{E}[u_n^2\dot{\varrho}]| \le \frac{C}{\sqrt{n}}, \qquad |\mathbb{E}[u_n\dot{\varrho}]| \le C\left\{\sqrt{\frac{\mathcal{A}}{n}} + \frac{1}{n}\right\}.
$$

Applying the triangle inequality to the two identities above and substituting these bounds and the preceding formula for $\mathbb{E}[\varrho]$ gives

$$|\mathbb{E}[\varrho(1-q)]| + |\mathbb{E}[\varrho(1-q)u_n^2]| \leq C\left\{\sqrt{\frac{\mathcal{A}}{n}} + \frac{\bar{d}}{n} + \frac{1}{n}\right\}, \tag{C.26}$$

where the implicit constant uses $h \geq \underline{h}$ by (G.14) and $H \geq \underline{H}$ by Assumption 1.

Step 4: Combining bounds.

Expand $\mathbb{E}[\Gamma^2]$ using $\Gamma = L_n + U$ from (C.8). Since $\mathbb{E}[L_n^2] = \sigma_\varepsilon^2/H$ by (F.8), comparing with (C.4) leaves

$$S = 2\mathbb{E}[L_n U] + \mathbb{E}[U^2]. \tag{C.27}$$

Expanding Ψ and substituting (C.27) and (F.9) gives

$$\begin{aligned} \mathbb{E}[\Psi^2] - C_{0n} - S &= 2\mathbb{E}[(1-q)L_n U] - 2\mathbb{E}[L_n U] = -2\mathbb{E}[qL_n U], \\ |\mathbb{E}[\Psi^2] - C_{0n} - S| &\leq 2\mathbb{E}[q^2 L_n^2]^{1/2}\mathbb{E}[U^2]^{1/2} \leq C\sqrt{\frac{S}{n}}, \end{aligned} \tag{C.28}$$

where the second inequality uses (C.5) and (F.8).

Substitute e and L_n as defined by (A.6) and (A.8), and u_n as defined immediately above (C.10), to write

$$L_n = \frac{\sigma_{\varepsilon v}}{\sigma_v \sqrt{H}}\, u_n + \frac{\Pi' e}{\sqrt{n}\, H}.$$

Because e is independent of v with $\mathrm{Var}(e_i) = \sigma_\varepsilon^2 - \sigma_{\varepsilon v}^2/\sigma_v^2$, conditioning on v yields

$$\mathbb{E}[L_n^2 \mid v] = \frac{\sigma_\varepsilon^2 - \sigma_{\varepsilon v}^2/\sigma_v^2}{H} + \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2 H} u_n^2. \tag{C.29}$$

Since ϱ, q, and u_n are functions of v alone, iterated expectations conditional on v, (C.18), and (C.29) give

$$\mathbb{E}[\varrho^2 L_n^2] = \mathbb{E}\Big[\varrho^2\, \mathbb{E}[L_n^2 \mid v]\Big] \leq C\, \mathbb{E}[\varrho^2(1+u_n^2)] \leq \frac{C}{n}. \tag{C.30}$$

The same conditioning step with (C.29) bounds $|\mathbb{E}[\varrho(1-q)L_n^2]|$ by $C\{|\mathbb{E}[\varrho(1-q)]| + |\mathbb{E}[\varrho(1-q)u_n^2]|\}$. Hence, by (C.26),

$$|\mathbb{E}[\varrho(1-q)L_n^2]| \leq C\left\{\sqrt{\frac{\mathcal{A}}{n}} + \frac{\bar{d}}{n} + \frac{1}{n}\right\} \leq C\left(\sqrt{\frac{S}{n}} + \frac{1}{n}\right), \tag{C.31}$$

where the second inequality applies the first bound in (G.24) of Proposition G.1. By Cauchy–Schwarz, (C.5), and (C.30),

$$|\mathbb{E}[\varrho L_n U]| \leq \mathbb{E}[\varrho^2 L_n^2]^{1/2}\mathbb{E}[U^2]^{1/2} \leq C\sqrt{\frac{S}{n}}. \tag{C.32}$$

Combining (C.31) and (C.32) over the two pieces of Ψ in (C.8) gives

$$|\mathbb{E}[\varrho L_n \Psi]| \leq C\left(\sqrt{\frac{S}{n}} + \frac{1}{n}\right). \tag{C.33}$$

By (A.13) and (C.9),

$$\begin{aligned}|\bar{Q}-S| &= |\mathbb{E}[\Lambda^2]-C_{0n}-S|\\ &\le |\mathbb{E}[\Psi^2]-C_{0n}-S|+2|\mathbb{E}[\varrho L_n\Psi]|+\mathbb{E}[\varrho^2 L_n^2]\\ &\le C\left(\sqrt{\frac{S}{n}}+\frac{1}{n}\right),\end{aligned}\tag{C.34}$$

where (C.28), (C.30), and (C.33) bound the first, third, and second terms, respectively. By (G.23) of Proposition G.1, $\min_c nS_c \to \infty$. Since $[J]$ is finite and the constants do not depend on the candidate, (C.34) yields, uniformly over candidates,

$$\left|\frac{\bar{Q}}{S}-1\right| \le C\left\{\frac{1}{\sqrt{nS}}+\frac{1}{nS}\right\}\to 0.$$

Restoring the candidate subscript, $\max_{c\in[J]}|\bar{Q}_c/S_c-1|\to 0$, which is (C.6). Since $\bar{Q}_c/S_c\to 1$ uniformly, the conclusions in (G.23) of Proposition G.1 hold with $\bar{Q}_c$ in place of S_c, giving (C.7). Finally, since $\min_c nS_c\to\infty$ and $\max_c|\bar{Q}_c/S_c-1|\to 0$, for all sufficiently large n every candidate has $\bar{Q}_c>0$ and $S_c\le 2\bar{Q}_c$, so each bound in (G.24) holds with $\bar{Q}_c$ in place of S_c after adjusting the constant, the final statement of the lemma. □

D Supporting Results for Theorem 2

Lemma D.1 (Centered-Fit Moments). *Under Assumption 1, for each candidate,*

$$\mathbb{E}\left[\left\{\mathbb{E}_n[\Pi_i\widetilde{\Pi}_{c,i}]\right\}^2\right]^{1/2}\le \sigma_v\sqrt{\frac{H}{n}},\tag{D.1}$$

$$\mathbb{E}\left[\left\{\mathbb{E}_n[\widetilde{\Pi}_{c,i}^2]-\mathcal{A}_c\right\}^2\right]^{1/2}\le 2\sigma_v\sqrt{\frac{\mathcal{A}_c}{n}}.\tag{D.2}$$

Moreover, with $\widetilde{\zeta}_c := n\mathbb{E}_n[v_i\widetilde{\Pi}_{c,i}]-\sigma_v^2 d_c$,

$$\mathbb{E}[\widetilde{\zeta}_c^2]=\sigma_v^2 n\mathcal{A}_c+\sigma_v^4\bar{d}_c.\tag{D.3}$$

Proof. Fix a candidate and suppress its subscript. Take $b:=(h/H)\Pi$, which is nonrandom because the design is fixed and h is the population moment defined in (18), so that the map m of Lemma F.5 is the centered-fit map of (A.3) and $M=\widetilde{\Pi}$. For (D.1), apply (F.11) at $g=\Pi$: since $\|\Pi\|_2^2=nH$ and $\mathbb{E}[\Pi'\widetilde{\Pi}]=0$ by (A.4),

$$\mathbb{E}[(\Pi'\widetilde{\Pi})^2]=\mathrm{Var}(\Pi'\widetilde{\Pi})\le\sigma_v^2 nH,$$

which proves (D.1). For (D.2), $\mathbb{E}[\|\widetilde{\Pi}\|_2^2]=n\mathcal{A}$ from (A.4) and (F.12) give

$$\mathrm{Var}(\|\widetilde{\Pi}\|_2^2)\le 4\sigma_v^2 n\mathcal{A},$$

which proves (D.2). Finally, $\widetilde{\zeta}=v'\widetilde{\Pi}-\sigma_v^2 d$, so the second identity in (F.14), again with $\mathbb{E}[\|\widetilde{\Pi}\|_2^2]=n\mathcal{A}$, is exactly (D.3). □

Lemma D.2 (Normalized Fitted-Moment Expansion)**.** *Suppose Assumptions 1 to 3 hold, with $J < \infty$ fixed. On $\mathcal{H}_n$ defined by (A.7), define*

$$R^{\text{fit}}_{n,c} := \frac{\mathbb{E}_n[\widehat{\Pi}^2_{c,i}]}{\widehat{h}^2_c} - \frac{1}{H} + \frac{2\mathbb{E}_n[\Pi_i v_i]}{H^2} - \frac{\mathcal{A}_c}{h_c^2}, \tag{D.4}$$

and set $R^{\text{fit}}_{n,c} := 0$ on $\mathcal{H}_n^c$. Then

$$\max_{c\in[J]} \frac{|R^{\text{fit}}_{n,c}|}{\bar{Q}_c} \to_p 0. \tag{D.5}$$

Proof. Step 1: Exact identity for R^{fit}_n.

Fix a candidate and suppress its subscript. Recall $\delta_h = \widehat{h} - h$ and $q = \mathbb{E}_n[\Pi_i v_i]/H$ from (A.7) and (A.11). Since $\widehat{\Pi} = (h/H)\Pi + \widetilde{\Pi}$, $x = \Pi + v$, and $\mathbb{E}_n[\Pi_i^2] = H$,

$$\begin{aligned}
\mathbb{E}_n[\widehat{\Pi}_i^2] &= \frac{h^2}{H} + \frac{2h}{H}\mathbb{E}_n[\Pi_i\widetilde{\Pi}_i] + \mathbb{E}_n[\widetilde{\Pi}_i^2],\\
\delta_h &= hq + \mathbb{E}_n[\Pi_i\widetilde{\Pi}_i] + \mathbb{E}_n[\widetilde{\Pi}_i v_i].
\end{aligned}$$

Substituting into (D.4) gives, on $\mathcal{H}_n$,

$$\begin{aligned}
\widehat{h}^2 R^{\text{fit}}_n = \mathbb{E}_n[\widetilde{\Pi}_i^2] - \mathcal{A} - \frac{2h}{H}\mathbb{E}_n[\widetilde{\Pi}_i v_i] - \frac{2\mathcal{A}}{h}\delta_h\\
+ \frac{2q}{H}(2h\delta_h + \delta_h^2) - \left(\frac{1}{H} + \frac{\mathcal{A}}{h^2}\right)\delta_h^2.
\end{aligned} \tag{D.6}$$

Step 2: Bounds on the terms of (D.6).

By (C.3) and (F.10), $\delta_h = O_p(n^{-1/2})$ and $q = O_p(n^{-1/2})$. By (G.16), $d/\bar{d} = O_p(1)$ and $\bar{d} \to \infty$. The bounds in (G.24) hold with $\bar{Q}$ in place of $\mathcal{S}$ by Lemma C.3. Applying these resulting bounds in $\bar{Q}$ gives

$$\frac{\sqrt{\mathcal{A}/n}}{\bar{Q}} \le \frac{C}{\bar{d}}, \qquad \frac{\bar{d}/n}{\bar{Q}} \le \frac{C}{\bar{d}}, \qquad \frac{\sqrt{\bar{d}}/n}{\bar{Q}} \le \frac{C}{\bar{d}^{3/2}}. \tag{D.7}$$

Together with Lemma D.1, these imply

$$\begin{aligned}
\mathbb{E}\left[\left\{\frac{\mathbb{E}_n[\widetilde{\Pi}_i^2] - \mathcal{A}}{\bar{Q}}\right\}^2\right] &\le \frac{4\sigma_v^2\mathcal{A}}{n\bar{Q}^2} \le \frac{C}{\bar{d}^2},\\
\mathbb{E}\left[\left(\frac{\widetilde{\zeta}}{n\bar{Q}}\right)^2\right] &= \frac{\sigma_v^2 n\mathcal{A} + \sigma_v^4\bar{d}}{n^2\bar{Q}^2} \le C\left(\frac{1}{\bar{d}^2} + \frac{1}{\bar{d}^3}\right).
\end{aligned}$$

By Markov's inequality,

$$\begin{aligned}
\frac{|\mathbb{E}_n[\widetilde{\Pi}_i^2] - \mathcal{A}|}{\bar{Q}} &= o_p(1),\\
\frac{|\widetilde{\zeta}|}{n\bar{Q}} &= o_p(1).
\end{aligned}$$

Also, by (D.7),

$$\frac{\sigma_v^2 d}{n\bar{Q}} \le C\frac{d}{\bar{d}}\frac{1}{\bar{d}} = o_p(1).$$

Since $n\mathbb{E}_n[\widetilde{\Pi}_i v_i] = \widetilde{\zeta} + \sigma_v^2 d$ and $\mathcal{A} \le C\bar{Q}$ by (C.7),

$$\frac{|\mathbb{E}_n[\widetilde{\Pi}_i v_i]|}{\bar{Q}} = o_p(1), \qquad \frac{\mathcal{A}|\delta_h|}{\bar{Q}} \le C|\delta_h| = o_p(1).$$

Finally, since $\underline{h} \le h \le C$ and $\mathcal{A} = o(1)$ by Lemma G.2, while $\underline{H} \le H \le \overline{H}$ by Assumption 1(ii),

$$\frac{2q}{H}(2h\delta_h + \delta_h^2) - \left(\frac{1}{H} + \frac{\mathcal{A}}{h^2}\right)\delta_h^2 = O_p(n^{-1}) = o_p(\bar{Q}),$$

since $n\bar{Q} \to \infty$ by (C.7). Divide (D.6) by $\widehat{h}^2\bar{Q}$. On $\mathcal{H}_n$, $\widehat{h} \ge \underline{h}/2$, so the bounds above give $R_n^{\mathrm{fit}}/\bar{Q} = o_p(1)$. Since $R_n^{\mathrm{fit}} = 0$ on $\mathcal{H}_n^c$ and $[J]$ is finite, (D.5) follows. □

Lemma D.3 (Pilot and Error-Moment Rates)**.** *Suppose Assumptions 1 to 3 hold and $J < \infty$ is fixed. Let $\check{c}$ be any (possibly data-dependent) $[J]$-valued index, and let $\check{\beta} := \widehat{\beta}_{\check{c}}$ and $\widehat{\sigma}_\varepsilon^2, \widehat{\sigma}_{\varepsilon v}$ be as in (25). Then*

$$\check{\beta} - \beta = O_p(n^{-1/2}), \tag{D.8}$$
$$\widehat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2 = O_p(n^{-1/2}), \qquad \widehat{\sigma}_{\varepsilon v} - \sigma_{\varepsilon v} = O_p(n^{-1/2}). \tag{D.9}$$

Proof. Step 1: Root-n rate for $\check{\beta} - \beta$.

