

Bias-Aware Inference in Regularized Regression Models*

Timothy B. Armstrong[†]

Yale University

Michal Kolesár[‡]

Princeton University

Soonwoo Kwon[§]

Yale University

December 29, 2020

Abstract

We consider inference on a regression coefficient under a constraint on the magnitude of the control coefficients. We show that a class of estimators based on an auxiliary regularized regression of the regressor of interest on control variables exactly solves a tradeoff between worst-case bias and variance. We derive “bias-aware” confidence intervals (CIs) based on these estimators, which take into account possible bias when forming the critical value. We show that these estimators and CIs are near-optimal in finite samples for mean squared error and CI length. Our finite-sample results are based on an idealized setting with normal regression errors with known homoskedastic variance, and we provide conditions for asymptotic validity with unknown and possibly heteroskedastic error distribution. Focusing on the case where the constraint on the magnitude of control coefficients is based on an ℓ_p norm ($p \geq 1$), we derive rates of convergence for optimal estimators and CIs under high-dimensional asymptotics that allow the number of regressors to increase more quickly than the number of observations.

*Parts of this paper include material from Section 4 of the working paper [Armstrong and Kolesár \(2016\)](#), which was taken out in the final published version ([Armstrong and Kolesár, 2018](#)). An earlier version of this paper was circulated under the title “Optimal Inference in Regularized Regression Models.” We thank Mark Li and Ulrich Müller for sharing their code. Kolesár acknowledges support by the Sloan Research Fellowship.

[†]email: timothy.armstrong@yale.edu

[‡]email: mkolesar@princeton.edu

[§]email: soonwoo.kwon@yale.edu

1 Introduction

We are interested in estimation and inference on a scalar coefficient β in a linear regression model

$$Y_i = w_i\beta + z_i'\gamma + \varepsilon, \quad i = 1, \dots, n, \quad (1)$$

where the k -vector of controls may be large. In such settings, the classic ordinary least squares (OLS) estimator exhibits variance that is too large to yield informative results, and it is not even defined when $k > n$. To ameliorate this, the regularized regression literature has considered modifying the OLS objective function to penalize large values of γ , thereby lowering variance at the cost of increased bias.

The most popular of these approaches is to use the lasso (Tibshirani, 1996) or other variants of ℓ_1 penalization (e.g. Candès and Tao, 2007; Belloni et al., 2011). There is a large literature (see, e.g. Bühlmann and van de Geer, 2011, for a review) showing favorable mean squared error (MSE) properties of these estimates under the assumption of sparsity on γ . For inference, several papers have proposed CIs based on “double lasso” estimators (see, among others, Belloni et al., 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014), with asymptotic justification relying on rate conditions for the sparsity of γ . However, in many applications in economics, the sparsity assumption may not be compelling. Furthermore, it is unclear what sparsity bound this approach implicitly imposes in a given finite sample.

In this paper, we take a different approach. Our approach is based on imposing an a priori bound on the magnitude of the control coefficients, formalized using a penalty function $\text{Pen}(\cdot)$: we assume $\text{Pen}(\gamma) \leq C$. In our leading specifications, we take the penalty to be an ℓ_p norm, but our framework can incorporate any restrictions on γ that place it in a convex symmetric set.¹ For example, if $z_i'\gamma$ is a basis approximation to some smooth function, we can define $\text{Pen}(\gamma)$ to incorporate bounds on the derivatives of this function. The regularity parameter C plays a role analogous to sparsity bound.

In this setting, we obtain sharp finite-sample results deriving near-optimal estimators and CIs under the idealized assumption that the regression errors ε_i are Gaussian with a known homoskedastic variance. We also study the optimal rates of convergence under high-dimensional asymptotics when $k \gg n$. Finally, we discuss the use of heteroskedasticity-robust variance estimators to form feasible versions of our CIs, and give conditions for their asymptotic validity.

Our main finite-sample result shows that the class of estimators that exactly resolves the

¹While we rule out sparsity constraints (which are non-convex), our results have implications for this case as well. See Section 5 for a discussion and comparison.

trade-off between worst-case bias and variance can be obtained by (1) regressing w_i on z_i using $\text{Pen}(\cdot)$ as the penalty function with weight λ and then (2) using the residuals from this regression as the instrument in the regression of Y_i on w_i . CIs based on these estimators can be constructed by using a critical value that incorporates the worst-case bias of the estimator, which we show can be obtained automatically as a byproduct of the regularized regression in step (1). These CIs are “bias-aware” as they account for the potential finite-sample bias of the estimator, and they are consequently valid in finite samples in the idealized Gaussian setting. We show how to choose the tuning parameter λ to optimize the MSE of the resulting estimator, or the length of the resulting CI.

We also consider the behavior of the bias-aware CI under high dimensional asymptotics where $k \gg n$, and $(w_i, z_i)'$ are independent over i with eigenvalues of the variance matrix bounded away from zero and infinity. We derive the rate at which the optimal CI shrinks in the case where $\text{Pen}(\gamma)$ is an ℓ_p norm. We show that, in the case where $k \gg n$ and C does not shrink with n , the optimal CI shrinks more slowly than $n^{-1/2}$, so that the bias term asymptotically dominates. Furthermore, we show that, in the ℓ_1 case, this rate cannot be improved even if one imposes the same ℓ_1 bound in the regression of w_i on z_i , as well as a certain degree of sparsity in both of these regressions.

As a key input for our approach, we require the researcher to explicitly specify the regularity parameter C bounding the magnitude of $\text{Pen}(\gamma)$. Our efficiency bounds show that it is impossible to automate the choice of C when forming CIs. We therefore recommend varying C as a form of sensitivity analysis and reporting a “breakdown” value given by the largest value of C such that a given finding, such as rejecting a particular null hypothesis, holds. We discuss how the choice of C can be guided by relating it to regression R^2 , and we present a lower CI for C that can be used as a specification check to ensure that the chosen value is not too low. As we discuss further in Section 5.2, CIs that do not choose regularity constants (such as C or, for sparsity-based approaches, the sparsity bound) explicitly involve implicit choices of these parameters. Our finite-sample approach has the advantage of making such choices explicit. This ensures that our coverage guarantees and efficiency bounds are not merely based on “asymptotic promises” about tuning parameters that may be hard to evaluate in a particular sample.

Our results relate to several strands of literature. Our procedures and efficiency bounds apply the general theory for linear functionals in convex Gaussian models developed in [Ibragimov and Khas'minskii \(1985\)](#), [Donoho \(1994\)](#), [Low \(1995\)](#) and [Armstrong and Kolesár \(2018\)](#). In particular, the optimal estimators are linear in outcomes, and the CIs are “bias-aware” fixed-length confidence interval (FLCI) centered at such estimators. Our results add to a growing recent literature applying this approach to various settings, including [Armstrong](#)

and Kolesár (2020a,b), Kolesár and Rothe (2018), Imbens and Wager (2019), Rambachan and Roth (2019), Noack and Rothe (2020), and Kwon and Kwon (2020). Muralidharan et al. (2020) apply the approach in the present paper to experiments with factorial designs and bounds on interaction effects.

The class of estimators we derive and, in particular, the idea of incorporating a regression of w_i on z_i to estimate β , is related to various estimators proposed for this problem, going back at least to the work of Robinson (1988) on the partly linear model. Our results provide a novel finite-sample justification for this idea, as well as an exact result giving the optimal form of this regression and the optimal estimator that incorporates it. Our results allow for a general form of $\text{Pen}(\cdot)$, which reduce to existing estimators in a few special cases: our results can in such cases be used to derive novel bias-aware CIs to accompany these estimators. The results of Li (1982) imply that the optimal estimator uses ridge regression when the penalty corresponds to an ℓ_2 norm. Li and Müller (2020) consider the weighted ℓ_2 norm $\text{Pen}(\gamma) = (\sum_{i=1}^n (z_i' \gamma)^2)^{1/2}$. They take a somewhat different approach, which leverages the particular invariance properties of this penalty function. Heckman (1988) derives optimal linear estimators in the partly linear model, where the penalty function bounds the first or second derivative of a univariate nonparametric regression function.

The problem of estimation and CI construction for β is distinct from the problem of estimation of the regression function itself or the entire parameter vector, using global loss. For the latter problem, see Zhang (2013) for the case where $p \leq 1$ (which overlaps with the class of bounds on γ that we consider when $p = 1$) and Shao and Deng (2012) for $p = 2$. These papers also differ from the present paper in focusing on asymptotic results.

The rest of this paper is organized as follows. Section 2 presents finite-sample results in the idealized model with Gaussian errors. Section 3 discusses implementation in the more realistic setting with unknown error distribution. Section 4 presents asymptotic characterizations of the efficiency bounds in the high dimensional setting under bounds on an ℓ_p norm. Section 5 compares our approach to CIs designed for sparsity constraints. Proofs and auxiliary results are in appendices.

2 Finite-sample results

This section sets up an idealized version of our model with Gaussian homoskedastic errors. We then show how to construct estimators and CIs in this model that are near-optimal in finite samples.

2.1 Setup

We write the model in eq. (1) in vector form as

$$Y = w\beta + Z\gamma + \varepsilon, \quad (2)$$

where $w = (w_1, \dots, w_n)' \in \mathbb{R}^n$ is the variable of interest with coefficient $\beta \in \mathbb{R}$ and $Z = (z'_1, \dots, z'_n)' \in \mathbb{R}^{n \times k}$ is a matrix of control variables. The design matrix $X = (w, Z)$ is treated as fixed. To obtain finite-sample results, we further assume that the errors are normal and homoskedastic,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad (3)$$

with σ^2 known. We discuss implementation with possibly heteroskedastic and non-Gaussian errors in Section 3. To make inference on β informative in settings where k is large relative to n (including the case where $k > n$), the researcher needs to make *a priori* restrictions on the control coefficients γ . We assume that these restrictions can be formalized by restricting the parameter space for $(\beta, \gamma)'$ to be $\mathbb{R} \times \Gamma$ where, for some linear subspace \mathcal{G} of \mathbb{R}^k ,

$$\Gamma = \Gamma(C) = \{\gamma \in \mathcal{G}: \text{Pen}(\gamma) \leq C\}, \quad \text{where } \text{Pen}(\cdot) \text{ is a seminorm on } \mathcal{G}. \quad (4)$$

The requirement that $\text{Pen}(\cdot)$ be a seminorm means that it satisfies the triangle inequality ($\text{Pen}(\gamma + \tilde{\gamma}) \leq \text{Pen}(\gamma) + \text{Pen}(\tilde{\gamma})$), and homogeneity ($\text{Pen}(c\gamma) = |c| \text{Pen}(\gamma)$ for any scalar c), but, unlike a norm, it is not necessarily positive definite ($\text{Pen}(\gamma) = 0$ does not imply $\gamma = 0$). This allows us to cover settings where only a subset of the control coefficients is restricted. To illustrate our methods, we focus on the case where $\text{Pen}(\cdot)$ corresponds to a weighted ℓ_p norm. To this end, partition the controls into a set of $k_1 \geq 0$ unrestricted baseline controls and a set of $k_2 = k - k_1$ additional controls, $Z = (Z_1, Z_2)$. Partition $\gamma = (\gamma'_1, \gamma'_2)'$ accordingly. Let H_A denote the projection matrix onto the column space of a matrix A .

Example 2.1 (ℓ_2 penalty). We specify the penalty as

$$\text{Pen}(\gamma) = \|M\gamma\|_2 = \sqrt{\gamma' M' M \gamma}, \quad (5)$$

where the $k_2 \times k$ matrix M incorporates scaling the variables and picking out which variables are to be constrained. If $M = (0, I_{k_2})$, then $\text{Pen}(\gamma) = \|\gamma_2\|_2$, with γ_1 unconstrained. Setting $M = (0, (Z'_2(I - H_{Z_1})Z_2/n)^{1/2})$ corresponds to the specification considered in [Li and Müller \(2020\)](#), which restricts the average of the squared mean effects $z'_{2i}\gamma_2$ on Y_i , after controlling for the baseline controls Z_1 . \square

Example 2.2 (ℓ_1 penalty). We replace the norm in eq. (5) with an ℓ_1 norm. For simplicity, we focus on the unweighted case, setting $M = (0, I_{k_2})$. This leads to $\text{Pen}(\gamma) = \|\gamma_2\|_1 = \sum_{j=k_1+1}^k |\gamma_j|$. \square

In addition to the choice of the penalty, the specification of Γ also requires the researcher to choose the regularity parameter C ; here we take it as given, and defer a discussion of this choice to Section 3.

While we have formulated the parameter space Γ in terms of a seminorm, this formulation is not restrictive in the sense that essentially any convex set Γ that is symmetric ($\gamma \in \Gamma$ implies $-\gamma \in \Gamma$) can be defined in this way (see Yosida, 1995, Proposition 5, p. 26). While we rule out non-convex constraints on Γ , such as sparsity, our results nonetheless have implications for such settings, as we discuss in Section 5.