Since $\check{c} \in [J]$, (B.2) of Lemma B.1 gives

$$|\sqrt{n}(\check{\beta} - \beta)| \le \max_{c\in[J]} |\sqrt{n}(\widehat{\beta}_c - \beta)| = O_p(1).$$

Thus (D.8) holds.

Step 2: $n^{-1/2}$ rates for the empirical second moments.

Under Assumption 1(i), Gaussian fourth moments give $\mathrm{Var}(g_i) \le C(1 + \Pi_i^2)$ for $g_i \in \{\varepsilon_i^2, \varepsilon_i x_i, x_i^2\}$, with C depending only on Ω. Independence across i and $\mathbb{E}_n[\Pi_i^2] = H \le \overline{H}$ from Assumption 1(ii) imply $\mathrm{Var}(\mathbb{E}_n[g_i]) \le C/n$. Chebyshev's inequality yields

$$\begin{aligned}
\mathbb{E}_n[\varepsilon_i^2] &= \sigma_\varepsilon^2 + O_p(n^{-1/2}), \\
\mathbb{E}_n[\varepsilon_i x_i] &= \sigma_{\varepsilon v} + O_p(n^{-1/2}), \\
\mathbb{E}_n[x_i^2] &= H + \sigma_v^2 + O_p(n^{-1/2}).
\end{aligned}$$

Step 3: $n^{-1/2}$ rates for $\widehat{\sigma}_\varepsilon^2$ and $\widehat{\sigma}_{\varepsilon v}$.

Finally, set $\delta_\beta := \check{\beta} - \beta = O_p(n^{-1/2})$ by (D.8). Substituting the three rates above gives

$$\begin{aligned}
\widehat{\sigma}^2_\varepsilon &= \mathbb{E}_n[\varepsilon_i^2] - 2\delta_\beta \mathbb{E}_n[\varepsilon_i x_i] + \delta_\beta^2 \mathbb{E}_n[x_i^2] \\
&= \sigma^2_\varepsilon + O_p(n^{-1/2}), \\
\widehat{\sigma}_{\varepsilon v} &= \mathbb{E}_n[\varepsilon_i x_i] - \delta_\beta \mathbb{E}_n[x_i^2] \\
&= \sigma_{\varepsilon v} + O_p(n^{-1/2}).
\end{aligned}$$

□

E LASSO Fitted-Value Geometry

Fix a candidate and suppress its subscript. Write z_i' for row i of Z, let $u(w) := w - \mu(w)$ denote the LASSO residual, and let $E(w)$ be the equicorrelation set of (9). Since the fitted vector $\mu(w)$ is unique, neither $u(w)$ nor $E(w)$ depends on which minimizer of (5) is chosen. Define the residual polyhedron

$$\mathcal{K} := \left\{ q \in \mathbb{R}^n : \|\mathbb{E}_n[z_i q_i]\|_\infty \le \lambda \right\}.$$

The set $\mathcal{K}$ is a nonempty closed convex polyhedron because $0 \in \mathcal{K}$ and it is the intersection of finitely many closed linear half-spaces. Thus its Euclidean projection is well defined and unique. Throughout this section $\widehat{\pi}(w)$ denotes an arbitrary minimizer of the problem in (5); Lemma E.2 fixes a particular measurable choice.

Lemma E.1 (LASSO Projection Geometry)**.** *For every $w \in \mathbb{R}^n$ and every minimizer $\widehat{\pi}(w)$ of the problem in* (5)*,*

$$|\mathbb{E}_n[z_{ij} u_i(w)]| \le \lambda, \qquad \widehat{\pi}_j(w) \mathbb{E}_n[z_{ij} u_i(w)] = \lambda |\widehat{\pi}_j(w)|, \quad j \in [p], \tag{E.1}$$

and hence $\operatorname{supp}\{\widehat{\pi}(w)\} \subseteq E(w)$ *and*

$$\mathbb{E}_n[\mu_i(w) w_i] = \mathbb{E}_n[\mu_i(w)^2] + \lambda \|\widehat{\pi}(w)\|_1. \tag{E.2}$$

Moreover,

$$u(w) = P_{\mathcal{K}}(w), \qquad \mu(w) = w - P_{\mathcal{K}}(w), \tag{E.3}$$

where $P_{\mathcal{K}}$ is Euclidean projection onto $\mathcal{K}$. The map μ is one-Lipschitz, $\mu(0) = 0$, and

$$\|\mu(w)\|_2 \le \|w\|_2. \tag{E.4}$$

For Lebesgue-almost every $w \in \mathbb{R}^n$,

$$\dot{\mu}(w) = P_{\operatorname{col}(Z_{E(w)})}, \qquad \dot{\mu}(w)' = \dot{\mu}(w) = \dot{\mu}(w)^2, \qquad \dot{\mu}(w)\mu(w) = \mu(w). \tag{E.5}$$

Proof. Begin with the KKT conditions for the LASSO problem defined by (5):

$$\mathbb{E}_n[z_i u_i(w)] = \lambda s, \qquad s_j \in \begin{cases} \{\operatorname{sign}\{\widehat{\pi}_j(w)\}\}, & \widehat{\pi}_j(w) \ne 0, \\ [-1, 1], & \widehat{\pi}_j(w) = 0. \end{cases}$$

Hence (E.1) holds and $\operatorname{supp}\{\widehat{\pi}(w)\} \subseteq E(w)$. Summing the second equality in (E.1) over $j \in [p]$ and using $\mu(w) = Z\widehat{\pi}(w)$ and $u(w) = w - \mu(w)$ gives (E.2). Multiplying the objective in (5) by n puts it in the standard form,

$$n\left\{\frac{1}{2}\mathbb{E}_n[(w_i - z_i'\pi)^2] + \lambda\|\pi\|_1\right\} = \frac{1}{2}\|w - Z\pi\|_2^2 + n\lambda\|\pi\|_1.$$

Lemma 3 of Tibshirani and Taylor (2012), applied with $X = Z$ and penalty $n\lambda$, gives (E.3). In view of Lemma 1 of the same paper and (E.3),

$$\|\mu(w) - \mu(\widetilde{w})\|_2 \leq \|w - \widetilde{w}\|_2.$$

At $w = 0$ the objective in (5) is $\frac{1}{2}\mathbb{E}_n[(z_i'\pi)^2] + \lambda\|\pi\|_1 \geq 0$, and it vanishes at $\pi = 0$, so $\pi = 0$ is a minimizer. Therefore,

$$\mu(0) = 0, \qquad \|\mu(w)\|_2 = \|\mu(w) - \mu(0)\|_2 \leq \|w\|_2,$$

which proves (E.4).

The derivative calculation in the proof of Theorem 1 of Tibshirani and Taylor (2012) yields $\dot{\mu}(w) = P_{\mathrm{col}(Z_{E(w)})}$ for Lebesgue-almost every w; orthogonal projections are symmetric and idempotent, so $\dot{\mu}(w)' = \dot{\mu}(w) = \dot{\mu}(w)^2$. Since $\mathrm{supp}\{\widehat{\pi}(w)\} \subseteq E(w)$ by (E.1), $\mu(w) = Z_{E(w)}\widehat{\pi}_{E(w)}(w) \in \mathrm{col}(Z_{E(w)})$, and hence

$$\dot{\mu}(w)\mu(w) = P_{\mathrm{col}(Z_{E(w)})}\mu(w) = \mu(w),$$

which proves (E.5). □

Lemma E.2 (Measurable Full-Rank LASSO Solution)**.** *There exists a Borel-measurable LASSO solution $w \mapsto \widehat{\pi}(w)$ such that, for every $w \in \mathbb{R}^n$,*

$$\mathrm{rank}\Big(Z_{\mathrm{supp}\{\widehat{\pi}(w)\}}\Big) = \big|\mathrm{supp}\{\widehat{\pi}(w)\}\big| \leq \mathrm{rank}(Z) \leq n, \tag{E.6}$$

and $Z\widehat{\pi}(w)$ equals the unique fitted vector $\mu(w)$.

Proof. Step 1: Support reduction.

The objective in (5) is convex and coercive since $\lambda > 0$, so the minimum is attained. Fix w, let b be a LASSO minimizer for response w, define $B := \mathrm{supp}(b)$, and suppose that the columns of Z_B are linearly dependent. Choose $0 \neq \delta \in \mathbb{R}^{|B|}$ such that $Z_B\delta = 0$. Since b is a minimizer, Lemma E.1 gives $\mathbb{E}_n[z_{ij}u_i(w)] = \lambda\,\mathrm{sign}(b_j)$ for $j \in B$, so

$$\mathrm{sign}(b_B)'\delta = \frac{u(w)'Z_B\delta}{n\lambda} = 0.$$

For $b_B(t) := b_B + t\delta$, define

$$\mathcal{I} := \{t : \mathrm{sign}(b_j)(b_j + t\delta_j) \geq 0,\ j \in B\}.$$

The constraints in the display above are linear in t, so $\mathcal{I}$ is a closed interval, and each holds strictly at $t = 0$ since $b_j \neq 0$ for $j \in B$. Hence 0 is interior to $\mathcal{I}$. For every $t \in \mathcal{I}$, the fitted values are unchanged because $Z_Bb_B(t) = Z_Bb_B + tZ_B\delta = Z_Bb_B$, while $|b_j + t\delta_j| = \mathrm{sign}(b_j)(b_j + t\delta_j)$ for $j \in B$, so that

$$\|b_B(t)\|_1 = \|b_B\|_1 + t\,\mathrm{sign}(b_B)'\delta = \|b_B\|_1.$$

Since $\delta \neq 0$ there is a $j \in B$ with $\delta_j \neq 0$, and the constraint for that index bounds $\mathcal{I}$ on one side. At a finite endpoint t_* of $\mathcal{I}$ at least one constraint holds with equality, so $b_k + t_*\delta_k = 0$ for the corresponding index

$k \in B$. Since $\mathcal{I}$ is closed, $t_* \in \mathcal{I}$, so neither the fit nor $\|b_B(t_*)\|_1$ has changed, and extending $b_B(t_*)$ by zeros off B yields a LASSO minimizer with strictly smaller support. Repeating this reduction at most $|\operatorname{supp}(b)|$ times yields a LASSO minimizer whose support indexes linearly independent columns of Z.

Step 2: Borel selection.

Define

$$\mathcal{B} := \{B \subseteq \{1, \ldots, p\} : \operatorname{rank}(Z_B) = |B|\}.$$

For nonempty $B \in \mathcal{B}$, define

$$b_B^{(B)}(w) := (Z_B' Z_B)^{-1} Z_B' \mu(w), \qquad b_{B^c}^{(B)}(w) := 0,$$

and put $b^{(\varnothing)}(w) := 0$. Define

$$\mathcal{R}_B := \left\{ w : \begin{array}{l} |\mathbb{E}_n[z_{ij}\{w_i - z_i' b^{(B)}(w)\}]| \leq \lambda, \quad j = 1, \ldots, p, \\ b_j^{(B)}(w)\mathbb{E}_n[z_{ij}\{w_i - z_i' b^{(B)}(w)\}] = \lambda |b_j^{(B)}(w)|, \quad j = 1, \ldots, p \end{array} \right\}.$$

The conditions defining $\mathcal{R}_B$ are (E.1) evaluated at $b^{(B)}(w)$, and these are sufficient for a minimum of the convex objective in (5). Hence $b^{(B)}(w)$ solves (5) for every $w \in \mathcal{R}_B$, and since the fitted vector is unique, $Zb^{(B)}(w) = \mu(w)$. The map μ is one-Lipschitz by Lemma E.1, so $b^{(B)}$ is continuous; each $\mathcal{R}_B$ is then the preimage of a closed set under a continuous map, hence closed. By Step 1, for each w there is a LASSO minimizer b for which $Z_{\operatorname{supp}(b)}$ has full column rank. Set $B := \operatorname{supp}(b)$. If $B = \varnothing$, then $\mu(w) = Zb = 0$, so $b^{(\varnothing)}(w) = 0 = b$ and $w \in \mathcal{R}_\varnothing$ by (E.1). Otherwise, since $Z_B b_B = Zb = \mu(w)$ and Z_B has full column rank,

$$b_B = (Z_B' Z_B)^{-1} Z_B' \mu(w) = b_B^{(B)}(w),$$

so $b^{(B)}(w) = b$ satisfies the conditions defining $\mathcal{R}_B$ by Lemma E.1, and $w \in \mathcal{R}_B$. Hence $\bigcup_{B \in \mathcal{B}} \mathcal{R}_B = \mathbb{R}^n$.

Enumerate the elements of $\mathcal{B}$ in lexicographic order as $B_1, \ldots, B_L$, and define

$$\ell(w) := \min\{\ell : w \in \mathcal{R}_{B_\ell}\}, \qquad \widehat{\pi}(w) := b^{(B_{\ell(w)})}(w).$$

For each ℓ,

$$\{w : \ell(w) = \ell\} = \mathcal{R}_{B_\ell} \setminus \bigcup_{j < \ell} \mathcal{R}_{B_j}$$

is Borel. Thus $\widehat{\pi}$ is a Borel-measurable LASSO solution. Moreover $\operatorname{supp}\{\widehat{\pi}(w)\} \subseteq B_{\ell(w)} \in \mathcal{B}$, so

$$\operatorname{rank}(Z_{\operatorname{supp}\{\widehat{\pi}(w)\}}) = |\operatorname{supp}\{\widehat{\pi}(w)\}| \leq |B_{\ell(w)}| = \operatorname{rank}(Z_{B_{\ell(w)}}) \leq \operatorname{rank}(Z) \leq n.$$

This is (E.6). Uniqueness of the fitted vector gives $Z\widehat{\pi}(w) = \mu(w)$. $\square$

F Gaussian Identities and Rank Variance

By Rademacher's theorem, any locally Lipschitz $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable almost everywhere. Define $\nabla f(w)$ to be its gradient where it is differentiable and zero otherwise, and set $\partial_i f(w) := \{\nabla f(w)\}_i$.

F.1 Gaussian Inequalities

Lemma F.1 (Gaussian Poincaré Inequality)**.** *Let $W \sim N(0, \sigma^2 I_n)$ with $\sigma > 0$. If $g : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz and*

$$\mathbb{E}[g(W)^2] < \infty, \qquad \mathbb{E}[\|\nabla g(W)\|_2^2] < \infty,$$

then

$$\mathrm{Var}\{g(W)\} \leq \sigma^2 \mathbb{E}[\|\nabla g(W)\|_2^2]. \tag{F.1}$$

Proof. Write $W = \sigma G$ with $G \sim N(0, I_n)$, and set $\widetilde{g}(a) := g(\sigma a)$, so that $\nabla \widetilde{g}(a) = \sigma \nabla g(\sigma a)$ almost everywhere. Then $\widetilde{g}$ is locally Lipschitz with $\mathbb{E}[\widetilde{g}(G)^2] < \infty$ and $\mathbb{E}[\|\nabla \widetilde{g}(G)\|_2^2] < \infty$, so $\widetilde{g} \in W^{2,1}(\gamma_n)$ for $\gamma_n := N(0, I_n)$ by Proposition 1.5.2 of Bogachev (1998), with Sobolev derivative equal almost everywhere to $\nabla \widetilde{g}$. Theorem 1.6.4 of Bogachev (1998) then gives

$$\mathrm{Var}\{g(\sigma G)\} = \mathrm{Var}\{\widetilde{g}(G)\} \leq \mathbb{E}[\|\nabla \widetilde{g}(G)\|_2^2] = \sigma^2 \mathbb{E}[\|\nabla g(\sigma G)\|_2^2],$$

which is (F.1) since $W = \sigma G$. □

Lemma F.2 (Gaussian Tail Bound)**.** *Let $G \sim N(0, 1)$ and define $\Phi(u) := \mathbb{P}(G \leq u)$. For $u \geq 0$,*

$$\Phi(-u) = \mathbb{P}(G \geq u) \leq \exp(-u^2/2), \qquad \mathbb{P}(|G| > u) \leq 2\exp(-u^2/2). \tag{F.2}$$

Proof. For $a \geq 0$, Markov's inequality and the standard-normal moment generating function give

$$\mathbb{P}(G \geq u) \leq \exp(-au)\mathbb{E}[\exp(aG)] = \exp(a^2/2 - au).$$

Setting $a = u$ gives the first bound in (F.2); the second follows by symmetry of G. □

F.2 Stein Identities

Lemma F.3 (Stein Identities)**.** *Let $W \sim N(0, \sigma^2 I_n)$ with $\sigma > 0$, and let m be a Lipschitz map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Set $A(w) := Dm(w)$ where m is differentiable and $A(w) := 0$ otherwise, and define $\operatorname{div} m(w) := \operatorname{tr}\{A(w)\}$. Then*

$$\mathbb{E}[W' m(W)] = \sigma^2 \mathbb{E}[\operatorname{div} m(W)], \tag{F.3}$$

$$\mathbb{E}[\{W' m(W) - \sigma^2 \operatorname{div} m(W)\}^2] = \sigma^2 \mathbb{E}[\|m(W)\|_2^2] + \sigma^4 \mathbb{E}[\operatorname{tr}\{A(W)^2\}]. \tag{F.4}$$

If, in addition, $g : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz and there are constants $C < \infty$ and $r \geq 0$ such that

$$|g(w)| + \|\nabla g(w)\|_2 \leq C(1 + \|w\|_2^r) \quad \textit{almost everywhere},$$

then the covariance identity is

$$\mathbb{E}[\{W' m(W) - \sigma^2 \operatorname{div} m(W)\} g(W)] = \sigma^2 \mathbb{E}[m(W)' \nabla g(W)]. \tag{F.5}$$

The identities (F.3)–(F.5) follow from the Gaussian integration-by-parts formula of Stein (1981) for the first-order and covariance identities, and from Bellec and Zhang (2021) for the second-order identity.

Proof. Let L be the Lipschitz constant of m. By Lipschitz continuity and Rademacher's theorem,

$$\|m(w)\|_2 \le \|m(0)\|_2 + L\|w\|_2, \qquad \|A(w)\|_{\mathrm{op}} \le L \quad \text{almost everywhere.} \tag{F.6}$$

Since W has a Lebesgue density, the nondifferentiability sets in Rademacher's theorem are $\mathbb{P}$-null. All Gaussian polynomial moments are finite.

Step 1: Coordinate identity.

We show that for any locally Lipschitz $a : \mathbb{R}^n \to \mathbb{R}$ satisfying

$$|a(w)| + |\partial_i a(w)| \le C(1 + \|w\|_2^r) \quad \text{almost everywhere}$$

for some $C < \infty$ and $r \ge 0$,

$$\mathbb{E}[W_i a(W)] = \sigma^2 \mathbb{E}[\partial_i a(W)]. \tag{F.7}$$

Write $w = (w_i, w_{-i})$, and let φ_σ be the $N(0, \sigma^2)$ density. By Rademacher's theorem and Fubini's theorem, for almost every w_{-i}, the section $h(t) := a(t, w_{-i})$ satisfies $h'(t) = \partial_i a(t, w_{-i})$ and the polynomial bounds almost everywhere. Since h is Lipschitz on compact intervals, it is absolutely continuous there by Theorem 3.3 of Heinonen (2005). The bound on $|h(t)|$ extends to every t by continuity, so $h(t)\varphi_\sigma(t) \to 0$ as $t \to \pm\infty$. Integration by parts, using $\varphi'_\sigma(t) = -t\varphi_\sigma(t)/\sigma^2$, gives

$$\begin{aligned}\int_{-\infty}^{\infty} t h(t)\varphi_\sigma(t)\, dt &= -\sigma^2 \int_{-\infty}^{\infty} h(t)\varphi'_\sigma(t)\, dt \\ &= \sigma^2 \int_{-\infty}^{\infty} \partial_i a(t, w_{-i})\varphi_\sigma(t)\, dt.\end{aligned}$$

The polynomial bounds ensure absolute convergence. Integrating over w_{-i} against the law of W_{-i}, independence and Fubini's theorem give (F.7).

Step 2: First-order and covariance identities.

For (F.3), take $a = m_i$ in (F.7). Its polynomial bounds follow from (F.6). Thus

$$\mathbb{E}[W' m(W)] = \sum_{i=1}^{n} \mathbb{E}[W_i m_i(W)] = \sigma^2 \sum_{i=1}^{n} \mathbb{E}[\partial_i m_i(W)] = \sigma^2 \mathbb{E}[\operatorname{div} m(W)].$$

Take $a = m_i g$ instead. By the product rule for locally Lipschitz functions, $\partial_i(m_i g) = (\partial_i m_i)g + m_i \partial_i g$ almost everywhere, while (F.6) with the polynomial bound on g controls $|m_i g|$ and $|\partial_i(m_i g)|$ by polynomials in $\|W\|_2$, which are Gaussian integrable. Summing (F.7) over i therefore gives

$$\mathbb{E}[W' m(W) g(W)] = \sigma^2 \mathbb{E}[\{\operatorname{div} m(W)\} g(W)] + \sigma^2 \mathbb{E}[m(W)' \nabla g(W)].$$

Rearranging the above display gives (F.5).

Step 3: Second-order identity.

Let $G \sim N(0, I_n)$, $W = \sigma G$, and $\widetilde{m}(u) := \sigma m(\sigma u)$. Then $\widetilde{m}$ is Lipschitz with constant $\sigma^2 L$, and by the chain rule $D\widetilde{m}(u) = \sigma^2 A(\sigma u)$ at its differentiability points. Hence

$$G'\widetilde{m}(G) - \operatorname{div}\widetilde{m}(G) = W'm(W) - \sigma^2 \operatorname{div} m(W), \qquad D\widetilde{m}(G) = \sigma^2 A(W) \quad \text{almost surely.}$$

Theorem 2.1(ii) of Bellec and Zhang (2021) gives, for a Lipschitz $f : \mathbb{R}^n \to \mathbb{R}^n$ with $\mathbb{E}[\|f(G)\|_2^2] + \mathbb{E}[\|Df(G)\|_F^2] < \infty$,

$$\mathbb{E}[\{G'f(G) - \operatorname{div} f(G)\}^2] = \mathbb{E}[\|f(G)\|_2^2] + \mathbb{E}[\operatorname{tr}\{Df(G)^2\}].$$

Both hypotheses hold at $f = \widetilde{m}$: since $\widetilde{m}(G) = \sigma m(W)$ and $D\widetilde{m}(G) = \sigma^2 A(W)$, it is enough to control the unscaled moments, and by (F.6), $(a+b)^2 \le 2a^2 + 2b^2$, and $\mathbb{E}[\|W\|_2^2] = n\sigma^2$,

$$\mathbb{E}[\|m(W)\|_2^2] \le 2\|m(0)\|_2^2 + 2L^2 n\sigma^2 < \infty, \qquad \|A(W)\|_F^2 \le n\|A(W)\|_{\mathrm{op}}^2 \le nL^2.$$

The trace term $\mathbb{E}[\operatorname{tr}\{Df(G)^2\}]$ in the Bellec–Zhang display above is finite as well, since Cauchy–Schwarz for the Frobenius inner product and $\|A'\|_F = \|A\|_F$ give $|\operatorname{tr}(A^2)| = |\langle A', A\rangle_F| \le \|A\|_F^2$. Hence

$$\begin{aligned}\mathbb{E}[\{G'\widetilde{m}(G) - \operatorname{div}\widetilde{m}(G)\}^2] &= \mathbb{E}[\|\widetilde{m}(G)\|_2^2] + \mathbb{E}[\operatorname{tr}\{D\widetilde{m}(G)^2\}] \\ &= \sigma^2\mathbb{E}[\|m(W)\|_2^2] + \sigma^4\mathbb{E}[\operatorname{tr}\{A(W)^2\}],\end{aligned}$$

which is (F.4), since $G'\widetilde{m}(G) - \operatorname{div}\widetilde{m}(G) = W'm(W) - \sigma^2 \operatorname{div} m(W)$ almost surely. □

F.3 Oracle Baseline

Lemma F.4 (Oracle Baseline)**.** *Under Assumption 1, with L_n, q, and C_{0n} defined by (A.8), (A.11), and (A.13),*

$$\mathbb{E}[L_n^2] = \frac{\sigma_\varepsilon^2}{H}, \qquad \mathbb{E}[qL_n^2] = 0, \qquad \mathbb{E}[q^2 L_n^2] = \frac{\sigma_\varepsilon^2\sigma_v^2 + 2\sigma_{\varepsilon v}^2}{nH^2}, \tag{F.8}$$

and

$$\mathbb{E}[\{(1-q)L_n\}^2] = C_{0n}. \tag{F.9}$$

Proof. Under Assumption 1(i), the pair $(\Pi'\varepsilon, \Pi'v)$ is centered Gaussian. Since $\|\Pi\|_2^2 = nH$,

$$\mathbb{E}[(\Pi'\varepsilon)^2] = \sigma_\varepsilon^2 nH, \qquad \mathbb{E}[(\Pi'v)^2] = \sigma_v^2 nH, \qquad \mathbb{E}[(\Pi'\varepsilon)(\Pi'v)] = \sigma_{\varepsilon v} nH. \tag{F.10}$$

By (A.8) and (A.11), $L_n = \Pi'\varepsilon/(\sqrt{n}\,H)$ and $q = \Pi'v/(nH)$, so the first identity in (F.10) yields $\mathbb{E}[L_n^2] = \sigma_\varepsilon^2/H$. Next, qL_n^2 is a cubic form in the centered Gaussian pair $(\Pi'\varepsilon, \Pi'v)$, so $\mathbb{E}[qL_n^2] = 0$ because odd moments of a centered Gaussian vector vanish. Finally, for a centered Gaussian pair (X, Y), $\mathbb{E}[X^2Y^2] = \mathbb{E}[X^2]\mathbb{E}[Y^2] + 2\mathbb{E}[XY]^2$ (Isserlis, 1918). Applying this identity at $X = \Pi'v$ and $Y = \Pi'\varepsilon$ and substituting (F.10),

$$\mathbb{E}[(\Pi'v)^2(\Pi'\varepsilon)^2] = n^2H^2\left\{\sigma_\varepsilon^2\sigma_v^2 + 2\sigma_{\varepsilon v}^2\right\}.$$