Our goal is to construct estimators and CIs for β . To evaluate estimators $\hat{\beta}$ of β , we consider their worst-case performance over the parameter space $\mathbb{R} \times \Gamma$ under the MSE criterion,

$$R_{\text{MSE}}(\hat{\beta}; \Gamma) = \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} E_{\beta, \gamma} [(\hat{\beta} - \beta)^2],$$

where $E_{\beta, \gamma}$ denotes expectation under $(\beta, \gamma)'$. To evaluate CIs, we first require that they satisfy a coverage requirement. A $100 \cdot (1 - \alpha)\%$ CI with half-length $\hat{\chi} = \hat{\chi}(Y, X)$ is an interval $\{\hat{\beta} \pm \hat{\chi}\}$ that satisfies

$$\inf_{\beta \in \mathbb{R}, \gamma \in \Gamma} P_{\beta, \gamma} \left(\beta \in \{\hat{\beta} \pm \hat{\chi}\} \right) \geq 1 - \alpha,$$

where $P_{\beta, \gamma}$ denotes probability under $(\beta, \gamma)'$. To compare two CIs under a particular parameter vector $(\beta, \gamma)'$, we prefer the one with shorter expected length $E_{\beta, \gamma}[2\hat{\chi}]$. Note that optimizing expected length will not necessarily lead to CIs centered at an estimator $\hat{\beta}$ that is optimal under the MSE criterion.

2.2 Linear estimators CIs

We start by considering estimators that are linear in the outcomes Y , $\hat{\beta} = a'Y$, and we show how to construct CIs based on such estimators. The n -vector of weights a may depend on the design matrix X or the known variance σ^2 . In Section 2.3 below, we show how to choose the weights a optimally, and in Section 2.4 we show that when a is optimally chosen, the resulting estimators and CIs are optimal among all procedures, not just linear ones.

Under a given parameter vector $(\beta, \gamma)'$, the bias of $\hat{\beta} = a'Y$ is given by $a'(w\beta + Z\gamma) - \beta$. As $(\beta, \gamma)'$ ranges over the parameter space $\mathbb{R} \times \Gamma$, the bias ranges over the

set $[-\overline{\text{bias}}_{\Gamma}(\hat{\beta}), \overline{\text{bias}}_{\Gamma}(\hat{\beta})]$, where

$$\overline{\text{bias}}_{\Gamma}(\hat{\beta}) = \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} a'(w\beta + Z\gamma) - \beta \quad (6)$$

denotes the worst-case bias. The variance of $\hat{\beta}$ does not depend on $(\beta, \gamma)'$, and is given by $\text{var}(\hat{\beta}) = \sigma^2 a'a$.

To form a CI centered at $\hat{\beta}$, note that the t -statistic $(\hat{\beta} - \beta) / \text{var}(\hat{\beta})^{1/2}$ follows a $\mathcal{N}(b, 1)$ distribution where $|b| \leq \overline{\text{bias}}_{\Gamma}(\hat{\beta}) / \text{var}(\hat{\beta})^{1/2}$. Thus, denoting the $1 - \alpha$ quantile of a $|\mathcal{N}(B, 1)|$ distribution by $\text{cv}_{\alpha}(B)$, a two-sided CI can be formed as²

$$\hat{\beta} \pm \chi, \quad \text{where } \chi = \text{var}(\hat{\beta})^{1/2} \cdot \text{cv}_{\alpha} \left(\overline{\text{bias}}_{\Gamma}(\hat{\beta}) / \text{var}(\hat{\beta})^{1/2} \right). \quad (7)$$

We refer to this as a fixed-length confidence interval (FLCI), following the terminology in [Donoho \(1994\)](#), since its length 2χ is fixed: it depends only on the non-random design matrix X , and known variance σ^2 , but not on Y or the parameter vector $(\beta, \gamma)'$.

2.3 Optimal weights

Both the MSE $R(\hat{\beta}; \Gamma) = \overline{\text{bias}}_{\Gamma}(\hat{\beta})^2 + \text{var}(\hat{\beta})$ and the CI half-length χ given in eq. (7) are increasing in the variance of $\hat{\beta}$ and in its worst-case bias $\overline{\text{bias}}_{\Gamma}(\hat{\beta})$. Therefore, to find the optimal weights, it suffices to minimize variance subject to a bound B on worst-case bias, which we can write as

$$\min_{a \in \mathbb{R}} a'a \quad \text{s.t.} \quad \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} a'(w\beta + Z\gamma) - \beta \leq B. \quad (8)$$

We can then vary the bound B to find the optimal tradeoff for the given criterion (MSE or CI length). Since this optimization does not depend on the outcome data Y , optimizing the weights in this way does not affect the coverage properties of the resulting CI.

Our main computational result shows that the optimization problem in eq. (8) can be computed using regularized regression of w on Z . With slight abuse of terminology, we refer to this regression as a propensity score regression (even though we do not require w_i to be binary). To state the result, let π_{λ}^* denote the coefficient estimate on Z in a regularized propensity score regression with penalty $\text{Pen}(\pi)$,

$$\min_{\pi} \|w - Z\pi\|_2^2 \quad \text{s.t.} \quad \text{Pen}(\pi) \leq t_{\lambda}, \quad (9)$$

²The critical value $\text{cv}_{1-\alpha}(B)$ can be computed as the square root of the $1 - \alpha$ quantile of a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter B^2 .

where t_λ is a bound on the penalty term. Here, λ indexes the weight placed on the constraint in eq. (9). It can be the Lagrange multiplier in a Lagrangian formulation of (9), or we can solve (9) directly and take $t_\lambda = \lambda$.

Theorem 2.1. *Let π_λ^* be a solution to (9), and suppose that $\|w - Z\pi_\lambda^*\|_2 > 0$. Then $a_\lambda^* = \frac{w - Z\pi_\lambda^*}{(w - Z\pi_\lambda^*)'w}$ solves (8) with the bound given by $B = \frac{C}{t_\lambda} \cdot \frac{(w - Z\pi_\lambda^*)'Z\pi_\lambda^*}{(w - Z\pi_\lambda^*)'w}$. Consequently, the worst-case bias and variance of the estimator*

$$\hat{\beta}_\lambda = a_\lambda^{*'}Y = \frac{(w - Z\pi_\lambda^*)'Y}{(w - Z\pi_\lambda^*)'w} \quad (10)$$

are given by

$$\overline{\text{bias}}_\Gamma(\hat{\beta}_\lambda) = C\overline{B}_\lambda, \quad V_\lambda = \frac{\sigma^2\|w - Z\pi_\lambda^*\|_2^2}{[(w - Z\pi_\lambda^*)'w]^2}, \quad \text{where } \overline{B}_\lambda = \frac{1}{\text{Pen}(\pi_\lambda^*)} \frac{(w - Z\pi_\lambda^*)'Z\pi_\lambda^*}{(w - Z\pi_\lambda^*)'w}. \quad (11)$$

The result follows by applying the general theory of [Ibragimov and Khas'minskii \(1985\)](#), [Donoho \(1994\)](#), [Low \(1995\)](#), and [Armstrong and Kolesár \(2018\)](#) to our setting, which allows us to rewrite (8) as a convex optimization problem. Solving this convex problem then yields the result. Theorem 2.1 shows that the class of linear estimators that optimally trade off bias and variance (i.e. they solve eq. (8) for some B) can be obtained by a simple two-step procedure. In the first step, we estimate a penalized propensity score regression (9), indexed by the penalty term λ , with the penalty given by the penalty Pen that determines Γ . In the second step, we use the residuals $w - Z\pi_\lambda^*$ from this regression as instruments in a regression of Y on w . The penalties λ_{MSE}^* and λ_{FLCI}^* that yield linear estimators $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ and $\hat{\beta}_{\lambda_{\text{FLCI}}^*}$ that optimize the MSE criterion, and yield the shortest CI length (which for linear estimators is fixed; see eq. (7)), correspond to the solutions to the univariate optimization problems

$$\lambda_{\text{MSE}}^* = \underset{\lambda}{\text{argmin}} V_\lambda + (C\overline{B}_\lambda)^2, \quad \lambda_{\text{FLCI}}^* = \underset{\lambda}{\text{argmin}} \text{cv}_\alpha(C\overline{B}_\lambda/\sqrt{V_\lambda})\sqrt{V_\lambda}, \quad (12)$$

respectively, where V_λ and \overline{B}_λ are given in (11).

As $t_\lambda \rightarrow 0$, then, provided that $\text{Pen}(\cdot)$ is a norm on Z_2 , $\hat{\beta}_\lambda$ converges to the short regression estimate $\hat{\beta}_{\text{short}} = \frac{w'(I - H_{Z_1})Y}{w'(I - H_{Z_1})w}$ that only includes the unrestricted controls Z_1 . This estimator minimizes variance among all linear estimators with finite worst-case bias. In the other direction, as $t_\lambda \rightarrow \infty$, $\hat{\beta}_\lambda$ converges to the long regression estimate $\hat{\beta}_{\text{long}} = \frac{w'(I - H_Z)Y}{w'(I - H_Z)w}$, provided that w is not in the column space of Z (which ensures that the condition $\|w - Z\pi_\lambda^*\|_2 > 0$ in Theorem 2.1 holds for all λ). This estimator minimizes variance among all linear estimators that are unbiased, so Theorem 2.1 reduces to the Gauss-Markov theorem in this case. In other words, the short and long regressions are corner solutions of the bias-variance tradeoff,

in which weight is entirely placed on variance, or on bias.

Example 2.1 (ℓ_2 penalty, continued). In this case, a convenient Lagrangian formulation for (9) is

$$\pi_\lambda^* = \underset{\pi}{\operatorname{argmin}} \|w - Z\pi\|_2^2 + \lambda \|M\pi\|_2^2,$$

If $Z'Z + \lambda M'M$ is invertible³, taking first order conditions immediately leads to the closed form solution

$$\pi_\lambda^* = (Z'Z + \lambda M'M)^{-1} Z'w$$

which is a (generalized) ridge regression estimator of the propensity score.⁴ Simple algebra shows that

$$\hat{\beta}_\lambda = \frac{(w - Z\pi_\lambda^*)'Y}{(w - Z\pi_\lambda^*)'w} = e_1' \left(X'X + \lambda \begin{pmatrix} 0 & 0 \\ 0 & M'M \end{pmatrix} \right)^{-1} X'Y, \quad (13)$$

where $e_1 = (1, 0, \dots, 0)'$ is the first standard basis vector. Thus, the optimal estimate can also be obtained from a generalized ridge regression of Y onto X . The optimality of ridge regression in this setting was shown by Li (1982), and the above derivation gives this result as a special case of Theorem 2.1. If $M = (0, (Z_2'(I - H_{Z_1})Z_2/n)^{1/2})$, then the estimator further simplifies to a weighted average of the short and long regression estimates,

$$\hat{\beta}_\lambda = \omega(\lambda)\hat{\beta}_{\text{short}} + (1 - \omega(\lambda))\hat{\beta}_{\text{long}},$$

with weights

$$\omega(\lambda) = \frac{\lambda/n}{\lambda/n + \zeta^2}, \quad \zeta^2 = \frac{w'(I - H_Z)w}{w'(I - H_{Z_1})w} = \frac{\operatorname{var}(\hat{\beta}_{\text{short}})}{\operatorname{var}(\hat{\beta}_{\text{long}})}.$$

The weight on the short regression increases with λ (as the relative weight on variance in the bias-variance tradeoff increases), and decreases with ζ^2 . \square

Example 2.2 (ℓ_1 penalty, continued). In this case, the solution to (9) is given by a variant of the lasso estimate (Tibshirani, 1996) that only penalizes γ_2 .

The resulting estimator $\hat{\beta}_\lambda$ is related to estimators recently proposed for constructing CIs using the lasso (see, among others, Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Belloni et al., 2014). These papers propose estimators for β that combine lasso estimates from the outcome regression of Y onto X with lasso estimates

³This holds so long as no element $\pi \neq 0$ satisfies $Z\pi = 0$ and $M\pi = 0$ simultaneously. Intuitively, if Z has rank less than k , then the data is not informative about certain directions π , and we require the matrix M to place sufficient restrictions on π in these directions.