Divide by n^3H^4 to obtain the third identity in (F.8). For (F.9), expand $\{(1-q)L_n\}^2 = L_n^2 - 2qL_n^2 + q^2L_n^2$ and apply (F.8) to get $\mathbb{E}[\{(1-q)L_n\}^2] = \sigma_\varepsilon^2/H + (\sigma_\varepsilon^2\sigma_v^2 + 2\sigma_{\varepsilon v}^2)/(nH^2)$, which is C_{0n} by (A.13). □

F.4 LASSO Fit Moments

Lemma F.5 (Offset-Fit Moments)**.** *Suppose Assumption 1(i) holds. Fix a candidate, suppress its subscript, and let $b \in \mathbb{R}^n$ be nonrandom. Define $m(a) := \mu(\Pi + a) - b$ for $a \in \mathbb{R}^n$ and $M := m(v) = \widehat{\Pi} - b$. Let d be as in (10) and $\bar{d} := \mathbb{E}[d]$ as in (A.2). Then m is one-Lipschitz, $\mathbb{E}[\|M\|_2^r] < \infty$ for every fixed $r > 0$, and, for all nonrandom $g, c_0 \in \mathbb{R}^n$,*

$$\operatorname{Var}(g'M) \leq \sigma_v^2 \|g\|_2^2, \tag{F.11}$$

$$\operatorname{Var}(\|M\|_2^2) \leq 4\sigma_v^2 \mathbb{E}[\|M\|_2^2], \tag{F.12}$$

$$\operatorname{Var}\{(\Pi + v + c_0)'M\} \leq 2\sigma_v^2 \left\{ \mathbb{E}[\|M\|_2^2] + \mathbb{E}[\|\Pi + v + c_0\|_2^2] \right\}, \tag{F.13}$$

$$\mathbb{E}[v'M] = \sigma_v^2 \bar{d},$$

$$\mathbb{E}[(v'M - \sigma_v^2 d)^2] = \sigma_v^2 \mathbb{E}[\|M\|_2^2] + \sigma_v^4 \bar{d}. \tag{F.14}$$

Proof. The map μ is one-Lipschitz by Lemma E.1 and b is nonrandom, so m is one-Lipschitz. That lemma also gives $\|\widehat{\Pi}\|_2 \leq \|\Pi + v\|_2$, so $\|M\|_2 \leq \|\Pi + v\|_2 + \|b\|_2$, and $\mathbb{E}[\|M\|_2^r] < \infty$ for every fixed $r > 0$ by finiteness of Gaussian polynomial moments. At almost every a, $Dm(a) = \dot{\mu}(\Pi + a)$ is an orthogonal projection by Lemma E.1, hence a contraction. Write $P := Dm(v)$; by (A.1), with the candidate subscript suppressed, $P = P' = P^2$ and $\operatorname{tr}(P) = d$ almost surely, hence also $\operatorname{tr}(P^2) = d$.

Step 1: Variance bounds.

Each map below is locally Lipschitz, and its square and squared gradient are bounded by polynomials in $\|a\|_2$ with nonrandom coefficients; since Gaussian polynomial moments are finite, Lemma F.1 applies with $\sigma^2 = \sigma_v^2$. All gradients below are taken at almost every a, without further qualification. First, for nonrandom g, the map $a \mapsto g'm(a)$ has gradient $Dm(a)'g$, of norm at most $\|g\|_2$ by contraction, so

$$\operatorname{Var}(g'M) \leq \sigma_v^2 \mathbb{E}[\|Dm(v)'g\|_2^2] \leq \sigma_v^2 \|g\|_2^2.$$

Next, $a \mapsto \|m(a)\|_2^2$ has gradient $2Dm(a)'m(a)$, of norm at most $2\|m(a)\|_2$, whence

$$\operatorname{Var}(\|M\|_2^2) \leq 4\sigma_v^2 \mathbb{E}[\|M\|_2^2].$$

Finally, the map $a \mapsto (\Pi + a + c_0)'m(a)$ has gradient $m(a) + Dm(a)'(\Pi + a + c_0)$, of squared norm at most $2\|m(a)\|_2^2 + 2\|\Pi + a + c_0\|_2^2$ by $\|x + y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$ and contraction of $Dm(a)$. Hence

$$\operatorname{Var}\{(\Pi + v + c_0)'M\} \leq 2\sigma_v^2 \left\{ \mathbb{E}[\|M\|_2^2] + \mathbb{E}[\|\Pi + v + c_0\|_2^2] \right\}.$$

Step 2: Stein identities.

Since $\operatorname{div} m(v) = \operatorname{tr}(P) = d$ and $\operatorname{tr}(P^2) = d$ almost surely, (F.3) and (F.4) with $\sigma^2 = \sigma_v^2$ give

$$\mathbb{E}[v'M] = \sigma_v^2 \bar{d},$$

$$\mathbb{E}[(v'M - \sigma_v^2 d)^2] = \sigma_v^2 \mathbb{E}[\|M\|_2^2] + \sigma_v^4 \bar{d}.$$

□

Lemma F.6 (Gaussian LASSO Moments)**.** *Under Assumption 1(i), fix a candidate and suppress its subscript. Then*

$$\mathbb{E}[(\varepsilon'\widehat{\Pi})^2] = \sigma_\varepsilon^2 \mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_{\varepsilon v}^2 \mathbb{E}[d^2 + d]. \tag{F.15}$$

Define

$$\zeta := v'\widehat{\Pi} - \sigma_v^2 d. \tag{F.16}$$

Then, with ζ so defined,[8]

$$\mathbb{E}[v'\widehat{\Pi}] = \sigma_v^2 \mathbb{E}[d], \tag{F.17}$$

$$\mathbb{E}[\zeta^2] = \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_v^4 \mathbb{E}[d], \tag{F.18}$$

$$\mathbb{E}[d\zeta] = 0. \tag{F.19}$$

Proof. Lemma F.5 at $b = 0$ gives $\mathbb{E}[\|\widehat{\Pi}\|_2^r] < \infty$ for every fixed $r > 0$, and (F.14) gives (F.17) and (F.18). Write $m(a) := \mu(\Pi + a)$ and $P := Dm(v) = \dot{\mu}(\Pi + v)$. The map m is one-Lipschitz by that lemma, while (A.1) gives $P = P' = P^2$, $\mathrm{tr}(P) = d$, and $P\widehat{\Pi} = \widehat{\Pi}$ almost surely; also $d \le \mathrm{rank}(Z) \le n$.

Step 1: $\mathbb{E}[d\zeta] = 0$.

Define $g(a) := a'm(a)$. Since m is Lipschitz, g is locally Lipschitz with $\nabla g(a) = m(a) + Dm(a)'a$ almost everywhere, and since $\|m(a)\|_2 \le \|\Pi\|_2 + \|a\|_2$ while $Dm(a)$ is an orthogonal projection by Lemma E.1, Cauchy–Schwarz and the triangle inequality give

$$\begin{aligned}
|g(a)| + \|\nabla g(a)\|_2 &\le \|a\|_2 \|m(a)\|_2 + \|m(a)\|_2 + \|Dm(a)'a\|_2 \\
&\le \|a\|_2(\|\Pi\|_2 + \|a\|_2) + \|\Pi\|_2 + 2\|a\|_2 \\
&\le C_\Pi(1 + \|a\|_2^2)
\end{aligned}$$

for a finite constant C_Π depending only on $\|\Pi\|_2$, verifying the growth condition of Lemma F.3 at $r = 2$. At $a = v$, (A.1) gives $P = P'$ and $P\widehat{\Pi} = \widehat{\Pi}$, so almost surely $\nabla g(v) = \widehat{\Pi} + Pv$ and $\widehat{\Pi}'Pv = (P\widehat{\Pi})'v = \widehat{\Pi}'v$. By (F.5) together with (F.17)–(F.18),

$$\begin{aligned}
\sigma_v^2 \mathbb{E}[d\zeta] &= \mathbb{E}[(v'\widehat{\Pi})\zeta] - \mathbb{E}[\zeta^2] \\
&= \sigma_v^2 \mathbb{E}[\widehat{\Pi}'(\widehat{\Pi} + Pv)] - \mathbb{E}[\zeta^2] \\
&= \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_v^4 \mathbb{E}[d] - \mathbb{E}[\zeta^2] = 0.
\end{aligned}$$

Step 2: Second moment of $\varepsilon'\widehat{\Pi}$.

Under Assumption 1(i), the conditional law of ε given v is

$$\varepsilon \mid v \sim N\left\{ \frac{\sigma_{\varepsilon v}}{\sigma_v^2} v, \left(\sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2} \right) I_n \right\},$$

[8] All quantities in (F.15)–(F.19) depend on the LASSO only through $\widehat{\Pi}$ and d, so no selection among non-unique coefficient solutions is involved.

and both $\widehat{\Pi}$ and d are $\sigma(v)$-measurable. Hence

$$\mathbb{E}[(\varepsilon'\widehat{\Pi})^2 \mid v] = \widehat{\Pi}'\mathbb{E}[\varepsilon\varepsilon' \mid v]\widehat{\Pi} = \frac{\sigma_{\varepsilon v}^2}{\sigma_v^4}(v'\widehat{\Pi})^2 + \left(\sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2}\right)\|\widehat{\Pi}\|_2^2.$$

Moreover,

$$\begin{aligned}\mathbb{E}[(v'\widehat{\Pi})^2] &= \mathbb{E}[(\zeta + \sigma_v^2 d)^2] \\ &= \mathbb{E}[\zeta^2] + 2\sigma_v^2\mathbb{E}[d\zeta] + \sigma_v^4\mathbb{E}[d^2] \\ &= \sigma_v^2\mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_v^4\mathbb{E}[d^2 + d],\end{aligned}$$

where (F.18) and (F.19) respectively evaluate $\mathbb{E}[\zeta^2]$ and remove the cross term. Taking expectations in the conditional second moment and substituting the above display,

$$\begin{aligned}\mathbb{E}[(\varepsilon'\widehat{\Pi})^2] &= \frac{\sigma_{\varepsilon v}^2}{\sigma_v^4}\mathbb{E}[(v'\widehat{\Pi})^2] + \left(\sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2}\right)\mathbb{E}[\|\widehat{\Pi}\|_2^2] \\ &= \sigma_\varepsilon^2\mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_{\varepsilon v}^2\mathbb{E}[d^2 + d].\end{aligned}$$ □

F.5 Variance of the Effective Dimension

The rank bound below follows from the variance inequality of Theorem 2.2 of Bellec and Zhang (2021), applied to the LASSO fitted-value map.

Lemma F.7 (Effective-Dimension Variance Bound)**.** *Fix Π, $Z \in \mathbb{R}^{n\times p}$, and $\lambda > 0$, and let μ denote the LASSO fitted-value map defined by (5). For $v \sim N(0, \sigma_v^2 I_n)$ with $\sigma_v^2 > 0$, define*

$$d := \operatorname{rank}(Z_{E(\Pi+v)}), \qquad \bar{d} := \mathbb{E}[d].$$

Then $d = \operatorname{div}\mu(\Pi + v)$ $\mathbb{P}$-almost surely. There is a universal constant $C < \infty$ such that

$$\operatorname{Var}(d) \le C\left[1 + \bar{d}\log\left\{\frac{ep}{\bar{d}\vee 1}\right\}\right]. \tag{F.20}$$

Proof. Set

$$G := \frac{v}{\sigma_v}, \qquad \mathcal{F}(u) := \frac{\mu(\Pi + \sigma_v u)}{\sigma_v}.$$

Then $G \sim N(0, I_n)$, and Lemma E.1 implies that $\mathcal{F}$ is 1-Lipschitz. At differentiability points the chain rule yields $D\mathcal{F}(u) = \dot{\mu}(\Pi + \sigma_v u)$, which (E.5) identifies as the orthogonal projection onto $\operatorname{col}(Z_{E(\Pi+\sigma_v u)})$ at Lebesgue-almost every u. Since G has a Lebesgue density, this identification holds almost surely, so, as in (A.1),

$$P := D\mathcal{F}(G) = P_{\operatorname{col}(Z_{E(\Pi+\sigma_v G)})}, \qquad P' = P = P^2, \qquad d = \operatorname{tr}(P).$$

In particular, $d = \operatorname{div}\mu(\Pi + v)$ almost surely, the first claim of the lemma.

Theorem 2.2 of Bellec and Zhang (2021), in its display (2.13), states that for a 1-Lipschitz f such that $g(u) := u'f(u)$ and its gradient are square integrable under $\gamma_n := N(0, I_n)$,

$$\operatorname{Var}\{\operatorname{div} f(G)\} \le \mathbb{E}[\operatorname{tr}\{Df(G)^2\}] + \mathbb{E}[\|Df(G)G\|_2^2].$$

At $f = \mathcal{F}$, the bounds

$$|u'\mathcal{F}(u)| \leq \|u\|_2\{\|\mathcal{F}(0)\|_2 + \|u\|_2\},$$
$$\big\|\nabla\{u'\mathcal{F}(u)\}\big\|_2 \leq \|\mathcal{F}(0)\|_2 + 2\|u\|_2$$

and finiteness of Gaussian polynomial moments verify these conditions; the transpose convention in their ∇f is immaterial here because $P' = P$. Since $D\mathcal{F}(G) = P$ and $\mathrm{tr}(P^2) = \mathrm{tr}(P) = d$ almost surely,

$$\mathrm{Var}(d) \leq \mathbb{E}[\mathrm{tr}(P^2)] + \mathbb{E}[\|PG\|_2^2] = \bar{d} + \mathbb{E}[\|PG\|_2^2]. \tag{F.21}$$

Step 1: Bounding the projection norm.