⁴The term ‘‘ridge regression’’ is sometimes reserved for the case where $M'M = I_k$. Here, we use the term to include generalizations such as this one.

from the propensity score regression, which yields an estimate that is non-linear in Y . In contrast, our estimator only uses lasso estimates for the propensity score regression, and is linear in Y . We give a detailed comparison between our estimator and this “double lasso” approach in Section 5. \square

Example 2.3 (Partly linear model). To flexibly control for a low-dimensional set of covariates \tilde{z}_i , one may specify a semiparametric model

$$y_i = w_i\beta + h(\tilde{z}_i) + \varepsilon_i, \quad \widetilde{\text{Pen}}(h) \leq \tilde{C},$$

where the penalty $\widetilde{\text{Pen}}(h)$ is a seminorm on functions $h(\cdot)$ that penalizes the “roughness” of h , such as the Hölder or Sobolev seminorm of order q . Minimax linear estimation in this model for particular choices of $\widetilde{\text{Pen}}(h)$ has been considered in Heckman (1988). This setting also falls directly into our setup by defining $Z = I_n$, $\gamma_i = h(\tilde{z}_i)$, and $\text{Pen}(\gamma) = \min_{h: h(\tilde{z}_i)=\gamma_i, i=1,\dots,n} \widetilde{\text{Pen}}(h)$ (assuming the minimum is taken). Theorem 2.1 then implies that the optimal estimator takes the form

$$\hat{\beta}_\lambda = \frac{\sum_{i=1}^n (w_i - g_\lambda^*(\tilde{z}_i)) Y_i}{\sum_{i=1}^n (w_i - g_\lambda^*(\tilde{z}_i)) w_i},$$

where $g_\lambda^*(\cdot)$ is analogous to the regularized regression estimate π_λ^* in (9): it solves

$$\min_g \sum_{i=1}^n (w_i - g(\tilde{z}_i))^2 \quad \text{s.t.} \quad \widetilde{\text{Pen}}(g) \leq t_\lambda.$$

When $\widetilde{\text{Pen}}$ is the Sobolev seminorm, this yields a spline estimate g_λ^* (see, for example Wahba, 1990). The partly linear model was treated in an influential paper by Robinson (1988), as well as earlier papers cited therein. Interestingly, the estimator proposed by Robinson (1988) takes a similar form to the estimator $\hat{\beta}_\lambda$, involving residuals from a nonparametric regression of w on \tilde{z}_i . While the analysis in Robinson (1988) is asymptotic, our results imply that a version of this estimator has sharp finite-sample optimality properties. \square

2.4 Efficiency among non-linear procedures

So, far we have restricted attention to procedures that are linear in the outcomes Y . We now show that the estimator $\hat{\beta}_{\lambda^*_{\text{MSE}}}$, and CIs based on the estimator $\hat{\beta}_{\lambda^*_{\text{FLCI}}}$ are in fact highly efficient among all procedures, not just linear ones. This is due to the fact that the parameter space Γ is convex and symmetric, and follows from the general results in Donoho (1994), Low (1995) and Armstrong and Kolesár (2018) for estimation of linear functionals in normal

models with convex parameter spaces.

Corollary 2.1. *Let λ_{MSE}^* and λ_{FLCI}^* be given in eq. (12), where the optimization is over all λ with $t_\lambda > 0$ such that $\|w - Z\pi_\lambda^*\|_2 > 0$. Let $\hat{\beta}_\lambda$, \overline{B}_λ and V_λ be given in (11). Let $\tilde{\beta}$ and $\tilde{\beta} \pm \tilde{\chi}$ denote some other (possibly non-linear) estimator and some other (possibly non-linear, variable-length) CI.*

(i) *For any λ , $\sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} \text{var}_{\beta, \gamma}(\tilde{\beta}) \leq V_\lambda$ implies $\overline{\text{bias}}_\Gamma(\tilde{\beta}) \geq C\overline{B}_\lambda$, and $\overline{\text{bias}}_\Gamma(\tilde{\beta}) \leq C\overline{B}_\lambda$ implies $\sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} \text{var}_{\beta, \gamma}(\tilde{\beta}) \geq V_\lambda$.*

(ii) *The worst-case MSE improvement of $\tilde{\beta}$ over $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ is bounded by*

$$\frac{\text{R}_{\text{MSE}}(\tilde{\beta})}{\text{R}_{\text{MSE}}(\hat{\beta}_{\lambda_{\text{MSE}}^*})} \geq \kappa_{\text{MSE}}^*(X, \sigma, \Gamma) \geq 0.8,$$

where $\kappa_{\text{MSE}}^*(X, \sigma, \Gamma)$ is given in Appendix A.2.

(iii) *The improvement of the expected length of the CI $\tilde{\beta} \pm \tilde{\chi}$ over the optimal linear FLCI $\hat{\beta}_{\lambda_{\text{FLCI}}^*} \pm \text{cv}_\alpha(C\overline{B}_{\lambda_{\text{FLCI}}^*}/V_{\lambda_{\text{FLCI}}^*}^{1/2})V_{\lambda_{\text{FLCI}}^*}^{1/2}$ at $\gamma = 0$ and any β is bounded by*

$$\frac{E_{\beta, 0}[\tilde{\chi}]}{\text{cv}_\alpha(C\overline{B}_{\lambda_{\text{FLCI}}^*}/V_{\lambda_{\text{FLCI}}^*}^{1/2})V_{\lambda_{\text{FLCI}}^*}^{1/2}} \geq \kappa_{\text{FLCI}}^*(X, \sigma, \Gamma),$$

where $\kappa_{\text{FLCI}}^*(X, \sigma, \Gamma)$ is given in Appendix A.2 and is at least 0.717 when $\alpha = 0.05$.

By construction, the estimator $\hat{\beta}_\lambda$ minimizes variance among all linear estimators with a bound $C\overline{B}_\lambda$ on the bias (or equivalently, it minimizes bias among all linear estimators with a bound V_λ on the variance). Corollary 2.1(i) shows that this optimality property is retained if we enlarge the class of estimators to all estimators, including non-linear ones. As a result, the minimax linear estimator $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ (i.e. the estimator attaining the lowest worst-case MSE in the class of linear estimators) continues to perform well among all estimators, including non-linear ones: by Corollary 2.1(ii), its worst-case MSE efficiency is at least 80%. The exact efficiency bound $\kappa_{\text{MSE}}^*(X, \sigma, \Gamma)$ depends on the design matrix, noise level, and particular choice of the parameter space, and can be computed explicitly in particular applications. We have found that typically the efficiency is considerably higher.

Finally, Corollary 2.1(iii) shows that it is not possible to substantively improve upon the FLCI based on $\hat{\beta}_{\lambda_{\text{FLCI}}^*}$ in terms of expected length when $\gamma = 0$, even if we consider variable length CIs that “direct power” at $\gamma = 0$ (potentially at the expense of longer expected length when $\gamma \neq 0$). The construction of the FLCI may appear conservative: its length depends on the worst-case bias over the parameter space for $(\beta, \gamma)'$, which, as the proof of

Theorem 2.1 shows, attains at $\gamma = Ct_{\lambda_{\text{FLCI}}^*}^{-1} \pi_{\lambda_{\text{FLCI}}^*}^*$, with $\text{Pen}(\gamma) = C$. Therefore, one may be concerned that when the magnitude of γ is much smaller than C , the FLCI is too long. Corollary 2.1(iii) shows that this is not the case, and the efficiency of the FLCI is at least 71.7% relative to variable-length CIs that optimize their expected length when $\gamma = 0$. The exact efficiency bound $\kappa_{\text{MSE}}^*(X, \sigma, \Gamma)$ can be computed explicitly in particular applications, and we have found that it is typically considerably higher than 71.7%.

A consequence of Corollary 2.1(iii) is that it is impossible to form a CI that is adaptive with respect to the regularity parameter C that bounds $\text{Pen}(\gamma)$. In the present setting, an adaptive CI would have length that automatically reflects the true regularity $\text{Pen}(\gamma)$ while maintaining coverage under a conservative a priori bound on $\text{Pen}(\gamma)$. However, according to Corollary 2.1(iii), any CI must have expected width that reflects the conservative a priori bound C rather than the true regularity $\text{Pen}(\gamma)$, even when $\text{Pen}(\gamma)$ is much smaller than the conservative a priori bound C . In particular, it is impossible to automate the choice of the regularity parameter C when forming a CI. We therefore recommend varying C as a form of sensitivity analysis, or using auxiliary information to choose C ; see Remark 3.3.

3 Implementation with non-Gaussian and heteroskedastic errors

We now discuss practical implementation issues, allowing ε to be non-Gaussian and heteroskedastic. As a baseline, we propose the following implementation:

Algorithm 3.1 (Baseline implementation).

Input Data (Y, X) , penalty $\text{Pen}(\cdot)$, regularity parameter C , and initial estimates of residuals

$$\hat{\varepsilon}_{\text{init},1}, \dots, \hat{\varepsilon}_{\text{init},n}.$$

Output Estimator and CI for β .

1. Compute an initial variance estimator, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_{\text{init},i}^2$, assuming homoskedasticity.
2. Compute the solution path $\{\pi_\lambda^*\}_{\lambda>0}$ for the regularized propensity score regression in eq. (9), indexed by the penalty weight λ . For each λ , compute $\hat{\beta}_\lambda$ as in eq. (10), and \bar{B}_λ , and V_λ as in eq. (11), with $\hat{\sigma}^2$ in place of σ^2 in the formula for V_λ .
3. Compute λ_{MSE}^* and λ_{FLCI}^* as in eq. (12), and compute the robust variance estimate $\hat{V}_{\lambda, \text{rob}} = \sum_{i=1}^n a_{\lambda,i}^*{}^2 \hat{\varepsilon}_{\text{init},i}^2$, where $a_\lambda^* = \frac{w - Z\pi_\lambda^*}{(w - Z\pi_\lambda^*)'w}$.

Return the estimator $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ and the CI $\hat{\beta}_{\lambda_{\text{FLCI}}^*} \pm \text{cv}_\alpha \left(C\bar{B}_{\lambda_{\text{FLCI}}^*} / \hat{V}_{\lambda_{\text{FLCI}}^*}^{1/2} \right) \cdot \hat{V}_{\lambda_{\text{FLCI}}^*}^{1/2}$. \square

Let us now discuss the implementation choices and the optimality and validity properties of the procedure in a series of remarks.

Remark 3.1 (Validity). As the initial residual estimates $\hat{\epsilon}_{\text{init},i}$, we can take residuals from a regularized outcome regression of Y on X . We give conditions for asymptotic validity of the resulting CIs in Appendix B.2. The key requirement is that the maximal Lindeberg weight $\text{Lind}(a_\lambda^*) = \max_{1 \leq i \leq n} a_{\lambda,i}^{*2} / \sum_{j=1}^n a_{\lambda,j}^{*2}$ associated with the estimator $\hat{\beta}_\lambda$ shrink quickly enough relative to error in the estimator used to form the residuals. Ensuring that $\text{Lind}(a_\lambda^*)$ is small prevents the estimator from putting too much weight on a particular observation, so that the Lindeberg condition for the central limit theorem holds.

Whether these conditions hold for the optimal estimator will in general depend on the form of $\text{Pen}(\gamma)$ and on the magnitude of C relative to n . To ensure that $\text{Lind}(a_\lambda^*)$ is small enough in a particular sample for a normal approximation to work well, one may impose a bound on this term by only minimizing eq. (12) over λ such that $\text{Lind}(a_\lambda^*)$ is small enough when computing λ_{FLCI}^* . This is similar to proposals by Noack and Rothe (2020), and Javanmard and Montanari (2014) in other settings. See Appendix B.2 for further discussion.

Remark 3.2 (Efficiency). The weights $a_{\lambda_{\text{FLCI}}^*}^*$ and $a_{\lambda_{\text{MSE}}^*}^*$ are not optimal under heteroskedasticity. One could in principle generalize the feasible generalized least squares (FGLS) approach used for unconstrained estimation by deriving optimal weights under the assumption $\epsilon \sim \mathcal{N}(0, \Sigma)$ (which simply follows the above analysis after pre-multiplying by $\Sigma^{-1/2}$), and derive conditions under which the estimator and CI that plug in an estimate of Σ are optimal asymptotically when the assumption of known variance and Gaussian errors is dropped. We instead generalize the common approach of reporting OLS with Eicker-Huber-White (EHW) standard errors in the unconstrained setting. The optimal weights a_λ^* are computed under the assumption of homoskedasticity, but we use a robust standard error to compute the CI to ensure its validity when this assumption is violated.

Remark 3.3 (Choice of C). By Corollary 2.1(iii), one cannot use a data-driven rule to automate the choice of C when forming a CI. We therefore recommend varying C as a form of sensitivity analysis, and reporting a “breakdown value” C^* as the largest value of C such that some empirical finding holds.

In settings where the plausible values of γ cannot be assessed using prior knowledge, one can relate the magnitude of $\text{Pen}(\gamma)$ to other quantities. One possibility is to run a regularized outcome regression of Y on X , with the constraint $\text{Pen}(\gamma) \leq C$ and report $R^2(C) = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_{i,C}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$ as a function of C , where $\{\hat{\epsilon}_{i,C}\}$ are the residuals from this regression, and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. The quantity $R^2(0)$ corresponds to the R^2 in the regression with only the baseline controls Z_1 included. One can then examine how $R^2(C)$ varies with C to relate bounds on

$\text{Pen}(\gamma)$ to R^2 . This mirrors the common practice in empirical applications in economics of examining how the magnitude of regression estimates and R^2 change when regressors are added (see [Oster, 2019](#), for further discussion and references). We note, however, that due to the impossibility result described above, additional assumptions would be needed to formally justify choosing C based on such a procedure.

Finally, one can form a lower CI $[\hat{C}, \infty)$ for C to assess the plausibility of a given bound on $\text{Pen}(\gamma)$. We present such a CI in [Appendix B.3](#) for the case where $\text{Pen}(\gamma)$ imposes an ℓ_p constraint. Such CIs can be useful as a specification check to ensure that the chosen value of the regularity parameter C is not too small.

Remark 3.4 (Computational issues). Step 2 involves computing the solution path of a regularized regression estimator. Efficient algorithms exist for computing these paths under ℓ_1 penalties and its variants ([Efron et al., 2004](#); [Rosset and Zhu, 2007](#)). Under ℓ_2 penalty, the regularized regression has a closed form, so that our algorithm can again be implemented in a computationally efficient manner. For other types of penalties, the convexity of optimization problem in [eq. \(9\)](#) can be exploited to yield efficient implementation. We also note that since the solution path π_λ^* does not depend on C , it only needs to be computed once, even when multiple choice of C are considered in a sensitivity analysis.

4 Rates of convergence

We now consider the asymptotic behavior of CIs and efficiency bounds as $n \rightarrow \infty$. For ease of notation, we assume all coefficients are constrained, and focus on the case $\text{Pen}(\gamma) = \|\gamma\|_p$ for some $p \geq 1$, and the case $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$ (see [Example 2.1](#)). We allow for sequences $C = C_n$ for the bound on $\text{Pen}(\gamma)$, which may go to 0 or ∞ with the sample size, as well as high dimensional asymptotics where $k = k_n \gg n$. We consider a standard “high dimensional” setting, placing conditions on the design matrix X that hold with high probability when w_i, z_i are drawn i.i.d. over i , with the eigenvalues of $\text{var}((w_i, z_i)')$ bounded away from zero and infinity.

Let $q \in [0, \infty]$ denote the Hölder conjugate of p , satisfying $1/p + 1/q = 1$. We will show that when $\text{Pen}(\gamma) = \|\gamma\|_p$, the optimal linear FLCI shrinks at the rate

$$n^{-1/2} + Cr_q(k, n) \quad \text{where} \quad r_q(k, n) = \begin{cases} k^{1/q}/\sqrt{n} & \text{if } q < \infty, \\ \sqrt{\log k}/\sqrt{n} & \text{if } q = \infty. \end{cases} \quad (14)$$

Furthermore, for $p = 1$ and $p = 2$, we will show that no other CI can shrink at a faster rate. For $p = 1$, we will in fact prove a stronger result showing that imposing sparsity bounds on

the outcome and propensity score regressions, in addition to the bound on $\text{Pen}(\gamma)$, does not help achieve a faster rate, unless one assumes sparsity of order greater than $C_n \sqrt{n/\log(k)}$ (termed the “ultra sparse” case in [Cai and Guo \(2017\)](#)). For the case $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$, we will show that the optimal rate is given by $n^{-1/2} + C$ when $k > n$.

In the case where $C = C_n$ does not decrease to zero with n , these rates require $p < 2$ (so that $q > 2$) for consistent estimation when $k/n \rightarrow \infty$. In the case where $p = 1$, we can then allow k to grow exponentially with n , whereas the cases $1 < p < 2$ allows for $k/n \rightarrow \infty$ with k growing at a polynomial rate in n that depends on p . Since taking $C_n \rightarrow 0$ rules out even a single coefficient being bounded away from zero, this suggests taking $p < 2$ in “high dimensional” settings, with $p = 1$ offering the best rate conditions. It also follows from these rate results that if $C_n = C$ does not decrease to zero with n , the bias term can dominate asymptotically, making it necessary to explicitly account for bias in CI construction even in large samples.

4.1 Upper bounds

To state the result, given $\eta > 0$, let $\mathcal{E}_n(\eta)$ denote the set of design matrices X for which there exists $\delta \in \mathbb{R}^k$ such that

$$\frac{1}{n} \|w - Z\delta\|_2^2 \leq \frac{1}{\eta}, \quad \frac{1}{n} w'(w - Z\delta) \geq \eta, \quad \frac{1}{n} \|Z'(w - Z\delta)\|_q \leq \frac{r_q(k, n)}{\eta}.$$

Let $R_{\text{FLCI}}^*(X, C) = 2 \text{cv}_\alpha(C \overline{B}_{\lambda_{\text{FLCI}}^*} / V_{\lambda_{\text{FLCI}}^*}^{1/2}) \cdot V_{\lambda_{\text{FLCI}}^*}^{1/2}$ denote the length of the optimal linear FLCI.

Theorem 4.1. (i) Suppose $\text{Pen}(\gamma) = \|\gamma\|_p$. There exists a finite constant K_η depending only on η such that $R_{\text{FLCI}}^*(X, C) \leq K_\eta n^{-1/2} (1 + Ck^{1/q})$ for $p > 1$, and $R_{\text{FLCI}}^*(X, C) \leq K_\eta n^{-1/2} (1 + C\sqrt{\log k})$ for $p = 1$ for any $X \in \mathcal{E}_n(\eta)$. (ii) Suppose $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$. There exists a finite constant K_η depending only on η such that $R_{\text{FLCI}}^*(X, C) \leq K_\eta (n^{-1/2} + C)$ for any X such that $\eta \leq w'w/n$.

The second part of the theorem follows since the short regression without any controls achieves a bias that is of the order C . The first part shows that the upper bounds on the rate of convergence match those in eq. (14) if the high-level condition $X \in \mathcal{E}_n(\eta)$ holds. The next lemma shows that this high-level condition holds with high probability when w_i, Z_i are drawn i.i.d. from a distribution satisfying mild conditions on moments and covariances.

Lemma 4.1. Suppose w_i, z_i are drawn i.i.d. over i , and let $\delta = \text{argmin}_b E[(w_i - z_i' b)^2]$ so that $z_i' \delta$ is the population best linear predictor error of w_i . Suppose that the linear prediction

error $E[(w_i - z_i'\delta)^2]$ is bounded away from zero as $k \rightarrow \infty$, $E[w_i^2] < \infty$, and that $\sup_j E[|(w_i - z_i'\delta)z_{ij}|^{\max\{2,q\}}] < \infty$ when $p > 1$, and, for some $c > 0$, $P(|(w_i - z_i'\delta)z_{ij}| \geq t) \leq 2 \exp(-ct^2)$ for all j when $p = 1$. Then, for any $\tilde{\eta} > 0$, there exists η such that $X \in \mathcal{E}_n(\eta)$ with probability at least $1 - \tilde{\eta}$ for large enough n .

4.2 Lower bounds

We now show that the rates in eq. (14) are sharp when $p = 2$, or $p = 1$.

4.2.1 $p = 2$

As with the upper bound in Section 4.1, we derive a bound that holds when the design matrix X is in some set, and then show that this set has high probability when w_i, z_i are drawn i.i.d. from a sequence of distributions satisfying certain conditions. We focus on the case $k \geq n$. Let $\tilde{\mathcal{E}}_n(\eta)$ denote the set of design matrices X such that

$$\eta \leq \frac{1}{n} w'w \leq \eta^{-1}, \quad \min \text{eig}(ZZ'/k) \geq \eta,$$

where $\text{eig}(A)$ denotes the set of eigenvalues of a square matrix A .

Theorem 4.2. *Let $\hat{\beta} \pm \hat{\chi}$ be a CI with coverage at least $1 - \alpha$ under $\text{Pen}(\gamma) \leq C$. (i) If $\text{Pen}(\gamma) = \|\gamma\|_2$, there exists a constant $c_\eta > 0$ depending only on η such that the expected length under $\beta = 0, \gamma = 0$ satisfies $E_{0,0}[\hat{\chi}] \geq c_\eta n^{-1/2}(1 + Ck^{1/2})$ for any $X \in \tilde{\mathcal{E}}_n(\eta)$. (ii) If $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$, there exists a constant $c_\eta > 0$ depending only on η such that the expected length under $\beta = 0, \gamma = 0$ satisfies $E_{0,0}[\hat{\chi}] \geq c_\eta n^{-1/2}(1 + C)$ for any $X \in \tilde{\mathcal{E}}_n(\eta)$.*

If z_i is i.i.d. over i , then EZZ'/k is equal to the $n \times n$ identity matrix times the scalar $\frac{1}{k} \sum_{j=1}^k E[z_{ij}^2]$. Thus, the condition on the minimum eigenvalue of ZZ'/k will hold under concentration conditions on the matrix $Z'Z$ so long as the second moments of the covariates are bounded from below. Here, we state a result for a special case where the z_{ij} 's are i.i.d. normal, which is immediate from Donoho (2006, Lemma 3.4).

Lemma 4.2. *Suppose that w_i are i.i.d. over i and that z_{ij} are i.i.d. normal over i and j . Then, for any $\tilde{\eta} > 0$, there exists $\eta > 0$ such that $X \in \tilde{\mathcal{E}}_n(\eta)$ with probability at least $1 - \tilde{\eta}$ once n and k/n are large enough.*

4.2.2 $p = 1$

We now consider the case where $p = 1$, as in Example 2.2. Rather than imposing conditions on X in a fixed design setting that hold with high probability (as in Section 4.1 and

Section 4.2.1), we directly consider a random design setting, and we do not condition on X when requiring coverage of CIs. This allows us to strengthen the conclusion of our theorem by showing that the rate in Theorem 4.1 is sharp even if one imposes a linear model for w_i given z_i along with sparsity and ℓ_1 bounds on the coefficients in this model.

We introduce some additional notation to cover the random design setting, which we use only in this section. We consider a random design model

$$\begin{aligned} Y &= w\beta + Z\gamma + \varepsilon, \quad \varepsilon \mid Z, w \sim \mathcal{N}(0, \sigma^2 I_n), \\ w &= Z\delta + v, \quad v \mid Z \sim \mathcal{N}(0, \sigma_v^2 I_n), \\ z_{ij} &\sim \mathcal{N}(0, 1) \quad \text{i.i.d. over } i, j. \end{aligned}$$

We use P_ϑ and E_ϑ for probability and expectation when Y, X follow this model with parameters $\vartheta = (\beta, \gamma', \delta', \sigma^2, \sigma_v^2)'$. Let $\sigma_0^2 > 0$ and $\sigma_{v,0}^2 > 0$ be given and let $\Theta(C, s, \eta)$ denote the set of parameters $\vartheta = (\beta, \gamma', \delta', \sigma^2, \sigma_v^2)$ where $|\sigma^2 - \sigma_0^2| \leq \eta$, $|\sigma_v^2 - \sigma_{v,0}^2| \leq \eta$, $\|\gamma\|_1 \leq C$, $\|\delta\|_1 \leq C$, $\|\gamma\|_0 \leq s$ and $\|\delta\|_0 \leq s$.

Theorem 4.3. *Let $\hat{\beta} \pm \hat{\chi}$ be a CI satisfying $P_\vartheta(\beta \in \{\hat{\beta} \pm \hat{\chi}\}) \geq 1 - \alpha$ for all ϑ in $\Theta(C_n, C_n \cdot K\sqrt{n/\log k}, \eta_n)$ where $\alpha < 1/2$. Suppose $k \rightarrow \infty$, $C_n\sqrt{\log k}/n \rightarrow 0$ and $C_n \leq \sqrt{k/n} \cdot k^{-\tilde{\eta}}$ for some $\tilde{\eta} > 0$. Then, there exists c such that, if K is large enough and $\eta_n \rightarrow 0$ slowly enough, the expected length of this CI under the parameter vector ϑ^* given by $\beta = 0$, $\gamma = 0$, $\delta = 0$, $\sigma^2 = \sigma_0^2$, $\sigma_v^2 = \sigma_{v,0}^2$ satisfies $E_{\vartheta^*}[\hat{\chi}] \geq c \cdot n^{-1/2}(1 + C_n\sqrt{\log k})$ once n is large enough.*

Theorem 4.3 follows from similar arguments to Cai and Guo (2017) and Javanmard and Montanari (2018), who provide similar bounds for the case where only a sparsity bound is imposed. According to Theorem 4.3, imposing sparsity does not allow one to improve upon the CIs that uses only the ℓ_1 bound $\|\gamma\|_1 \leq C_n$ (thereby attaining the rate in Theorem 4.1), unless one imposes sparsity of order greater than $C_n\sqrt{n/\log k}$. We provide further comparison with CIs that impose sparsity in the next section.

5 Comparison with sparsity constraints

Several authors have considered CIs for β using “double lasso” estimators (see, among others, Belloni et al., 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014). These CIs are valid under the parameter space

$$\tilde{\Gamma}(s) = \{\gamma: \|\gamma\|_0 \leq s\}, \tag{15}$$

where $\|\gamma\|_0 = \#\{j: \gamma_j \neq 0\}$ is the ℓ_0 “norm,” which indexes the sparsity of γ , and with s increasing slowly enough relative to n and k . Since $\|\gamma\|_0$ is not a true norm or seminorm (it is non-convex), this falls outside our setting. Here, we discuss some connections with the optimal estimators we derive under ℓ_1 constraints with these double lasso estimators (Section 5.1), and we provide a discussion comparing our approach to CIs based on these estimators (Section 5.2).

5.1 Connection between double lasso and optimal estimator under ℓ_1 constraints

In the case where $\text{Pen}(\gamma) = \|\gamma\|_1$ (example 2.2), the solution π_λ^* to (9) is the lasso estimate in the propensity score regression of w on Z , and our estimator (10) uses residuals from this lasso regression. This is related to recently proposed “double lasso” estimators used to form CIs for β under sparsity constraints on γ (see, among others, Belloni et al., 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014). For concreteness, we focus on the estimator in Zhang and Zhang (2014), which is given by

$$\hat{\beta}_{ZZ} = \hat{\beta}_{\text{lasso}} + \frac{(w - Z\pi_\lambda^*)'(Y - w\hat{\beta}_{\text{lasso}} - Z\hat{\gamma}_{\text{lasso}})}{(w - Z\pi_\lambda^*)w},$$

where $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$ are the lasso estimates from regressing Y on X :

$$\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}} = \underset{\beta, \gamma}{\text{argmin}} \|Y - w\beta - Z\gamma\|_2^2 + \tilde{\lambda}(|\beta| + \|\gamma\|_1)$$

for some penalty parameter $\tilde{\lambda} > 0$.