To bound $\mathbb{E}[\|PG\|_2^2]$ in (F.21), define, for $1 \leq r \leq \mathrm{rank}(Z)$,

$$\mathscr{P}_r := \Big\{P_{\mathrm{col}(Z_B)} : B \subseteq [p],\ |B| = \mathrm{rank}(Z_B) = r\Big\}.$$

Each element of $\mathscr{P}_r$ is determined by an r-element subset of $[p]$. On $\{d = r\}$ the span $\mathrm{col}(Z_{E(\Pi+v)})$ has rank r, so it admits an r-column basis Z_B with $B \subseteq E(\Pi + v)$. Together these give

$$|\mathscr{P}_r| \leq \binom{p}{r} \leq \left(\frac{ep}{r}\right)^r, \qquad P \in \mathscr{P}_r \quad \text{on } \{d = r\}. \tag{F.22}$$

For fixed $P_0 \in \mathscr{P}_r$, $\|P_0G\|_2^2 \sim \chi_r^2$. Display (4.3) of Laurent and Massart (2000) states that, if $U \sim \chi_r^2$, then for every $u > 0$,

$$\mathbb{P}\big\{U > r + 2\sqrt{ru} + 2u\big\} \leq e^{-u}. \tag{F.23}$$

For $L_r := \log(ep/r)$ and $t \geq 0$, a union bound over $\mathscr{P}_r$, using (F.22) and (F.23), gives

$$\mathbb{P}\bigg\{\max_{P_0 \in \mathscr{P}_r} \|P_0G\|_2^2 > 11\{rL_r + t\}\bigg\} \leq e^{-r-t}. \tag{F.24}$$

Indeed, set $u := 2rL_r + r + t$. Since $2\sqrt{ru} \leq r + u$ and $L_r \geq 1$ (as $r \leq p$), the threshold in (F.23) satisfies

$$r + 2\sqrt{ru} + 2u \leq 2r + 3u = 5r + 6rL_r + 3t \leq 11(rL_r + t),$$

so the event in (F.24) is contained in the union over $\mathscr{P}_r$ of the Laurent–Massart events, whose probability is at most

$$|\mathscr{P}_r|e^{-u} \leq e^{-rL_r - r - t} \leq e^{-r-t}.$$

Step 2: Taking expectations.

Define

$$\psi_p(x) := x \log\left(\frac{ep}{x \vee 1}\right), \qquad 0 \leq x \leq p.$$

For every $t \geq 0$, the event $\{\|PG\|_2^2 > 11\{\psi_p(d) + t\}\}$ is empty on $\{d = 0\}$, because there $P = 0$ and $\psi_p(d) = 0$. Summing (F.24) over $r \geq 1$ therefore gives

$$\mathbb{P}\Big\{\|PG\|_2^2 > 11\{\psi_p(d) + t\}\Big\} \leq \frac{e^{-t}}{e - 1}.$$

For $X := \|PG\|_2^2$, the pointwise bound $X \leq 11\psi_p(d) + \{X - 11\psi_p(d)\}_+$ and Fubini's theorem imply

$$\mathbb{E}[X] \leq 11\mathbb{E}[\psi_p(d)] + 11 \int_0^\infty \mathbb{P}\{X > 11(\psi_p(d) + t)\}\, dt,$$

so the tail bound above yields

$$\mathbb{E}[\|PG\|_2^2] \leq 11\mathbb{E}[\psi_p(d)] + \frac{11}{e-1}. \tag{F.25}$$

The function ψ_p is continuous, with

$$\psi_p'(x) = \begin{cases} \log(ep), & 0 < x < 1, \\ \log(p/x), & 1 < x < p. \end{cases}$$

Thus ψ_p is concave on $[0, p]$. Since $d \leq \operatorname{rank}(Z) \leq p$ almost surely, Jensen's inequality gives

$$\mathbb{E}[\psi_p(d)] \leq \psi_p(\bar{d}). \tag{F.26}$$

Since $\bar{d} \leq \psi_p(\bar{d})$, combining (F.21), (F.25), and (F.26) yields

$$\operatorname{Var}(d) \leq 12\psi_p(\bar{d}) + \frac{11}{e-1} \leq 12\{1 + \psi_p(\bar{d})\},$$

which is (F.20) with $C = 12$. □

G Sparse First-Stage Bounds

Following Belloni et al. (2012), define the score event

$$\mathcal{G}_c := \left\{ \left\| \mathbb{E}_n[z_{c,i} v_i] \right\|_\infty \leq \frac{\lambda_c}{c_\lambda} \right\}. \tag{G.1}$$

Under Assumptions 1 and 2, Step 1 of the proof of Lemma G.2 shows that $\mathbb{P}(\mathcal{G}_c^c) \leq \eta_{n,c} \to 0$, where $\eta_{n,c}$ is the sequence defined there.

Lemma G.1 (LASSO Bounds and Strong-Core Recovery)**.** *Under Assumption 2, for all sufficiently large n, the following hold on the event $\mathcal{G}_c$ defined by (G.1). First, the prediction norm satisfies*

$$\mathbb{E}_n[\{z_{c,i}'(\widehat{\pi}_c - \pi_c^0)\}^2]^{1/2} \leq 2\mathbb{E}_n[\xi_{c,i}^2]^{1/2} + \frac{2(1 + c_\lambda^{-1})\lambda_c \sqrt{s_c}}{\kappa_0}. \tag{G.2}$$

Next, for the measurable solution $\widehat{\pi}_c$ of Lemma E.2,

$$\|\widehat{\pi}_c - \pi_c^0\|_2 \leq \frac{2}{\sqrt{\underline{\phi}}} \left\{ \mathbb{E}_n[\xi_{c,i}^2]^{1/2} + \frac{(1 + c_\lambda^{-1})\lambda_c \sqrt{s_c}}{\kappa_0} \right\}. \tag{G.3}$$

The effective dimension d_c defined in (10) does not depend on which minimizer is taken, and satisfies

$$d_c < M s_c + 1. \tag{G.4}$$

If Assumption 3 holds as well, then, again on $\mathcal{G}_c$,

$$k_c \leq d_c < Ms_c + 1, \tag{G.5}$$

and, for the measurable solution $\widehat{\pi}_c$ above,

$$A_c \subseteq \mathrm{supp}(\widehat{\pi}_c), \qquad \mathrm{sign}(\widehat{\pi}_{c,j}) = \mathrm{sign}(\pi^0_{c,j}), \quad j \in A_c. \tag{G.6}$$

Proof. Suppress the candidate subscript c throughout. Define $\Delta := \widehat{\pi} - \pi^0$, $a := \mathbb{E}_n[\xi_i^2]^{1/2}$, and

$$t := \mathbb{E}_n[(z_i'\Delta)^2]^{1/2}, \qquad \theta := c_\lambda^{-1}, \qquad \theta_+ := 1 + \theta, \qquad \theta_- := 1 - \theta.$$

By construction, $L_\lambda = \theta_+/\theta_-$, while Assumption 2(ii)–(iii) give

$$\rho_n := \frac{a}{\lambda\sqrt{s}} \leq \frac{C_a}{c_\lambda\sqrt{2}\sigma_v\{\log(2p) + \alpha\log\log(p \vee n)\}^{1/2}} \to 0. \tag{G.7}$$

The limit holds because $\log(2p) + \alpha\log\log(p \vee n) \geq \alpha\log\log n \to \infty$.

Step 1: Prediction norm.

Substituting $x_i = z_i'\pi^0 + \xi_i + v_i$ from (1) and (11) into the objective of (5), which $\widehat{\pi}$ minimizes, gives

$$\begin{aligned}
0 &\geq \frac{1}{2}\mathbb{E}_n[(x_i - z_i'\widehat{\pi})^2 - (x_i - z_i'\pi^0)^2] + \lambda\{\|\widehat{\pi}\|_1 - \|\pi^0\|_1\} \\
&= \frac{1}{2}\mathbb{E}_n[(z_i'\Delta)^2] - \mathbb{E}_n[(\xi_i + v_i)z_i'\Delta] + \lambda\{\|\pi^0 + \Delta\|_1 - \|\pi^0\|_1\}, \\
\frac{1}{2}t^2 &\leq \mathbb{E}_n[(\xi_i + v_i)z_i'\Delta] + \lambda\{\|\pi^0\|_1 - \|\pi^0 + \Delta\|_1\}.
\end{aligned} \tag{G.8}$$

On $\mathcal{G}$, Hölder's inequality bounds the score term,

$$|\mathbb{E}_n[v_i z_i'\Delta]| \leq \theta\lambda\|\Delta\|_1,$$

while Cauchy–Schwarz bounds the approximation term,

$$|\mathbb{E}_n[\xi_i z_i'\Delta]| \leq at.$$

Because $\pi^0_{T^c} = 0$, the triangle inequality gives

$$\begin{aligned}
\|\pi^0 + \Delta\|_1 &= \|\pi^0_T + \Delta_T\|_1 + \|\Delta_{T^c}\|_1 \\
&\geq \|\pi^0_T\|_1 - \|\Delta_T\|_1 + \|\Delta_{T^c}\|_1,
\end{aligned}$$

and hence

$$\|\pi^0\|_1 - \|\pi^0 + \Delta\|_1 \leq \|\Delta_T\|_1 - \|\Delta_{T^c}\|_1.$$

Combining these bounds with (G.8) yields

$$\frac{1}{2}t^2 + \theta_-\lambda\|\Delta_{T^c}\|_1 \leq \theta_+\lambda\|\Delta_T\|_1 + at. \tag{G.9}$$

Assume $t > 0$, the bound (G.2) being trivial otherwise. If $\|\Delta_{T^c}\|_1 \le L_\lambda \|\Delta_T\|_1$, then Δ lies in the cone of (12), so $\kappa \ge \kappa_0$ from Assumption 2(iv) gives $\|\Delta_T\|_1 \le \sqrt{s}\, t/\kappa_0$; dropping the nonnegative $\theta_-\lambda\|\Delta_{T^c}\|_1$ term in (G.9) and dividing by t leaves $\frac{1}{2}t \le a + \theta_+\lambda\sqrt{s}/\kappa_0$, so that

$$t \le 2a + \frac{2\theta_+\lambda\sqrt{s}}{\kappa_0}.$$

If instead $\|\Delta_{T^c}\|_1 > L_\lambda\|\Delta_T\|_1$, then $\theta_+\|\Delta_T\|_1 - \theta_-\|\Delta_{T^c}\|_1 < 0$, so (G.9) gives $\frac{1}{2}t^2 < at$ and hence $t < 2a$. In either case t satisfies (G.2).

Step 2: Rank upper bound.

For (G.4), define $E := E(x)$, which, like the LASSO residual, does not depend on which minimizer of (5) is taken (Appendix E), so neither does $d := \operatorname{rank}(Z_E) = d_c$ of (10). Set $\bar{s} := \lceil Ms \rceil$ and $u(x) := x - Z\widehat{\pi}$. Suppose $d \ge \bar{s}$. Then $\bar{s} \le d \le p$, so the upper sparse-eigenvalue restriction in Assumption 2(iv) applies at $\bar{s}$. Choose $D \subseteq E$ with $|D| = \bar{s}$ such that Z_D has full column rank. Since $D \subseteq E$ and E is defined by (9), $\|Z_D' u(x)/n\|_2 = \lambda\sqrt{\bar{s}}$. Each coordinate of $Z_D'v/n$ is bounded in absolute value by $\|\mathbb{E}_n[z_i v_i]\|_\infty$, so on $\mathcal{G}$

$$\left\|\frac{Z_D'v}{n}\right\|_2 \le \sqrt{\bar{s}}\,\|\mathbb{E}_n[z_i v_i]\|_\infty \le \theta\lambda\sqrt{\bar{s}}.$$

Since $u(x) = \xi + v - Z\Delta$, the triangle inequality and the above display give

$$\begin{aligned}
\theta_-\lambda\sqrt{\bar{s}} &\le \left\|\frac{Z_D'\xi}{n}\right\|_2 + \left\|\frac{Z_D'Z\Delta}{n}\right\|_2 \\
&\le \left\|\frac{Z_D}{\sqrt{n}}\right\|_{\mathrm{op}} (a+t) \\
&\le \sqrt{\phi_{\max}(\bar{s})}(a+t) \\
&\le \sqrt{\overline{\phi}}\left(3\rho_n + \frac{2\theta_+}{\kappa_0}\right)\lambda\sqrt{s},
\end{aligned} \tag{G.10}$$

where we use $\|Z_D/\sqrt{n}\|_{\mathrm{op}}^2 \le \phi_{\max}(\bar{s}) \le \overline{\phi}$ from Assumption 2(iv) and (13), together with (G.2) and (G.7). Since $\rho_n \to 0$ by (G.7) and $M > 4L_\lambda^2\overline{\phi}/\kappa_0^2 = (\overline{\phi}/\theta_-^2)(2\theta_+/\kappa_0)^2$ by the constants in Assumption 2, for all sufficiently large n,

$$\bar{s} \le \frac{\overline{\phi}}{\theta_-^2}\left(3\rho_n + \frac{2\theta_+}{\kappa_0}\right)^2 s < Ms \le \bar{s},$$

a contradiction, so $d < \bar{s} = \lceil Ms \rceil < Ms + 1$.