Remark 5.1. Note that $\hat{\beta}_{ZZ}$ is non-linear in Y , due to non-linearity of the lasso estimates $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$, which is consistent with the goal of efficiency in the non-convex parameter space (15). In contrast, Corollary 2.1 shows that under the convex parameter space $\Gamma = \{\gamma: \|\gamma\|_1 \leq C\}$, the estimator $\hat{\beta}_\lambda$ in (10) which only uses lasso in the propensity score regression of w on Z , is already highly efficient among all estimators, so that there is no further role for substantive efficiency gains from the lasso regression of Y on X , or from the use of other non-linear estimators.

To further understand the connection between these estimators, we note that Zhang and Zhang (2014) motivate their approach by bounds of the form

$$\|\hat{\gamma}_{\text{lasso}} - \gamma\|_1 \leq \tilde{C} \quad \text{where} \quad \tilde{C} = \text{const.} \cdot s\sqrt{\log k}/\sqrt{n}, \quad (16)$$

which hold with high probability with the constant depending on certain “compatibility constants” that describe the regularity of the design matrix X (see [Bühlmann and van de Geer, 2011](#), Theorem 6.1, and references in the surrounding discussion). This suggests correcting the initial estimate $\hat{\beta}_{\text{lasso}}$ by estimating $\tilde{\beta} = \beta - \hat{\beta}_{\text{lasso}}$ in the regression

$$\tilde{Y} = w(\beta - \hat{\beta}_{\text{lasso}}) + Z(\gamma - \hat{\gamma}_{\text{lasso}}) + \varepsilon = w\tilde{\beta} + Z\tilde{\gamma} + \varepsilon,$$

where $\tilde{Y} = Y - \hat{\beta}_{\text{lasso}} - Z\hat{\gamma}_{\text{lasso}}$. Heuristically, we can treat the bound (16) as a constraint $\|\tilde{\gamma}\|_1 \leq \tilde{C}$ on the unknown parameter $\tilde{\gamma} = \gamma - \hat{\gamma}_{\text{lasso}}$ and search for an optimal estimator of $\tilde{\beta} = (\beta - \hat{\beta}_{\text{lasso}})$ under this constraint. Applying the optimal estimator derived in Theorem 2.1 then suggests estimating $\beta - \hat{\beta}_{\text{lasso}}$ with $\frac{(w - Z\pi_\lambda^*)'\tilde{Y}}{(w - Z\pi_\lambda^*)'w}$. Adding this estimate to $\hat{\beta}_{\text{lasso}}$ gives the estimate $\hat{\beta}_{\text{ZZ}}$ proposed by [Zhang and Zhang \(2014\)](#). Whereas [Zhang and Zhang \(2014\)](#) motivate their approach as one possible way of correcting the initial estimate $\hat{\beta}_{\text{lasso}}$ using the bound (16), the above analysis shows that their correction is in fact identical to an approach in which one optimizes this correction numerically.⁵

Using the bound (16) it follows that $\hat{\beta}_{\text{ZZ}} - \beta = \tilde{b} + a_\lambda^{*\prime}\varepsilon$ where $a_\lambda^* = \frac{(w - Z\pi_\lambda^*)}{(w - Z\pi_\lambda^*)'w}$ are the optimal weights under the ℓ_1 constraint $\|\tilde{\gamma}\|_1 \leq \tilde{C}$, given in Theorem 2.1. Furthermore, $|\tilde{b}| \leq \tilde{C}\bar{B}_\lambda$, with \bar{B}_λ given in Theorem 2.1 and \tilde{C} given in (16), and the variance of the random term $a_\lambda^{*\prime}\varepsilon$ is given by V_λ in Theorem 2.1. Using arguments similar to those used to prove Theorem 4.1, it follows that $\tilde{C}\bar{B}_\lambda/\sqrt{V_\lambda}$ is bounded by a constant times $s(\log k)/\sqrt{n}$, so that one can ignore bias in large samples as long as this term converges to zero. This leads to the CI proposed by [Zhang and Zhang \(2014\)](#), which takes the form

$$\{\hat{\beta}_{\text{ZZ}} \pm z_{1-\alpha/2}\hat{V}_\lambda^{1/2}\}, \tag{17}$$

where \hat{V}_λ is an estimate of the variance V_λ . We use the term “double lasso CI” to refer to this CI, and to related CIs such as those proposed in [Belloni et al. \(2014\)](#); [Javanmard and Montanari \(2014\)](#); [van de Geer et al. \(2014\)](#).

Remark 5.2. To avoid having to assume that $s(\log k)/\sqrt{n} \rightarrow 0$ one could, in principle,

⁵The estimator proposed by [Javanmard and Montanari \(2014\)](#) performs a numerical optimization of this form, but with the constraint (16) replaced by a constraint on $|\hat{\beta}_{\text{lasso}} - \beta| + \|\hat{\gamma}_{\text{lasso}} - \gamma\|_1$. Thus, Theorem 2.1 shows that a modification of the constraint used in [Javanmard and Montanari \(2014\)](#) yields the same estimator as [Zhang and Zhang \(2014\)](#).

extend our approach and the above analysis to form valid bias-aware CIs as⁶

$$\{\hat{\beta}_{\text{ZZ}} \pm [\tilde{C}\bar{B}_\lambda + z_{1-\alpha/2}\hat{V}_\lambda^{1/2}]\}$$

Unfortunately, finding a computable constant \tilde{C} in (16) that is sharp enough to yield useful bounds in practice appears to be difficult, although it is an interesting area for future research.

5.2 Comparison of our approach with CIs based on double lasso estimators

When should one use a double lasso CI, and when should one use the approach in the present paper? In principle, this depends on the a priori assumptions one is willing to make, and whether they are best captured by a sparsity bound or a bound on convex penalty function, such as the ℓ_1 or ℓ_2 norm. In many settings, it may be difficult to motivate the assumption that a regression function has a sparse approximation, whereas upper bounds on the magnitude of the coefficients may be more plausible.

A key advantage of the CIs and estimators we propose is that they have sharp finite-sample optimality properties and coverage guarantees in the fixed design Gaussian model with known error variance. While this is an idealized setting, the worst-case bias calculations do not depend on the error distribution, and remain the same under non-Gaussian, heteroskedastic errors. Our approach directly accounts for the potential finite-sample bias of the estimator, rather than relying on “asymptotic promises” about rates at which certain constants involved in bias terms converge to zero.

A flip side of this approach is that our CIs require an explicit choice of the regularity parameter C in order to form a “bias-aware” CI. In contrast, CIs based on double lasso estimators do not require explicitly choosing the regularity (in this case, the sparsity s), since they ignore bias. This is justified under asymptotics in which s increases more slowly than $\sqrt{n}/\log k$, which lead to the bias of $\hat{\beta}_{\text{ZZ}}$ decreasing more quickly than its standard deviation. Thus, one can say that the CI in eq. (17) is “asymptotically valid” without explicitly specifying the sparsity index s : one need only make an “asymptotic promise” that s increases slowly enough. However, such asymptotic promises are difficult to evaluate in a given finite-sample setting. Indeed, shown in, for example, [Li and Müller \(2020\)](#), the double lasso CI leads to undercoverage in finite samples even in relatively sparse settings. To ensure good finite-sample coverage of the CI in eq. (17), one needs to ensure that the actual finite-

⁶We use the slightly more conservative approach of adding and subtracting the bound $\tilde{C}\bar{B}_\lambda$ rather than using the critical value $\text{cv}_\alpha(\tilde{C}\bar{B}_\lambda/\hat{V}_\lambda^{1/2})$ as in eq. (7), since the “bias” term for $\hat{\beta}_{\text{ZZ}}$ is correlated with ε through the first step estimates $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$.

sample bias is negligible relative to the standard deviation of the estimator. Since any bias bound will depend on the sparsity index s (as in the bound in eq. (16)), this gets us back to having to specify s .

Thus, CIs that ignore bias such as conventional CIs based on double lasso estimators do not avoid the problem of specifying s or C : they merely make such choices implicit in asymptotic promises. These issues show up formally in the asymptotic analysis of such CIs. In particular, double lasso CIs require the “ultra sparse” asymptotic regime $s = o(\sqrt{n}/\log k)$, and they undercover asymptotically in the “moderately sparse” regime where s increases more slowly than n with $s \gg \sqrt{n}/\log k$. Indeed, Theorem 4.3 above, as well as the results of Cai and Guo (2017) and Javanmard and Montanari (2018) show that it is impossible to avoid explicitly specifying s if one allows for the moderately sparse regime. On the other end of the spectrum, in the “low dimensional” regime where $k \ll n$, the double lasso CI is asymptotically equivalent to the usual CI based on the long regression. Thus, the double lasso CI cannot be used when the goal is to use a priori information on γ to improve upon the CI based on the long regression (as in, for example, Muralidharan et al., 2020), even if s is small enough that such improvements would be warranted with prior knowledge of s . In contrast, our approach optimally incorporates the bound C regardless of the asymptotic regime.

Appendix A Proofs

This Appendix gives proofs for all results in the main text.

A.1 Proof of Theorem 2.1

To prove Theorem 2.1, we first explain how our results fall into the general setup used in Donoho (1994), Low (1995) and Armstrong and Kolesár (2018). In the notation of Armstrong and Kolesár (2018), $(\beta, \gamma)'$ plays the role of the parameter f , the functional of interest is given by $L(\beta, \gamma)' = \beta$ and $K(\beta, \gamma)' = w\beta + Z\gamma$. The parameter space $\mathbb{R} \times \Gamma$ is centrosymmetric, so that the modulus of continuity (eq. (25) in Armstrong and Kolesár, 2018) is given by

$$\omega(\delta) = \sup_{\beta, \gamma} 2\beta \quad \text{s.t.} \quad \|w\beta + Z\gamma\|_2 \leq \delta/2, \quad \text{Pen}(\gamma) \leq C.$$

Using the substitution $\pi = -\gamma/\beta$, we can write this as

$$\omega(\delta) = \sup_{\beta, \pi} 2\beta \quad \text{s.t.} \quad \beta\|w - Z\pi\|_2 \leq \delta/2, \quad \beta \text{Pen}(\pi) \leq C. \quad (18)$$

Let $\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}}$ and $\pi_\delta^{\text{mod}} = -\gamma_\delta^{\text{mod}}/\beta_\delta^{\text{mod}}$ denote a solution to this problem when it exists. In the notation of Armstrong and Kolesár (2018), $(\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}})'$ plays the role of g_δ^* , and the solution (f_δ^*, g_δ^*) satisfies $f_\delta^* = -g_\delta^* = -(\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}})'$ by centrosymmetry.

This optimization problem is clearly related to the problem in eq. (9): we want to make $\|w - Z\pi\|_2$ and $\text{Pen}(\pi)$ small so that large values of β satisfy the constraint in (18). The following lemma formalizes the connection.

Lemma A.1. *If there exists $\pi \in \mathcal{G}$ such that $w = Z\pi$ and $\text{Pen}(\pi) = 0$, then $\omega(\delta) = \infty$ for all $\delta \geq 0$. Otherwise, (i) for any $\delta > 0$, the modulus problem (18) has a solution $\beta_\delta^{\text{mod}}, \pi_\delta^{\text{mod}}$ with $\beta_\delta^{\text{mod}} > 0$. For $t_\lambda = C/\beta_\delta^{\text{mod}} = 2C/\omega(\delta)$, this solution π_δ^{mod} is also a solution to the penalized regression (9) with optimized objective $\|w - Z\pi_\delta^{\text{mod}}\|_2 = \delta/(2\beta_\delta^{\text{mod}}) = \delta/\omega(\delta) > 0$; and (ii) for any $t_\lambda > 0$, the penalized regression problem (9) has a solution π_λ^* . Setting $\beta_\lambda^* = C/t_\lambda$ and $\delta_\lambda = 2\beta_\lambda^*\|w - Z\pi_\lambda^*\|_2 = (2C/t_\lambda)\|w - Z\pi_\lambda^*\|_2$, the pair $\beta_\lambda^*, \pi_\lambda^*$ solves the modulus problem (18) at $\delta = \delta_\lambda$, with optimized objective $\omega(\delta_\lambda) = 2C/t_\lambda$, so long as $\|w - Z\pi_\lambda^*\|_2 > 0$.*

Proof. If there exists $\pi \in \mathcal{G}$ such that $w = Z\pi$ and $\text{Pen}(\pi) = 0$, then the result is immediate. Suppose there does not exist such a π .

First, we show that the problem (9) has a solution. Let $\mathcal{G}^{(0)}$ denote the linear subspace of vectors $\pi \in \mathcal{G}$ such that $Z\pi = 0$ and $\text{Pen}(\pi) = 0$, and let $\mathcal{G}^{(1)}$ be a subspace such that $\mathcal{G} = \mathcal{G}^{(0)} \oplus \mathcal{G}^{(1)}$, so that we can write $\pi \in \mathcal{G}$ uniquely as $\pi = \pi^{(0)} + \pi^{(1)}$ where $\pi^{(0)} \in$

$\mathcal{G}^{(0)}$ and $\pi^{(1)} \in \mathcal{G}^{(1)}$. Note that $Z\pi = Z\pi^{(1)}$ and, applying the triangle inequality twice, $\text{Pen}(\pi^{(1)}) = \text{Pen}(\pi^{(1)}) - \text{Pen}(-\pi^{(0)}) \leq \text{Pen}(\pi) \leq \text{Pen}(\pi^{(0)}) + \text{Pen}(\pi^{(1)}) = \text{Pen}(\pi^{(1)})$ so that $\text{Pen}(\pi) = \text{Pen}(\pi^{(1)})$. Thus, the problem (9) can be written in terms of $\pi^{(1)} \in \mathcal{G}^{(1)}$ only. The level sets of this optimization problem are bounded and are closed by continuity of the seminorm $\text{Pen}(\cdot)$ (Goldberg, 2017), and so it has a solution, which is also a solution in the original problem. Similarly, to show that the problem (18) has a solution, note that feasible values of β are bounded by a constant times the inverse of the minimum of $\max\{\|w - Z\pi\|_2, \text{Pen}(\pi)\}$ over π , which is strictly positive by continuity of $\text{Pen}(\pi)$ and the fact that there does not exist π with $\max\{\|w - Z\pi\|_2, \text{Pen}(\pi)\} = 0$. Thus, we can restrict $\beta, \tilde{\pi}^{(1)}$ to a compact set without changing the optimization problem.

To show the first statement in the lemma, note that $\beta_\delta^{\text{mod}} > 0$, since it is feasible to set $\pi = 0$ and $\beta = \delta/(2\|w\|_2)$, and that $\|w - Z\pi_\delta^{\text{mod}}\|_2 > 0$, since otherwise a strictly larger value of β could be achieved by multiplying π_δ^{mod} by $1 - \eta$ for $\eta > 0$ small enough. Now, if the first statement did not hold, there would exist a $\tilde{\pi}$ with $\text{Pen}(\tilde{\pi}) \leq C/\beta_\delta^{\text{mod}}$ such that $\|w - Z\tilde{\pi}\|_2 \leq \|w - Z\pi_\delta^{\text{mod}}\|_2 - \nu$ for small enough $\nu > 0$. Then, letting $\tilde{\pi}_\eta = (1 - \eta)\tilde{\pi}$, we would have $\|w - Z\tilde{\pi}_\eta\|_2 \leq \|w - Z\tilde{\pi}\|_2 + \eta\|Z\tilde{\pi}\|_2 \leq \|w - Z\pi_\delta^{\text{mod}}\|_2 - \nu + \eta\|Z\tilde{\pi}\|_2 \leq \delta/(2\beta_\delta^{\text{mod}}) - \nu + \eta\|Z\tilde{\pi}\|_2$. Thus, for small enough η , $\|w - Z\tilde{\pi}_\eta\|_2$ will be strictly less than $\delta/(2\beta_\delta^{\text{mod}})$ for small enough η and $\text{Pen}(\tilde{\pi}_\eta) \leq (1 - \eta)C/\beta_\delta^{\text{mod}} < C/\beta_\delta^{\text{mod}}$. This is a contradiction, since it would allow a strictly larger value of β by setting $\pi = \tilde{\pi}_\eta$.

The second statement follows immediately, since any pair $\tilde{\beta}, \tilde{\pi}$ satisfying the constraints in the modulus (18) for $\delta = \delta_\lambda$ with $\tilde{\beta} > \beta_\lambda^*$ would have to have $\|w - Z\tilde{\pi}\|_2 < \|w - Z\pi_\lambda^*\|_2$ while maintaining the constraint $\text{Pen}(\pi_\lambda^*) \leq t_\lambda$. \square

We now prove Theorem 2.1. The class of bias-variance optimizing estimators, \hat{L}_δ in the notation of Armstrong and Kolesár (2018), is given by $\frac{(w\beta_\delta^{\text{mod}} + Z\gamma_\delta^{\text{mod}})'Y}{(w\beta_\delta^{\text{mod}} + Z\gamma_\delta^{\text{mod}})'w}$, where we use eq. (26) in Armstrong and Kolesár (2018) to compute the form of this estimator under centrosymmetry, and Lemma D.1 in Armstrong and Kolesár (2018) to calculate the derivative $\omega'(\delta)$, since the problem is translation invariant with ι given by the parameter $\beta = 1, \gamma = 0$. Given λ with $\|w - Z\pi_\lambda^*\|_2 > 0$, it follows from Lemma A.1 that, for δ_λ given in the lemma, this estimator \hat{L}_{δ_λ} is equal to $\hat{\beta}_\lambda = a_\lambda^*{}'Y$ where $a_\lambda^* = \frac{w - Z\pi_\lambda^*}{(w - Z\pi_\lambda^*)'w}$, as defined in Theorem 2.1. The worst-case bias formula in Theorem 2.1 then follows from the fact that the maximum bias is attained at $\gamma = -\gamma_{\delta_\lambda}^{\text{mod}} = Ct_\lambda^{-1}\pi_\lambda^*$ by Lemma A.1 in Armstrong and Kolesár (2018) (or Lemma 4 in Donoho, 1994).

A.2 Proof of Corollary 2.1

Part (i) of Corollary 2.1 follows from Low (1995). In particular, consider the one-dimensional submodel $\beta \in [-C/t_\lambda, C/t_\lambda]$, $\gamma = -\pi_\lambda^* \beta$. Let $b_\lambda = (w - Z\pi_\lambda^*)/\|w - Z\pi_\lambda^*\|_2^2$, and let $B \in \mathbb{R}^{(n-1) \times n}$ be an orthogonal matrix that's orthogonal to b_λ . Note that in this submodel, $B'Y = B'(w - Z\pi_\lambda^*)\beta + B'\varepsilon = B'\varepsilon$, which does not depend on the unknown parameter β , and is independent of $b_\lambda'Y$. Therefore, $b_\lambda'Y \sim \mathcal{N}(\beta, \|b_\lambda\|_2^2 \sigma^2)$ is a sufficient statistic in this submodel. By Theorem 1 in Low (1995), in this submodel, the estimator $\hat{\beta}_\lambda = a_\lambda^{*'}Y = \kappa b_\lambda'Y$, where $\kappa = \|w - Z\pi_\lambda^*\|_2^2 / (w - Z\pi_\lambda^*)'w$ minimizes $\sup_\beta \text{var}(\delta(Y))$ among all estimators $\delta(Y)$ with $\sup_\beta |E_\beta[\delta(Y)] - \beta| \leq (1 - \kappa)C/t_\lambda = C\bar{B}_\lambda$, and, likewise, it minimizes $\sup_\beta |E_\beta[\delta(Y)] - \beta|$ among all estimators with $\sup_\beta \text{var}(\delta(Y)) \leq \kappa^2 \sigma^2 \|b_\lambda\|_2^2 = V_\lambda$. Since the worst-case bias $\overline{\text{bias}}_\Gamma(\hat{\beta}_\lambda) \leq C\bar{B}_\lambda$ and variance $(\hat{\beta}_\lambda) = V_\lambda$ are the same in the full model by Theorem 2.1, the result follows.

Part (ii) of Corollary 2.1 is immediate from Donoho (1994). In particular, it holds with

$$\kappa_{\text{MSE}}^*(X, \sigma, \Gamma) = \frac{\sup_{\delta > 0} (\omega(\delta)/\delta)^2 \rho_N(\delta/2, \sigma)}{\sup_{\delta > 0} (\omega(\delta)/\delta)^2 \rho_A(\delta/2, \sigma)} \geq 0.8,$$

where $\omega(\delta)$ is defined in eq. (18), and ρ_A and ρ_N are the minimax risk among affine estimators, and among all estimators, respectively, in the bounded normal means problem $Y \sim \mathcal{N}(\theta, \sigma^2)$, $|\theta| \leq \tau$, defined in Donoho (1994), and the last inequality follows from eq. (4) in Donoho (1994).

Finally, Part (iii) of Corollary 2.1 follows from Corollary 3.3 in Armstrong and Kolesár (2018), with

$$\kappa_{\text{FLCI}}^*(X, \sigma, \Gamma) = \frac{(1 - \alpha)E[\omega(2(z_{1-\alpha} - Z)) \mid Z \leq z_{1-\alpha}]}{2 \min_\delta \text{cv}_\alpha \left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2} \right) \omega'(\delta)},$$

where $Z \sim \mathcal{N}(0, 1)$, $\omega(\delta)$ is given in eq. (18), and by Lemma D.1 in Armstrong and Kolesár, since the problem is translation invariant with ι given by the parameter $\beta = 1$, $\gamma = 0$, $\omega'(\delta) = \delta/[w'(w - Z\pi_\delta^{\text{mod}}) \cdot \omega(\delta)]$. The universal lower bound 0.717 when $\alpha = 0.05$ follows from Theorem 4.1 in Armstrong and Kolesár (2020b).

A.3 Proof of Theorem 4.1

To prove that the claimed upper bound holds for $X \in \mathcal{E}_n(\eta)$, we first note that, since the FLCI based on $\hat{\beta}_{\text{FLCI}}^*$ is shorter than the FLCI based on any linear estimator $a'Y$, it suffices to show that there exists a sequence of weight vectors a such that the worst-case bias and standard deviation are bounded by constants times $n^{-1/2}(1 + Ck^{1/q})$ when $p > 1$ or $n^{-1/2}(1 + C\sqrt{\log k})$ when $p = 1$. We consider the weights $\tilde{a}_i = \frac{v_i}{\sum_{j=1}^n v_j w_j}$, where $v_i = w_i - z_i' \delta$, with δ given in the

definition of $\mathcal{E}(\eta)$. The variance of the estimator $\tilde{a}'Y$ is $\frac{\sum_{i=1}^n v_i^2}{(\sum_{i=1}^n v_i w_i)^2} \leq \eta^{-3}/n$. The worst-case bias is

$$\sup_{\gamma: \|\gamma\|_p \leq C} \tilde{a}'Z\gamma = C\|Z'\tilde{a}\|_q = n^{-1/2}C \frac{n^{-1/2}\|Z'(w - Z\delta)\|_q}{n^{-1}|w'(w - Z\delta)|} \leq C \frac{r_q(k, n)}{\eta^2},$$

where the first equality follows by Hölder's inequality, and the last quality follows by definition of $\mathcal{E}_n(\eta)$. This yields the convergence rate $n^{-1/2} + Cr_q(k, n)$, as claimed. For part (ii), by analogous reasoning, it suffices to consider the short regression estimator $\hat{\beta}_0 = w'Y/w'w$. The variance of this estimator is $\sigma^2/w'w \leq \eta^{-1}\sigma^2/n$. The bias of the estimator is $w'Z\gamma/w'w$. By the Cauchy-Schwarz inequality, this quantity is bounded in absolute value by $\|w/w'w\|_2\|Z\gamma\|_2 = \|Z\gamma/\sqrt{n}\|_2/\sqrt{w'w/n} \leq \eta^{-1/2}C$. This yields the desired convergence rate.

A.4 Proof of Lemma 4.1

By the orthogonality condition for the best linear predictor, we have $E[w_i v_i] = E[v_i^2]$, where $v_i = w_i - z_i'\delta$, which is bounded from below uniformly over k by assumption. Since $E[w_i v_i]$ is bounded from above by $Ew_i^2 < \infty$, it follows from the law of large numbers for triangular arrays that $\frac{1}{n} \sum_{i=1}^n w_i v_i \geq \eta$ with probability approaching one once η is small enough. Similarly, $\frac{1}{n} \sum_{i=1}^n v_i^2 \leq 1/\eta$ for large enough η by the law of large numbers for triangular arrays.

For the last inequality in the definition of $\mathcal{E}_n(\eta)$, first consider the case $p > 1$ so that $q < \infty$. We then have $E\|\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i v_i\|_q^q = E \sum_{j=1}^k |\sum_{i=1}^n v_i z_{ij}/\sqrt{n}|^q \leq k \cdot K$ by [von Bahr \(1965\)](#), where K is a constant that depends only on an upper bound for $\max_j E[|v_i z_{ij}|^{\max\{q, 2\}}]$. Applying Markov's inequality gives the required bound. When $p = 1$, then $q = \infty$ so that

$$P \left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i v_i \right\|_q \geq \eta^{-1} \sqrt{\log k} \right) \leq \sum_{j=1}^k P \left(\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n v_i z_{ij} \right| > \eta^{-1} \sqrt{\log k} \right),$$

which is bounded by $2k \exp(-K \cdot \eta^{-2} \log k) = 2k^{1-K\eta^{-2}}$ for some constant K by Hoeffding's inequality for sub-Gaussian random variables ([Vershynin, 2018](#), Theorem 2.6.3). This can be made arbitrarily small uniformly in k by making η small, as required.