Step 3: ℓ_2 bound.

The bounds of Steps 1 and 2 hold for every minimizer; for the remainder of the proof, take $\widehat{\pi}$ to be the measurable solution from Lemma E.2, retain $\Delta := \widehat{\pi} - \pi^0$, and define $\widehat{T} := \operatorname{supp}(\widehat{\pi})$. By Lemma E.1 and (E.6),

$$|\widehat{T}| = \operatorname{rank}(Z_{\widehat{T}}) \le \operatorname{rank}(Z_E) = d.$$

Moreover, $\operatorname{supp}(\Delta) \subseteq T \cup \widehat{T}$, so $\|\Delta\|_0 \le s + |\widehat{T}| \le s + d \le s + \lceil Ms \rceil$, since $d < \bar{s} = \lceil Ms \rceil$ by Step 2. Since also $\|\Delta\|_0 \le p$ and s is an integer,

$$\|\Delta\|_0 \le \min\{p, s + \lceil Ms \rceil\} = \min\{p, \lceil (M+1)s \rceil\}.$$

If $\Delta = 0$ the ℓ_2 bound is immediate, so take $\Delta \neq 0$. The prior display, $\phi_{\min}(\lceil (M+1)s \rceil \wedge p) \geq \underline{\phi}$ from Assumption 2(iv), and (G.2) yield

$$\sqrt{\underline{\phi}}\|\Delta\|_2 \leq t \leq 2a + \frac{2\theta_+ \lambda \sqrt{s}}{\kappa_0}.$$

Dividing by $\sqrt{\underline{\phi}}$ gives (G.3).

Step 4: Strong-core recovery and the rank lower bound.

Suppose now that Assumption 3 also holds. For (G.6), Assumption 2(ii), the definition of $\underline{b}_n$ in Assumption 3, and (G.3) give

$$|\Delta_j| \leq \|\Delta\|_2 \leq \frac{2}{\sqrt{\underline{\phi}}} \left\{ C_a \sqrt{\frac{s}{n}} + \frac{\theta_+ \lambda \sqrt{s}}{\kappa_0} \right\} = \frac{\underline{b}_n}{2}, \qquad j \in A.$$

Since $|\pi_j^0| \geq \underline{b}_n$ for $j \in A$ by Assumption 3,

$$\widehat{\pi}_j \pi_j^0 = (\pi_j^0)^2 + \Delta_j \pi_j^0 \geq |\pi_j^0| \{ |\pi_j^0| - |\Delta_j| \} \geq \frac{\underline{b}_n^2}{2} > 0, \qquad j \in A,$$

so $\widehat{\pi}_j \neq 0$ and $\operatorname{sign}(\widehat{\pi}_j) = \operatorname{sign}(\pi_j^0)$ for every $j \in A$.

Since $A \subseteq \widehat{T}$, Step 3 gives

$$k = |A| \leq |\widehat{T}| = \operatorname{rank}(Z_{\widehat{T}}) \leq \operatorname{rank}(Z_E) = d.$$

Together with (G.4), this gives (G.5). $\square$

Lemma G.2 (First-Stage Moments and Rank Ratios). *Suppose Assumptions 1 and 2 hold, with $J < \infty$ fixed. Define*

$$\eta_{n,c} := 2p_c \Phi\left(-\frac{\sqrt{n}\lambda_c}{c_\lambda \sigma_v} \right)$$

and

$$a_{n,c} := \mathbb{E}_n[\xi_{c,i}^2] + s_c \lambda_c^2 + \sqrt{\eta_{n,c}}.$$

Uniformly over candidates,

$$\eta_{n,c} \to 0, \qquad a_{n,c} \to 0,$$

and

$$\overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2] = O(a_{n,c}), \tag{G.11}$$

$$h_c - H = O(\sqrt{a_{n,c}}), \qquad \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - H = O(\sqrt{a_{n,c}}), \tag{G.12}$$

$$\mathcal{A}_c = O(a_{n,c}). \tag{G.13}$$

There are constants $\underline{h} > 0$ and $C < \infty$, independent of n and c, such that, for all sufficiently large n,

$$0 < \underline{h} \leq \min_c h_c, \qquad \max_c \{ h_c + \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] \} \leq C. \tag{G.14}$$

If Assumption 3 also holds, then, for all sufficiently large n,

$$k_c(1-\eta_{n,c}) \le \bar{d}_c \le 2(Ms_c+1), \tag{G.15}$$

and, uniformly over candidates,

$$\begin{aligned} \bar{d}_c &\asymp k_c, & \bar{d}_c &\to \infty, \\ \frac{\mathrm{Var}(d_c)}{\bar{d}_c^2} &\to 0, & \frac{d_c}{\bar{d}_c} &\to_p 1, \\ \frac{R_{d,c}}{\bar{d}_c^2} &\to 1, & \frac{\bar{d}_c}{\sqrt{n}} &\to 0. \end{aligned} \tag{G.16}$$

Proof. Throughout Step 4, C is the universal constant of Lemma F.7, held fixed; elsewhere C is generic.

Step 1: $\mathbb{P}(\mathcal{G}_c^c) \le \eta_{n,c}$ and $a_{n,c} \to 0$.

Under Assumption 1(i) the scores are mean-zero Gaussian, and the column normalization $\mathbb{E}_n[z_{c,ij}^2] = 1$ of Assumption 2(i) fixes their variance, so

$$\mathbb{E}_n[z_{c,ij}v_i] \sim N\left(0, \frac{\sigma_v^2}{n}\right), \qquad j = 1, \ldots, p_c.$$

A union bound over $j \in [p_c]$ then gives

$$\mathbb{P}\left(\left\|\mathbb{E}_n[z_{c,i}v_i]\right\|_\infty > \frac{\lambda_c}{c_\lambda}\right) \le 2p_c\Phi\left(-\frac{\sqrt{n}\lambda_c}{c_\lambda\sigma_v}\right) = \eta_{n,c}. \tag{G.17}$$

By Lemma F.2 and the penalty restriction in Assumption 2(iii),

$$\eta_{n,c} \le 2p_c \exp\left\{-\frac{n\lambda_c^2}{2c_\lambda^2\sigma_v^2}\right\} \le \{\log(p_c \vee n)\}^{-\alpha} \to 0.$$

Parts (ii) and (v) of Assumption 2, together with $\eta_{n,c} \to 0$, then give $a_{n,c} \to 0$. Since $[J]$ is finite, $\eta_{n,c} \to 0$ and $a_{n,c} \to 0$ uniformly over candidates.

Step 2: Bound for $\overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2]$.

Define $U_c := \mathbb{E}_n[(\widehat{\Pi}_{c,i} - \Pi_i)^2]^{1/2}$. Since $\widehat{\Pi}_{c,i} = z_{c,i}'\widehat{\pi}_c$ and $\Pi_i = z_{c,i}'\pi_c^0 + \xi_{c,i}$,

$$\widehat{\Pi}_{c,i} - \Pi_i = z_{c,i}'(\widehat{\pi}_c - \pi_c^0) - \xi_{c,i}.$$

On $\mathcal{G}_c$, the triangle inequality and Lemma G.1 give

$$\mathbb{E}_n[(\widehat{\Pi}_{c,i} - \Pi_i)^2]^{1/2} \le 3\mathbb{E}_n[\xi_{c,i}^2]^{1/2} + \frac{2(1+c_\lambda^{-1})\lambda_c\sqrt{s_c}}{\kappa_0}. \tag{G.18}$$

A crude envelope, needed only off $\mathcal{G}_c$, holds for every realization: $\mathbb{E}_n[\widehat{\Pi}_{c,i}^2] \le \mathbb{E}_n[x_i^2]$ by (E.4), so the triangle inequality and $x_i = \Pi_i + v_i$ imply

$$U_c \le \mathbb{E}_n[\widehat{\Pi}_{c,i}^2]^{1/2} + H^{1/2} \le \mathbb{E}_n[x_i^2]^{1/2} + H^{1/2} \le 2H^{1/2} + \mathbb{E}_n[v_i^2]^{1/2}. \tag{G.19}$$

By (G.19) and $(a+b)^4 \le 8(a^4+b^4)$, $U_c^4 \le 8\{16H^2 + \mathbb{E}_n[v_i^2]^2\}$, while Assumption 1(i) gives $\mathbb{E}[\mathbb{E}_n[v_i^2]^2] = \sigma_v^4(1+2/n) \le 3\sigma_v^4$; with $H \le \overline{H}$ from (3), this gives $\mathbb{E}[U_c^4] \le C$. Decompose

$$\mathbb{E}[U_c^2] = \mathbb{E}[U_c^2 \mathbb{1}\{\mathcal{G}_c\}] + \mathbb{E}[U_c^2 \mathbb{1}\{\mathcal{G}_c^c\}].$$

By (G.18) and $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathbb{E}[U_c^2 \mathbb{1}\{\mathcal{G}_c\}] \le 18\mathbb{E}_n[\xi_{c,i}^2] + \frac{8(1+c_\lambda^{-1})^2\lambda_c^2 s_c}{\kappa_0^2}.$$

For the off-event term, Cauchy–Schwarz gives

$$\mathbb{E}[U_c^2 \mathbb{1}\{\mathcal{G}_c^c\}] \le \mathbb{E}[U_c^4]^{1/2}\mathbb{P}(\mathcal{G}_c^c)^{1/2}.$$

Adding the two bounds and using $\mathbb{E}[U_c^4] \le C$, $\mathbb{P}(\mathcal{G}_c^c) \le \eta_{n,c}$, and the definition of $a_{n,c}$ gives

$$\begin{aligned}\overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2] &= \mathbb{E}[U_c^2] \\ &\le C\{\mathbb{E}_n[\xi_{c,i}^2] + s_c\lambda_c^2\} + \mathbb{E}[U_c^4]^{1/2}\mathbb{P}(\mathcal{G}_c^c)^{1/2} \\ &\le C\{\mathbb{E}_n[\xi_{c,i}^2] + s_c\lambda_c^2 + \sqrt{\eta_{n,c}}\} = O(a_{n,c}),\end{aligned} \tag{G.20}$$

which is (G.11).

Step 3: Bounds for $h_c - H$, $\mathcal{A}_c$, and $\overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - H$.

With h_c defined by (18),

$$\begin{aligned} h_c - H &= \overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)\Pi_i], \\ \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - H &= 2(h_c - H) + \overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2], \\ \mathcal{A}_c &= \overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2] - \frac{(h_c - H)^2}{H}. \end{aligned}$$

Here the first two identities expand (18) and $\mathbb{E}_n[\Pi_i^2] = H$, while the third follows from the representation (A.5). By Cauchy–Schwarz under $\overline{\mathbb{E}}$, $\mathcal{A}_c \ge 0$ from (A.5), and (G.20),

$$|h_c - H|^2 \le H\overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2] = O(a_{n,c}), \qquad 0 \le \mathcal{A}_c \le \overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2] = O(a_{n,c}).$$

Taking square roots in the first bound gives the rate for $h_c - H$ in (G.12); the second bound is (G.13). Moreover, for the rate on $\overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - H$,

$$|\overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] - H| \le 2|h_c - H| + \overline{\mathbb{E}}[(\widehat{\Pi}_{c,i} - \Pi_i)^2] = O(\sqrt{a_{n,c}} + a_{n,c}) = O(\sqrt{a_{n,c}}),$$

where the last equality uses $a_{n,c} \to 0$, so (G.12) holds. Finally,

$$h_c \ge H - C\sqrt{a_{n,c}}, \qquad h_c + \overline{\mathbb{E}}[\widehat{\Pi}_{c,i}^2] \le 2H + C\sqrt{a_{n,c}},$$

so the bounds on H in (3) and $a_{n,c} \to 0$ give (G.14).

Step 4: Two-sided bounds for $\bar{d}_c$.

Suppose now that Assumption 3 holds. By (G.5), $k_c \le d_c < Ms_c + 1$ on $\mathcal{G}_c$. Set $b_c := Ms_c + 1$. Since $\mathbb{P}(\mathcal{G}_c) \ge 1 - \eta_{n,c}$,

$$\bar{d}_c \ge \mathbb{E}[d_c \mathbb{1}\{\mathcal{G}_c\}] \ge k_c \mathbb{P}(\mathcal{G}_c) \ge k_c(1 - \eta_{n,c}). \tag{G.21}$$

Moreover, $d_c \le b_c$ on $\mathcal{G}_c$ by (G.5), so $(d_c - \bar{d}_c)^2 \ge (\bar{d}_c - b_c)_+^2$ on that event, and hence

$$(1 - \eta_{n,c})(\bar{d}_c - b_c)_+^2 \le \mathbb{E}[(d_c - \bar{d}_c)^2 \mathbb{1}\{\mathcal{G}_c\}] \le \mathrm{Var}(d_c).$$

For all sufficiently large n, $\eta_{n,c} \le 1/2$ and $\bar{d}_c \ge k_c/2 \ge 1$, while $16C \log(2ep_c/k_c) < k_c/2$ by the growth condition (17). Fix such an n and suppose $\bar{d}_c > 2b_c$. Then $\bar{d}_c - b_c > \bar{d}_c/2$, so the prior display and $\eta_{n,c} \le 1/2$ give $\bar{d}_c^2/8 \le \mathrm{Var}(d_c)$. Since $d_c \le p_c$ we have $\log(ep_c/\bar{d}_c) \ge 1$, so Lemma F.7 gives

$$\begin{aligned}
\frac{\bar{d}_c^2}{8} &\le C\left[1 + \bar{d}_c \log\left(\frac{ep_c}{\bar{d}_c}\right)\right] \le 2C\bar{d}_c \log\left(\frac{ep_c}{\bar{d}_c}\right), \\
\bar{d}_c &\le 16C \log\left(\frac{ep_c}{\bar{d}_c}\right) \le 16C \log\left(\frac{2ep_c}{k_c}\right) < \frac{k_c}{2} \le \bar{d}_c,
\end{aligned}$$

which is impossible. Hence $\bar{d}_c \le 2b_c$; together with (G.21), this gives (G.15). Since $s_c \le C_s k_c$ and $k_c \to \infty$ by (17), the bounds (G.15) imply $\bar{d}_c \asymp k_c$ and $\bar{d}_c \ge k_c/2 \to \infty$.

Step 5: Concentration of d_c and the remaining rank limits.

Since $\bar{d}_c \asymp k_c$ by Step 4 and $k_c \le p_c$, $\log(ep_c/\bar{d}_c) = \log(ep_c/k_c) + O(1)$, so the growth condition (17) in Assumption 3 yields

$$\frac{1 + \log\{ep_c/\bar{d}_c\}}{\bar{d}_c} \le C \frac{1 + \log\{p_c/k_c\}}{k_c} \to 0. \tag{G.22}$$

For all sufficiently large n, Lemma F.7 gives

$$\frac{\mathrm{Var}(d_c)}{\bar{d}_c^2} \le C\left[\frac{1}{\bar{d}_c^2} + \frac{1 + \log\{ep_c/\bar{d}_c\}}{\bar{d}_c}\right] \to 0,$$

where the convergence uses $\bar{d}_c \to \infty$ and (G.22). Chebyshev's inequality then yields $d_c/\bar{d}_c \to_p 1$. By (A.2),

$$\left|\frac{R_{d,c}}{\bar{d}_c^2} - 1\right| = \frac{\mathrm{Var}(d_c)}{\bar{d}_c^2} + \frac{1}{\bar{d}_c} \to 0.$$

By Assumption 2(v) and (G.15),

$$\frac{\bar{d}_c}{\sqrt{n}} \le 2M \frac{s_c}{\sqrt{n}} + 2n^{-1/2} \to 0.$$

These limits, together with $\bar{d}_c \asymp k_c$ and $\bar{d}_c \to \infty$ from Step 4, hold uniformly over the finitely many candidates, giving (G.16). □

Proposition G.1 (Infeasible Criterion Rates)**.** *Suppose Assumptions 1 to 3 hold, with $J < \infty$ fixed. Then* (G.15) *and* (G.16) *hold and, uniformly over candidates,*

$$S_c \asymp \mathcal{A}_c + \frac{R_{d,c}}{n} \asymp \mathcal{A}_c + \frac{\bar{d}_c^2}{n}, \qquad \max_c S_c \to 0, \qquad \min_c nS_c \to \infty. \tag{G.23}$$

Moreover, for all sufficiently large n, uniformly over candidates,

$$\begin{aligned}\sqrt{\frac{\mathcal{A}_c}{n}}+\frac{\bar{d}_c}{n}&\le C\sqrt{\frac{S_c}{n}},\\ \sqrt{\frac{\mathcal{A}_c}{n}}+\frac{\bar{d}_c}{n}&\le \frac{CS_c}{\bar{d}_c},\\ \frac{\sqrt{\bar{d}_c}}{n}&\le \frac{CS_c}{\bar{d}_c^{3/2}}.\end{aligned}\tag{G.24}$$

Proof. The hypotheses of Lemma G.2 hold, giving (G.15) and (G.16). Since $\sigma_\varepsilon^2>0$ by the nonsingularity of Ω in Assumption 1(i) and $\sigma_{\varepsilon v}\neq 0$ by (3), (20), (G.14), and (G.16) give

$$C^{-1}\left(\mathcal{A}_c+\frac{R_{d,c}}{n}\right)\le S_c\le C\left(\mathcal{A}_c+\frac{R_{d,c}}{n}\right),\qquad \frac{R_{d,c}}{\bar{d}_c^2}\to 1.\tag{G.25}$$

Moreover, Assumption 2(v), the growth condition (17), and Lemma G.2 give

$$\mathcal{A}_c\to 0,\qquad \frac{\bar{d}_c}{\sqrt{n}}\to 0,\qquad \bar{d}_c\ge C^{-1}k_c\to\infty.$$

Combining these bounds with (G.25) yields

$$S_c\le C\left(\mathcal{A}_c+\frac{\bar{d}_c^2}{n}\right)\to 0,\qquad nS_c\ge C^{-1}\bar{d}_c^2\to\infty.$$

As $[J]$ is fixed, these bounds hold uniformly over candidates, proving (G.23).

It remains to prove (G.24). By (A.2), $R_{d,c}\ge \bar{d}_c^2$, while the lower bound in (G.25) gives $\mathcal{A}_c\le CS_c$ and $R_{d,c}/n\le CS_c$, so $\bar{d}_c^2/n\le CS_c$. The first bound follows from

$$\sqrt{\frac{\mathcal{A}_c}{n}}\le\sqrt{\frac{CS_c}{n}},\qquad \frac{\bar{d}_c}{n}=\frac{1}{\sqrt{n}}\sqrt{\frac{\bar{d}_c^2}{n}}\le\sqrt{\frac{CS_c}{n}}.$$

For the second, $\bar{d}_c\ge 1$ for large n by (G.15), and $2ab\le a^2+b^2$ with $a=\sqrt{\mathcal{A}_c}$ and $b=\bar{d}_c/\sqrt{n}$ gives $2\bar{d}_c\sqrt{\mathcal{A}_c/n}\le\mathcal{A}_c+\bar{d}_c^2/n\le CS_c$, so $\sqrt{\mathcal{A}_c/n}\le CS_c/\bar{d}_c$, while $\bar{d}_c/n=(\bar{d}_c^2/n)/\bar{d}_c\le CS_c/\bar{d}_c$. The third bound follows from $\sqrt{\bar{d}_c}/n=(\bar{d}_c^2/n)/\bar{d}_c^{3/2}\le CS_c/\bar{d}_c^{3/2}$. □

H Simulation Details and Complete Results

H.1 Implementation Details

We use the residualized movie sales and weather data for the 1,671 opening-weekend observations in Gilchrist and Sands (2016). The original date controls have already been partialled out. Four linearly dependent weather indicators are removed, leaving 34 temperature instruments and 14 rain, snow, and precipitation instruments. The Expanded dictionary adds all $34\times 14=476$ interactions between these two groups. Each instrument is centered and scaled to have unit empirical root mean square; this normalization is repeated within the fixed subsample when $n=800$.

For a dictionary with p_c columns, we use the homoskedastic BCCH plug-in penalty

$$\lambda_c = \frac{1.1\,\varkappa_c}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\gamma_c}{2p_c}\right), \qquad \gamma_c = \frac{0.1}{\log(\max\{p_c, n\})},$$

where Φ is the standard normal distribution function and the first-stage error standard deviation is known to be one. Each dictionary uses the multipliers $\varkappa_c = 2^a$ for $a = -4, -3.5, \ldots, 2$. The four identification strengths in Section 4.1 correspond to $H = 0.0803, 0.1606, 0.3156, 0.6885$. At $n = 800$, we restrict the full-sample first stage to one fixed subsample and rescale it to preserve H, without re-estimating its direction. The corresponding expected F statistics are 2.345, 3.687, 6.276, and 12.509.

We calibrate the error correlation using cumulative sales in weeks 2–6. The estimated error correlation in the data is $\widehat{\rho} = \widehat{\sigma}_{\varepsilon v}/(\widehat{\sigma}_\varepsilon \widehat{\sigma}_v) = -0.212$, with the moments in (25) computed from pilot residuals and $\widehat{\sigma}_v$ the residual standard deviation from the least-squares regression on the Original instruments. To account for attenuation from estimating the pilot coefficient, we choose $\rho = -0.291$ so that the mean of $\widehat{\rho}$ over 2,000 independent replications of the $n = 1{,}671$, $F = 3.8$ Gaussian design equals -0.212. For each sample size and identification strength, the same 1,000 evaluation draws are used at all seven correlations.

We select the pilot on the Original dictionary over a finer grid of 29 penalty multipliers from 2^{-12} to 2^2, choosing the fit with the largest absolute IV denominator $\widehat{h}_c$ in (24) among those with finite $|\widehat{h}_c| \geq n^{-1/2}$. The pilot estimate is used to construct estimators of the error moments and to replace a zero estimate for the first-stage when BCCH selects no instruments. This replacement occurs in 27.1% of Gaussian replications at $n = 800$ and $F = 3.8$, and in fewer than 1% in the other Gaussian designs.

Cross-validation uses one prespecified ten-fold assignment, keeping observations from the same opening weekend in the same fold. Each training fit uses the full-sample penalty and instrument normalization, and candidates are compared by the average prediction error across held-out observations. The selected candidate is then refit on the full sample. Effective dimension is computed as the rank of the numerical equicorrelation set in (10). We include all instruments with nonzero coefficients in this set to avoid omitting an active instrument because of numerical error.

For each observation in the Laplace designs, we draw mutually independent $W_i \sim \mathrm{Exp}(1)$ and $a_i, b_i \sim N(0, 1)$, and set

$$v_i = \sqrt{W_i}\,a_i, \qquad \varepsilon_i = \sqrt{W_i}\{\rho a_i + \sqrt{1-\rho^2}\,b_i\}.$$

The draws are independent across observations. Both errors have variance-one Laplace marginals and correlation ρ. Repeating the calibration under these errors gives $\rho = -0.290$. The simulation details are otherwise left completely unchanged. The same draws of (W_i, a_i, b_i) are used across correlations at each sample size and identification strength.

The oracle candidate minimizes the infeasible criterion S_c in (20), with its population moments estimated from a separate set of 1,000 independent replications. The Known moments rule instead evaluates $\widehat{S}_c$ at $\sigma_\varepsilon^2 = 1$ and $\sigma_{\varepsilon v} = \rho$. Both benchmarks are infeasible.

H.2 Complete Results

The following tables report relative risk and mean effective dimension for all Gaussian and Laplace designs. Risk is the mean of $n(\widehat{\beta} - \beta)^2$ over 1,000 replications. Calibrated denotes the correlation obtained above. The oracle and Known moments columns use the infeasible benchmarks described in the preceding subsection.

Table 3: Relative risk by simulation design under Gaussian errors.

	$n = 800$			$n = 1{,}671$		
Correlation	Criterion	Cross-validation	BCCH	Criterion	Cross-validation	BCCH
$F = 3.8$						
Calibrated (0.29)	1.266	1.000	1.532	1.128	1.000	2.045
0.40	1.226	1.000	1.273	1.047	1.000	1.579
0.50	1.105	1.000	1.111	1.000	1.009	1.275
0.60	1.000	1.008	1.007	1.000	1.097	1.133
0.70	1.007	1.086	1.000	1.000	1.212	1.048
0.80	1.000	1.174	1.017	1.019	1.352	1.000
0.90	1.000	1.244	1.030	1.086	1.551	1.000
$F = 6.6$						
Calibrated (0.29)	1.117	1.000	1.833	1.044	1.000	1.684
0.40	1.010	1.000	1.403	1.050	1.000	1.378
0.50	1.000	1.043	1.169	1.000	1.000	1.152
0.60	1.000	1.150	1.059	1.000	1.098	1.072
0.70	1.026	1.284	1.000	1.000	1.231	1.039
0.80	1.084	1.479	1.000	1.000	1.363	1.013
0.90	1.149	1.665	1.000	1.025	1.501	1.000
$F = 12$						
Calibrated (0.29)	1.050	1.000	1.676	1.004	1.000	1.445
0.40	1.040	1.000	1.371	1.035	1.000	1.300
0.50	1.000	1.025	1.175	1.055	1.000	1.173
0.60	1.000	1.118	1.089	1.019	1.000	1.062
0.70	1.000	1.227	1.034	1.000	1.039	1.006
0.80	1.001	1.348	1.000	1.000	1.130	1.005
0.90	1.053	1.505	1.000	1.000	1.214	1.000
$F = 25$						
Calibrated (0.29)	1.005	1.000	1.525	1.000	1.000	1.379
0.40	1.027	1.000	1.379	1.000	1.001	1.289
0.50	1.055	1.000	1.249	1.013	1.000	1.201
0.60	1.038	1.000	1.132	1.035	1.000	1.116
0.70	1.000	1.046	1.077	1.038	1.000	1.035
0.80	1.000	1.146	1.080	1.055	1.040	1.000
0.90	1.000	1.241	1.077	1.072	1.117	1.000
Average	1.045	1.135	1.189	1.026	1.106	1.194
Maximum	1.266	1.665	1.833	1.128	1.551	2.045

Notes: Risk is relative to the lowest risk among the three rules in each design. Summary rows report the average and maximum of the 28 ratios in each column.

Table 4: Relative risk by simulation design under Laplace errors.