A.5 Proof of Theorem 4.2

By Corollary 2.1(iii), it suffices to show the bound for $R_{\text{FLCI}}^*(X, C)$. We first note that any estimator $a'Y$ that does not have infinite worst-case bias must satisfy $a'w = 1$, which implies $1 \leq \|a\|_2 \cdot \|w\|_2$ by the Cauchy-Schwarz inequality, so that the variance $\sigma^2 a'a$ is bounded by

$\sigma^2/\|w\|_2^2 \leq \sigma^2\eta^{-1}/n$. It therefore suffices to show that the worst-case bias is bounded by a constant times $C(k/n)^{1/2}$ (for (i)), or a constant times C (for (ii)).

For part (i), let $\tilde{\gamma} = -C\eta\sqrt{k/n}Z'(ZZ)^{-1}w$. Observe $\text{Pen}(\gamma) = C(k/n)\eta\sqrt{w'(ZZ)^{-1}w} \leq C\eta \cdot (\max \text{eig}(Z'Z/k)^{-1})^{1/2}\sqrt{w'w/n} \leq C$. Let $\tilde{\beta} = C\eta\sqrt{k/n}$. Then $w\tilde{\beta} + Z\tilde{\gamma} = 0$. Thus, $\tilde{\beta}, \tilde{\gamma}$ is observationally equivalent to the parameter vector $\beta = 0, \gamma = 0$, which implies that the length of any CI must be at least $C\eta\sqrt{k/n}$.

Part (ii), follows by an analogous argument, with $\tilde{\gamma} = -Z'(ZZ')^{-1}w \cdot C\eta^{1/2}$ and $\tilde{\beta} = C\eta^{1/2}$.

A.6 Proof of Theorem 4.3

Since the lower bound $c \cdot n^{-1/2}$ follows from standard efficiency bounds with finite dimensional parameters (e.g. taking the submodel where $\delta = \gamma = 0$), we show the lower bound $E_{\vartheta^*} \hat{\chi} \geq C_n \cdot c \cdot \sqrt{\log k}/\sqrt{n}$. To show this, we follow essentially the same arguments as [Cai and Guo \(2017, Theorem 3\)](#) and [Javanmard and Montanari \(2018, Proposition 4.2\)](#), noting that the required bounds on $\|\delta\|$ and $\|\gamma\|$ hold for the distributions used in the lower bound. Under a given parameter vector $\vartheta = (\beta, \gamma', \delta', \sigma^2, \sigma_v^2)$, the data $(Y_i, w_i, z_i)'$ are i.i.d. normal with mean zero and variance matrix

$$\Sigma_{\vartheta} = \begin{pmatrix} \sigma^2 + \beta^2(\sigma_v^2 + \|\delta\|_2^2) + 2\beta\delta'\gamma + \|\gamma\|_2^2 & \beta(\sigma_v^2 + \|\delta\|_2^2) + \gamma'\delta & \beta\delta' + \gamma' \\ \beta(\sigma_v^2 + \|\delta\|_2^2) + \gamma'\delta & \sigma_v^2 + \|\delta\|_2^2 & \delta' \\ \beta\delta + \gamma & \delta & I_k \end{pmatrix}.$$

Let f_{π} denote the distribution of the data $\{Y_i, w_i, z_i\}_{i=1}^n$ when the parameters follow a prior distribution π , and let $\chi^2(f_{\pi_0}, f_{\pi_1})$ denote the chi-square distance between these distributions for prior distributions π_0 and π_1 . By Lemma 1 in [Cai and Guo \(2017\)](#), it suffices to find a prior distribution π_1 over the parameter space $\Theta(C_n, C_n \cdot K \sqrt{n/\log k}, \eta_n)$ such that π_1 places probability one on $\beta = \beta_{1,n}$ for some sequence with $|\beta_{1,n}|$ bounded from below by a constant times $C_n\sqrt{\log k}/\sqrt{n}$ and such that $\chi^2(f_{\pi_0}, f_{\pi_1}) \rightarrow 0$, where π_0 is the distribution that places probability one on ϑ^* given in the statement of the theorem.

To this end, we first note that we can assume $\sigma_0^2 = \sigma_{v,0}^2 = 1$ without loss of generality, since dividing Y_i and w_i by σ_0 and $\sigma_{v,0}$ leads to the same model with parameters multiplied by constants that depend only on σ_0 and $\sigma_{v,0}$.

Let π_1 be defined by a uniform prior for δ over the set with $\|\delta\|_0 = s$ and each element $\delta_j \in \{0, \nu\}$, where s and ν will be determined below. We then set the remaining parameters as deterministic functions of δ : $\beta = -\|\delta\|_2^2/(1 - \|\delta\|_2^2)$, $\gamma = (1 - \beta)\delta$, $\sigma_v^2 = 1 - \|\delta\|_2^2$ and $\sigma^2 = (1 - 2\|\delta\|_2^2)/(1 - \|\delta\|_2^2)$. We note that $\|\delta\|_2$ is constant under this prior, so that β is a

unit point mass as required. This leads to the variance matrix

$$\Sigma_{\vartheta} = \begin{pmatrix} 1 & 0 & \delta' \\ 0 & 1 & \delta' \\ \delta & \delta & I_k \end{pmatrix}$$

for ϑ in the support of π_1 , and $\Sigma_{\vartheta^*} = I_{k+2}$ under the point mass π_0 . It now follows from eqs. (118) and (119) in [Javanmard and Montanari \(2018\)](#) (which are applications of Lemmas 2 and 3 in [Cai and Guo \(2017\)](#)) that

$$\chi^2(f_{\pi_0}, f_{\pi_1}) \leq e^{\frac{s^2}{k-s}} \left(1 + \frac{s}{k}(e^{4n\nu^2} - 1)\right)^s - 1.$$

We set $\nu = (\sqrt{c_\nu}/2) \cdot \sqrt{\log k}/\sqrt{n}$ for some $c_\nu > 0$ so that $e^{4n\nu^2} = k^{c_\nu}$. We then set s to be the greatest integer less than $C_n/\nu = (2C_n/\sqrt{c_\nu}) \cdot (\sqrt{n}/\sqrt{\log k})$. The condition that $C_n \leq \sqrt{k/n} \cdot k^{-\tilde{\eta}}$ for some $\tilde{\eta} > 0$ then guarantees that $s \leq k^\psi$ for some $\psi < 1/2$, so that the above display is bounded by

$$e^{k^{2\psi-1}(1-k^{\psi-1})^{-1}} \left(1 + \frac{1}{s}k^{2\psi-1}(k^{c_\nu} - 1)\right)^s - 1.$$

This converges to zero as required if c_ν is chosen small enough so that $2\psi + c_\nu < 1$.

Finally, we note that, under π_1 , $\|\delta\|_2^2 = (1 + o(1))s\nu^2 = (1 + o(1))C_n\nu = (1 + o(1)) \cdot C_n(\sqrt{c_\nu}/2) \cdot \sqrt{\log k}/\sqrt{n}$ and $|\beta| = \|\delta\|_2^2(1 + o(1)) = (1 + o(1))C_n(\sqrt{c_\nu}/2) \cdot \sqrt{\log k}/\sqrt{n}$. Thus, we obtain a lower bound of $C_n \cdot c \cdot \sqrt{\log k}/\sqrt{n}$ as required.

Appendix B Additional results

We present some additional results that are useful for practical implementation with unknown error variance, and for assessing the plausibility of the assumption $\text{Pen}(\gamma) \leq C$. [Appendix B.1](#) considers the problem of estimating the regression function globally, and derives properties of a regularized regression estimator in this problem. [Appendix B.2](#) applies this estimator as an initial estimator used to construct residuals for standard errors with unknown error variance. [Appendix B.3](#) presents a lower CI for C that can be used to assess the plausibility of the assumption $\text{Pen}(\gamma) \leq C$.

Throughout most of this section, we focus on the case where $\text{Pen}(\gamma) = \|\gamma_2\|_p$, with $k_2 \rightarrow \infty$ and $k_1/n \rightarrow 0$. We use the following notation. Let $\theta = (\beta, \gamma)'$, and let $X = (X_1, X_2)$, where $X_1 = (w, Z_1)$, and $X_2 = Z_2$. We partition θ accordingly, with $\theta_1 = (\beta, \gamma_1)'$, and $\theta_2 = \gamma_2$.

Let $H_{X_1} = X_1(X_1'X_1)^{-1}X_1'$ and $M_{X_1} = I - H_{X_1}$ denote projections onto the column space of X_1 and its orthogonal complement.

For some results in this section, we allow for the possibility that the distribution of ε_i is unknown and possibly non-Gaussian, which requires some additional notation. We consider coverage over a class \mathcal{Q}_n of distributions for ε , and we use $P_{\theta,Q}$ and $E_{\theta,Q}$ to denote probability and expectation when the data Y are drawn according to $Q \in \mathcal{Q}_n$ and $\theta = (\beta, \gamma)' \in \mathbb{R} \times \Gamma$, and we use the notation P_Q and E_Q for expressions that depend on ε only and not on θ . We assume throughout that ε_i is independent across i under each $Q \in \mathcal{Q}_n$.

B.1 Estimating the regression function globally

Consider the regularized regression estimate of θ , given by

$$\hat{\theta} = \underset{\vartheta}{\operatorname{argmin}} \|Y - X\vartheta\|_2^2/n + \lambda\|\vartheta_2\|_p. \quad (19)$$

We first give an elementary property of $\hat{\theta}$, following standard arguments (see [Bühlmann and van de Geer \(2011, Section 6.2\)](#) and [van de Geer \(2000, Chapter 10.1\)](#)), and we derive rates of convergence for this estimator. In the remainder of the appendix, use the estimator to construct feasible CIs with unknown error distribution, and to construct a lower CI for the regularity parameter C .

Lemma B.1. *If $\|2X_2'M_{X_1}\varepsilon\|_q/n \leq \lambda_0$, then $\|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n + (\lambda - \lambda_0)\|\hat{\theta}_2\|_p \leq (\lambda + \lambda_0)\|\theta_2\|_p$.*

Proof. We can write the objective function as

$$\|H_{X_1}Y - X_1\vartheta_1 - H_{X_1}X_2\vartheta_2\|_2^2/n + \|M_{X_1}Y - M_{X_1}X_2\vartheta_2\|_2^2/n + \lambda\|\vartheta_2\|_p.$$

The first part of the objective can be set to zero for any ϑ_2 by taking $\vartheta_1 = (X_1'X_1)^{-1}X_1'Y - (X_1'X_1)^{-1}X_1'X_2\vartheta_2$. Therefore,

$$\hat{\theta}_2 = \underset{\vartheta}{\operatorname{argmin}} \|M_{X_1}Y - M_{X_1}X_2\vartheta_2\|_2^2/n + \lambda\|\vartheta_2\|_p,$$

with $\hat{\theta}_1 = (X_1'X_1)^{-1}X_1'Y + (X_1'X_1)^{-1}X_1'X_2\hat{\theta}_2$. This implies $H_{X_1}\varepsilon = H_{X_1}Y - H_{X_1}X'\theta = H_{X_1}X'(\hat{\theta} - \theta)$, so that

$$\|X(\hat{\theta} - \theta)\|_2^2/n = \|H_{X_1}\varepsilon\|_2^2/n + \|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n, \quad (20)$$

Using the fact that $\hat{\theta}_2$ attains a lower value of the objective than the true parameter value

θ_2 , we obtain an ℓ_p version of what in the ℓ_1 case [Bühlmann and van de Geer \(2011, Lemma 6.1\)](#) term “the Basic Inequality,”

$$\|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n + \lambda\|\hat{\theta}_2\|_p \leq 2\varepsilon' M_{X_1}X_2(\hat{\theta}_2 - \theta_2)/n + \lambda\|\theta_2\|_p.$$

By Hölder’s inequality, we have $2\varepsilon' M_{X_1}X_2(\hat{\theta}_2 - \theta_2) \leq \|2X_2' M_{X_1}\varepsilon\|_q\|\hat{\theta}_2 - \theta_2\|_p$ so that, on the event $\|2X_2' M_{X_1}\varepsilon\|_q/n \leq \lambda_0$, we have

$$\|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n + \lambda\|\hat{\theta}_2\|_p \leq \lambda_0\|\hat{\theta}_2 - \theta_2\|_p + \lambda\|\theta_2\|_p \leq \lambda_0\|\hat{\theta}_2\|_p + (\lambda + \lambda_0)\|\theta_2\|_p,$$

which implies the result. \square

We now use this result to derive rates of convergence for the regularized regression estimator in eq. (19) for estimating the regression function in ℓ_2 loss. For simplicity, we use a fixed sequence for the penalty parameter satisfying certain sufficient conditions. In practice, data-driven methods such as cross-validation may be appealing. We discuss another possible choice based on moderate deviations bounds in Remark B.1 in Appendix B.3 below. Our aim here is to present simple sufficient conditions to allow this estimator to be used for auxiliary purposes such as standard error construction, and we leave the analysis of such extensions for future research.