	$n = 800$			$n = 1{,}671$		
Correlation	Criterion	Cross-validation	BCCH	Criterion	Cross-validation	BCCH
$F = 3.8$						
Calibrated (0.29)	1.261	1.000	1.586	1.134	1.000	1.683
0.40	1.204	1.000	1.335	1.046	1.000	1.310
0.50	1.065	1.000	1.172	1.000	1.038	1.107
0.60	1.000	1.003	1.058	1.000	1.137	1.017
0.70	1.000	1.093	1.062	1.024	1.297	1.000
0.80	1.000	1.156	1.052	1.087	1.468	1.000
0.90	1.000	1.218	1.052	1.134	1.627	1.000
$F = 6.6$						
Calibrated (0.29)	1.170	1.000	2.010	1.058	1.000	1.694
0.40	1.057	1.000	1.524	1.058	1.000	1.415
0.50	1.000	1.010	1.219	1.023	1.000	1.207
0.60	1.000	1.111	1.090	1.000	1.068	1.115
0.70	1.005	1.221	1.000	1.000	1.179	1.081
0.80	1.088	1.427	1.000	1.000	1.288	1.051
0.90	1.169	1.631	1.000	1.000	1.385	1.018
$F = 12$						
Calibrated (0.29)	1.071	1.000	1.653	1.003	1.000	1.423
0.40	1.039	1.000	1.365	1.011	1.000	1.268
0.50	1.000	1.019	1.177	1.030	1.000	1.143
0.60	1.000	1.117	1.108	1.000	1.033	1.069
0.70	1.000	1.238	1.071	1.000	1.098	1.040
0.80	1.000	1.343	1.030	1.000	1.166	1.020
0.90	1.017	1.453	1.000	1.000	1.245	1.018
$F = 25$						
Calibrated (0.29)	1.004	1.000	1.462	1.005	1.000	1.435
0.40	1.033	1.000	1.308	1.013	1.000	1.355
0.50	1.042	1.000	1.175	1.021	1.000	1.269
0.60	1.008	1.000	1.058	1.036	1.000	1.180
0.70	1.000	1.080	1.038	1.049	1.000	1.091
0.80	1.000	1.212	1.068	1.031	1.000	1.007
0.90	1.000	1.296	1.059	1.040	1.079	1.000
Average	1.044	1.130	1.205	1.029	1.111	1.179
Maximum	1.261	1.631	2.010	1.134	1.627	1.694

Notes: Risk is relative to the lowest risk among the three rules in each design. The calibrated correlation is recomputed under Laplace errors.

Table 5: Mean effective dimension under Gaussian errors.

	$n = 800$			$n = 1{,}671$		
Correlation	Criterion	Cross-validation	BCCH	Criterion	Cross-validation	BCCH
$F = 3.8$						
Calibrated (0.29)	73.1	19.7	14.2	39.8	30.6	3.4
0.40	43.7	19.7	14.2	25.7	30.6	3.4
0.50	27.8	19.7	14.2	19.7	30.6	3.4
0.60	18.9	19.7	14.2	16.1	30.6	3.4
0.70	15.6	19.7	14.2	13.5	30.6	3.4
0.80	13.0	19.7	14.2	11.9	30.6	3.4
0.90	11.3	19.7	14.2	10.5	30.6	3.4
$F = 6.6$						
Calibrated (0.29)	42.7	30.3	3.4	35.5	38.2	6.0
0.40	27.2	30.3	3.4	27.6	38.2	6.0
0.50	21.2	30.3	3.4	21.5	38.2	6.0
0.60	17.4	30.3	3.4	17.3	38.2	6.0
0.70	14.5	30.3	3.4	14.7	38.2	6.0
0.80	12.5	30.3	3.4	13.0	38.2	6.0
0.90	11.4	30.3	3.4	11.8	38.2	6.0
$F = 12$						
Calibrated (0.29)	35.4	37.4	5.9	38.2	39.8	9.9
0.40	27.8	37.4	5.9	34.4	39.8	9.9
0.50	21.8	37.4	5.9	27.5	39.8	9.9
0.60	18.0	37.4	5.9	21.1	39.8	9.9
0.70	15.4	37.4	5.9	17.2	39.8	9.9
0.80	13.8	37.4	5.9	15.0	39.8	9.9
0.90	12.7	37.4	5.9	13.8	39.8	9.9
$F = 25$						
Calibrated (0.29)	38.3	39.7	10.2	40.4	41.2	12.6
0.40	35.0	39.7	10.2	39.3	41.2	12.6
0.50	28.3	39.7	10.2	37.0	41.2	12.6
0.60	22.0	39.7	10.2	32.2	41.2	12.6
0.70	18.1	39.7	10.2	26.0	41.2	12.6
0.80	16.1	39.7	10.2	20.4	41.2	12.6
0.90	14.9	39.7	10.2	17.6	41.2	12.6
Average	23.9	31.8	8.4	23.5	37.4	8.0
Maximum	73.1	39.7	14.2	40.4	41.2	12.6

Notes: Mean effective dimension over 1,000 replications. Cross-validation and the BCCH penalty use the first stage alone, so their entries do not vary with the correlation. When BCCH selects no instruments, the entries include the effective dimension of the pilot fit used instead.

Table 6: Mean effective dimension under Laplace errors.

	$n = 800$			$n = 1{,}671$		
Correlation	Criterion	Cross-validation	BCCH	Criterion	Cross-validation	BCCH
$F = 3.8$						
Calibrated (0.29)	75.9	19.9	14.1	41.6	31.2	3.5
0.40	45.7	19.9	14.1	26.7	31.2	3.5
0.50	25.9	19.9	14.1	20.7	31.2	3.5
0.60	19.0	19.9	14.1	16.2	31.2	3.5
0.70	14.8	19.9	14.1	13.3	31.2	3.5
0.80	12.7	19.9	14.1	11.7	31.2	3.5
0.90	11.4	19.9	14.1	10.6	31.2	3.5
$F = 6.6$						
Calibrated (0.29)	44.9	29.7	3.3	35.6	38.0	6.1
0.40	27.5	29.7	3.3	28.0	38.0	6.1
0.50	21.2	29.7	3.3	21.4	38.0	6.1
0.60	16.9	29.7	3.3	17.0	38.0	6.1
0.70	14.4	29.7	3.3	14.5	38.0	6.1
0.80	12.4	29.7	3.3	12.8	38.0	6.1
0.90	11.1	29.7	3.3	11.7	38.0	6.1
$F = 12$						
Calibrated (0.29)	37.6	37.5	5.9	38.2	39.8	9.9
0.40	28.1	37.5	5.9	34.5	39.8	9.9
0.50	22.4	37.5	5.9	27.6	39.8	9.9
0.60	18.2	37.5	5.9	21.1	39.8	9.9
0.70	15.5	37.5	5.9	17.1	39.8	9.9
0.80	13.7	37.5	5.9	15.2	39.8	9.9
0.90	12.6	37.5	5.9	14.0	39.8	9.9
$F = 25$						
Calibrated (0.29)	38.6	39.7	10.2	40.3	41.3	12.6
0.40	34.7	39.7	10.2	39.1	41.3	12.6
0.50	29.2	39.7	10.2	37.0	41.3	12.6
0.60	22.3	39.7	10.2	32.7	41.3	12.6
0.70	18.4	39.7	10.2	26.4	41.3	12.6
0.80	16.0	39.7	10.2	20.7	41.3	12.6
0.90	15.1	39.7	10.2	17.6	41.3	12.6
Average	24.1	31.7	8.4	23.7	37.6	8.0
Maximum	75.9	39.7	14.1	41.6	41.3	12.6

Notes: Mean effective dimension over 1,000 replications. When BCCH selects no instruments, the entries include the effective dimension of the pilot fit used instead.

Table 7: Risk relative to the oracle under Gaussian errors.

	$n = 800$				$n = 1{,}671$			
Correlation	Criterion	Known moments	Cross-validation	BCCH	Criterion	Known moments	Cross-validation	BCCH
$F = 3.8$								
Calibrated (0.29)	1.311	1.005	1.036	1.587	1.108	1.012	0.983	2.010
0.40	1.440	1.126	1.175	1.496	1.131	1.099	1.080	1.706
0.50	1.418	1.182	1.283	1.425	1.218	1.153	1.229	1.553
0.60	1.399	1.221	1.410	1.408	1.274	1.208	1.397	1.443
0.70	1.403	1.269	1.513	1.394	1.278	1.223	1.549	1.340
0.80	1.359	1.279	1.595	1.382	1.268	1.247	1.683	1.245
0.90	1.333	1.289	1.658	1.373	1.259	1.243	1.797	1.159
$F = 6.6$								
Calibrated (0.29)	1.117	1.002	1.000	1.833	1.048	1.036	1.005	1.691
0.40	1.147	1.087	1.135	1.592	1.074	1.052	1.022	1.409
0.50	1.218	1.117	1.270	1.424	1.151	1.096	1.150	1.325
0.60	1.247	1.166	1.434	1.321	1.159	1.097	1.272	1.243
0.70	1.262	1.196	1.580	1.230	1.165	1.129	1.434	1.210
0.80	1.250	1.213	1.705	1.153	1.164	1.148	1.587	1.179
0.90	1.250	1.221	1.811	1.088	1.215	1.206	1.779	1.185
$F = 12$								
Calibrated (0.29)	1.026	1.017	0.977	1.637	1.004	1.010	1.000	1.446
0.40	1.071	1.041	1.030	1.412	1.035	1.032	1.000	1.300
0.50	1.114	1.058	1.141	1.309	1.070	1.054	1.014	1.190
0.60	1.163	1.121	1.300	1.267	1.124	1.111	1.104	1.173
0.70	1.184	1.147	1.453	1.224	1.158	1.138	1.203	1.165
0.80	1.184	1.183	1.594	1.182	1.153	1.141	1.302	1.159
0.90	1.203	1.190	1.719	1.142	1.135	1.126	1.377	1.135
$F = 25$								
Calibrated (0.29)	1.005	1.013	1.000	1.525	1.000	0.999	1.000	1.379
0.40	1.027	1.041	1.000	1.379	0.999	0.999	1.000	1.288
0.50	1.053	1.046	0.998	1.247	1.017	1.020	1.005	1.207
0.60	1.112	1.081	1.071	1.213	1.043	1.054	1.008	1.125
0.70	1.116	1.079	1.167	1.202	1.075	1.080	1.036	1.072
0.80	1.097	1.079	1.257	1.184	1.115	1.098	1.099	1.057
0.90	1.112	1.104	1.380	1.198	1.128	1.122	1.175	1.052
Average	1.201	1.128	1.310	1.351	1.127	1.105	1.225	1.301
Maximum	1.440	1.289	1.811	1.833	1.278	1.247	1.797	2.010

Notes: Risk is relative to the oracle candidate in each design. The oracle is chosen using 1,000 independent replications to estimate S_c. Known moments uses the population error moments in the feasible criterion.

Table 8: Risk relative to the oracle under Laplace errors.

	$n = 800$				$n = 1{,}671$			
Correlation	Criterion	Known moments	Cross-validation	BCCH	Criterion	Known moments	Cross-validation	BCCH
$F = 3.8$								
Calibrated (0.29)	1.345	1.014	1.067	1.692	1.154	1.032	1.018	1.713
0.40	1.432	1.138	1.190	1.588	1.179	1.087	1.127	1.476
0.50	1.369	1.187	1.285	1.506	1.168	1.081	1.212	1.293
0.60	1.396	1.246	1.400	1.477	1.202	1.133	1.367	1.223
0.70	1.366	1.276	1.493	1.451	1.190	1.170	1.508	1.162
0.80	1.357	1.285	1.568	1.427	1.209	1.186	1.633	1.112
0.90	1.336	1.288	1.627	1.405	1.215	1.202	1.743	1.071
$F = 6.6$								
Calibrated (0.29)	1.083	0.974	0.926	1.860	1.058	1.036	1.001	1.695
0.40	1.149	1.082	1.087	1.657	1.074	1.045	1.016	1.437
0.50	1.227	1.133	1.239	1.496	1.155	1.105	1.129	1.364
0.60	1.260	1.178	1.400	1.373	1.143	1.098	1.220	1.274
0.70	1.271	1.229	1.545	1.265	1.153	1.126	1.360	1.246
0.80	1.274	1.246	1.672	1.172	1.160	1.142	1.495	1.220
0.90	1.278	1.260	1.784	1.093	1.192	1.187	1.651	1.213
$F = 12$								
Calibrated (0.29)	1.052	1.004	0.983	1.624	1.003	1.005	1.000	1.423
0.40	1.088	1.059	1.047	1.429	1.012	1.023	1.000	1.268
0.50	1.142	1.078	1.163	1.343	1.098	1.073	1.066	1.218
0.60	1.176	1.115	1.314	1.303	1.124	1.106	1.161	1.202
0.70	1.178	1.161	1.458	1.262	1.146	1.136	1.258	1.192
0.80	1.186	1.176	1.593	1.221	1.157	1.141	1.350	1.181
0.90	1.200	1.206	1.715	1.180	1.151	1.137	1.434	1.172
$F = 25$								
Calibrated (0.29)	1.004	1.005	1.000	1.462	1.005	1.006	1.000	1.435
0.40	1.033	1.033	1.000	1.308	1.013	1.016	1.000	1.355
0.50	1.090	1.091	1.046	1.229	1.021	1.022	1.000	1.270
0.60	1.142	1.099	1.133	1.199	1.017	1.024	0.982	1.158
0.70	1.141	1.081	1.233	1.184	1.059	1.061	1.010	1.102
0.80	1.120	1.102	1.358	1.197	1.096	1.084	1.063	1.070
0.90	1.131	1.124	1.466	1.198	1.110	1.108	1.153	1.068
Average	1.208	1.138	1.314	1.379	1.117	1.092	1.213	1.272
Maximum	1.432	1.288	1.784	1.860	1.215	1.202	1.743	1.713

Notes: Risk is relative to the oracle candidate in each design. Known moments uses the population error moments in the feasible criterion.