Theorem B.1. *Suppose that, for some $\eta > 0$, and for all n and all $Q \in \mathcal{Q}_n$, we have $E_Q[|\varepsilon_i|^{\max\{2+\eta, q\}}] < 1/\eta$ when $p > 1$ and $P_Q(|\varepsilon_i| > t) \leq 2\exp(-\eta t)$ when $p = 1$. Suppose the elements of $M_{X_1}X_2$ are bounded by some constant K_X uniformly over n . Let $\hat{\theta}$ be the penalized regression estimator defined in eq. (19) with $\lambda = K_n r_q(k_2, n)$, where $K_n \rightarrow \infty$ and $r_q(k, n)$ given in eq. (14). Then*

$$\sup_{\theta \in \mathbb{R}^{k+1}} \sup_{Q \in \mathcal{Q}_n} P_{\theta, Q} \left(\|X(\hat{\theta} - \theta)\|_2^2/n > K_n(k_1/n + 2\|\theta_2\|_p r_q(k_2, n)) \right) \rightarrow 0,$$

Proof. Lemma B.1 and Lemma B.2 below, we have $\|M_{X_1}X_2(\hat{\theta} - \theta)\|_2^2/n \leq 2K_n\|\theta_2\|_p r_q(k_2, n)$ with probability approaching one uniformly over $\theta \in \mathbb{R}^{k+1}$ and $Q \in \mathcal{Q}_n$. In addition, since H_{X_1} is idempotent with rank $(k_1 + 1)/n$ and $E_Q\varepsilon\varepsilon'$ is diagonal with elements bounded uniformly over $Q \in \mathcal{Q}_n$, we have $E_Q\|H_{X_1}\varepsilon\|_2^2/n \leq \tilde{K}k_1/n$ for some constant \tilde{K} . The result follows by Markov’s inequality and eq. (20). \square

Lemma B.2. *Under the conditions of Theorem B.1, for any sequence $K_n \rightarrow \infty$, we have $\inf_{Q \in \mathcal{Q}_n} P_Q(\|2X_2' M_{X_1}\varepsilon\|_q/n \leq K_n r_q(k_2, n)) \rightarrow 1$.*

Proof. Let $\tilde{x}_{ij} = (2M_{X_1}X_2)_{ij}$. For $q < \infty$, we have

$$E_Q \|2X_2' M_{X_1} \varepsilon\|_q^q = E_Q \sum_{j=1}^{k_2} \left(\sum_{i=1}^n \tilde{x}_{ij} \varepsilon_i \right)^q \leq k_2 \cdot K \cdot n^{q/2}$$

for some constant K that depends only on η , q and K_X , by [von Bahr \(1965\)](#). The result then follows by Markov's inequality. For $q = \infty$, we have

$$P_Q \left(\|2X_2' M_{X_1} \varepsilon\|_q / n > K_n \sqrt{\log k_2} / \sqrt{n} \right) = P_Q \left(\max_j \left| \sum_{i=1}^n \tilde{x}_{ij} \varepsilon_i \right| / n > K_n \sqrt{\log k_2} / \sqrt{n} \right),$$

which, for some $\tilde{K} > 0$, is bounded by $2k_2 \exp(-\tilde{K} \cdot K_n^2 \log k_2) = 2k_2^{1-\tilde{K} \cdot K_n^2} \rightarrow 0$ by Hoeffding's inequality for sub-Gaussian random variables ([Vershynin, 2018](#), Thm. 2.6.3). \square

B.2 Standard errors

We consider standard errors for linear estimators of the form $\hat{\beta} = a'Y$ considered in the main text. We assume that the weights a are nonrandom: they can depend on X but not on Y . Let $\hat{\theta}$ be an estimate of θ , and let $\hat{\varepsilon} = Y - X\hat{\theta}$. Consider the estimator $\hat{V} = \sum_{i=1}^n a_i^2 \hat{\varepsilon}_i^2$ of $V_Q = \text{var}_Q(a'Y) = \sum_{i=1}^n E_Q \varepsilon_i^2$. The weights a are allowed to depend on n so that a_1, \dots, a_n is a triangular array rather than a sequence, but we leave this implicit in the notation. We consider coverage of the feasible bias-aware CI

$$\hat{\beta} \pm \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}) / \sqrt{\hat{V}}) \cdot \sqrt{\hat{V}},$$

where $\overline{\text{bias}}_\Gamma(\hat{\beta})$ is the worst-case bias, given in (6), for the parameter space $\Theta = \mathbb{R} \times \Gamma$ for the parameter $\theta = (\beta, \gamma)'$. We first present a general result for an arbitrary parameter space Θ . We then specialize to the case where $\Theta = \mathbb{R} \times \mathbb{R}^{k_1} \times \{\gamma_2: \|\gamma_2\| \leq C_n\}$ and the residuals $\hat{\varepsilon}_i$ are formed using the regularized regression from [Appendix B.1](#).

Theorem B.2. *Suppose that, for some $\eta > 0$, $\eta \leq E_Q \varepsilon_i^2$ and $E_Q |\varepsilon_i|^{2+\eta} \leq 1/\eta$ for all i and all $Q \in \mathcal{Q}_n$, and that $\sqrt{n} c_n \max_{1 \leq i \leq n} a_i^2 / \sum_{j=1}^n a_j^2 \rightarrow 0$ and $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_{\theta, Q}(\|X(\hat{\theta} - \theta)\|_2 \leq c_n) \rightarrow 1$ for some sequence c_n such that c_n / \sqrt{n} is bounded from above. Then, for any $\delta > 0$, $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q \left(|(\hat{V} - V_Q) / V_Q| < \delta \right) \rightarrow 1$. Furthermore,*

$$\liminf_n \inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q \left(\beta \in \left\{ \hat{\beta} \pm \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}) / \sqrt{\hat{V}}) \cdot \sqrt{\hat{V}} \right\} \right) \geq 1 - \alpha. \quad (21)$$

Proof. We have

$$\frac{\hat{V} - V_Q}{V_Q} = \frac{\sum_{i=1}^n a_i^2 (\hat{\varepsilon}_i^2 - \varepsilon_i^2)}{V_Q} + \frac{\sum_{i=1}^n a_i^2 (\varepsilon_i^2 - E_Q \varepsilon_i^2)}{V_Q}.$$

Let $\tilde{b}_i = a_i^2 / \sum_{j=1}^n a_j^2$. The second term is bounded by $|\sum_{i=1}^n \tilde{b}_i (\varepsilon_i^2 - E_Q \varepsilon_i^2)| / \eta$. The absolute $1 + \eta$ moment of this quantity is bounded by a constant times $\sum_{i=1}^n \tilde{b}_i^{1+\eta} \cdot 1 / \eta^{1+\eta}$ by [von Bahr and Esseen \(1965\)](#). This is bounded by $\max_{1 \leq i \leq n} \tilde{b}_i^\eta \cdot \sum_{i=1}^n \tilde{b}_i / \eta^{1+\eta} = \max_{1 \leq i \leq n} \tilde{b}_i^\eta / \eta^{1+\eta} \rightarrow 0$. The first term is bounded by $\max_{1 \leq i \leq n} \tilde{b}_i / \eta$ times

$$\sum_{i=1}^n |\hat{\varepsilon}_i^2 - \varepsilon_i^2| = \sum_{i=1}^n |\hat{\varepsilon}_i + \varepsilon_i| \cdot |\hat{\varepsilon}_i - \varepsilon_i| \leq \|\hat{\varepsilon} + \varepsilon\|_2 \|\hat{\varepsilon} - \varepsilon\|_2 \leq (\|\hat{\varepsilon} - \varepsilon\|_2 + 2\|\varepsilon\|_2) \|\hat{\varepsilon} - \varepsilon\|_2.$$

For some constant K that depends only on η , we have $2\|\varepsilon\|_2 \leq K\sqrt{n}$ with probability approaching one uniformly over $Q \in \mathcal{Q}_n$. Since $\|\hat{\varepsilon} - \varepsilon\|_2 = \|X(\hat{\theta} - \theta)\|_2 \leq c_n$ it follows that the above display is bounded by $(K\sqrt{n} + c_n) \cdot c_n$ with probability approaching one uniformly over $\theta \in \Theta$, $Q \in \mathcal{Q}_n$. Plugging in the conditions on c_n , it follows that for any $\delta > 0$, $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q \left(\left| (\hat{V} - V_Q) / V_Q \right| < \delta \right) \rightarrow 1$. Coverage of the CI then follows from Theorem F.1 in [Armstrong and Kolesár \(2018\)](#), with the central limit theorem condition following by using the weights and moment bounds to verify the Lindeberg condition (see Lemma F.1 in [Armstrong and Kolesár \(2018\)](#)). \square

The condition on $\text{Lind}(a) = \max_{1 \leq i \leq n} a_i^2 / \sum_{j=1}^n a_j^2$ can be checked on a case-by-case basis, or it can be imposed directly by incorporating a bound on $\text{Lind}(a)$ in the optimization problem that defines the optimal weights. In the latter case, we note that, by the proof of Theorem 4.1, the optimal rate under ℓ_p constraints can be achieved by a linear FLCI with weights a proportional to $w_i - z_i' \delta$, where $\sum_{i=1}^n (w_i - z_i' \delta)^2$ is bounded from below by a constant times n . If $w_i - z_i' \delta$ is bounded, we can thus obtain the optimal rate of convergence if we impose the upper bound $n \text{Lind}(a) \leq \tilde{K}$ for some large constant \tilde{K} (or, more generally, we can allow the bound \tilde{K} to increase with n under appropriate conditions on C_n and p). The conditions of this theorem hold with high probability when the design matrix is drawn such that $w_i - z_i' \delta$ is the population best linear predictor error for predicting w_i with z_i , so this essentially requires tail conditions on this best linear predictor error.

For the setting in Theorem B.1, we can take $c_n = \sqrt{K_n n (k_1/n + C_n r_q(k_2, n))}$ for a slowly increasing constant K_n , so long as $\sqrt{K_n (k_1/n + C_n r_q(k_2, n))} \cdot n \text{Lind}(a) \rightarrow 0$. This gives the following result.

Corollary B.1. *Suppose that, for some $\eta > 0$, $\eta \leq E_Q \varepsilon_i^2$ and $E_Q \varepsilon_i^{2+\eta} \leq 1/\eta$ for all i and $Q \in \mathcal{Q}_n$, and let $\hat{\varepsilon}$ be the residuals from the regularized regression in (19), with λ*

given in Theorem B.1 for some $K_n \rightarrow \infty$, and suppose the conditions from Theorem B.1 hold. Then, if $\sqrt{K_n(k_1/n + C_n r_q(k_2, n))} \cdot n \text{Lind}(a) \rightarrow 0$, the coverage result (21) holds with $\Theta = \mathbb{R} \times \mathbb{R}^{k_1} \times \{\gamma_2: \|\gamma_2\| \leq C_n\}$.

To interpret the conditions on $\text{Lind}(a)$, consider the case where k_1 is fixed, $k/n \rightarrow \infty$ and C_n is bounded away from zero. Then the condition on $\text{Lind}(a)$ reduces to $\sqrt{K_n \cdot C_n r_q(k, n)} \cdot n \text{Lind}(a) \rightarrow 0$. Note also that, in this case, $C_n r_q(k, n)$ is the optimal rate of convergence for $\hat{\beta}$ from Theorem 4.1. Thus, in this case, we require $n \text{Lind}(a)$ to increase more slowly than the inverse of the square root of the optimal rate of convergence for $\hat{\beta}$. In particular, if $n \text{Lind}(a)$ is bounded as $n \rightarrow \infty$, then we can always construct a feasible bias-aware CI that is asymptotically valid. As described above, this bound can be imposed on the weights without affecting the rate of convergence of the width of the CI.

B.3 Lower CIs for C

We present a lower CI for the regularity parameter C , which can be used to assess the plausibility of the assumption $\text{Pen}(\gamma_2) \leq C$. Let $\hat{\theta}_2(\lambda)$ denote the regularized regression estimator of γ_2 , given in (19), with penalty λ . Let λ_α^* denote an upper bound for the $1 - \alpha$ quantile of $\|2X_2' M_{X_1} \varepsilon\|_q/n$. Let

$$\underline{\hat{C}} = \sup_{\lambda > \lambda_\alpha^*} \frac{\lambda - \lambda_\alpha^*}{\lambda + \lambda_\alpha^*} \|\hat{\theta}_2(\lambda)\|_p. \quad (22)$$

In the idealized finite sample setting with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ with σ^2 known, λ_α^* can be computed exactly, so that $\underline{\hat{C}}$ is feasible.

Theorem B.3. *Let $\underline{\hat{C}}$ be given in (22) with λ_α^* given by the $1 - \alpha$ quantile of $\|2X_2' M_{X_1} \varepsilon\|_q/n$. Then, for any $\beta, \gamma_1, \gamma_2$ with $\|\gamma_2\|_p \leq C$, we have $P_{\beta, \gamma_1, \gamma_2}(C \in [\underline{\hat{C}}, \infty)) \geq 1 - \alpha$.*

Proof. It follows from Lemma B.1 that, on the event $\|2X_2' M_{X_1} \varepsilon\|_q/n \leq \lambda_\alpha^*$ (which holds with probability at least $1 - \alpha$ by assumption), we have $\frac{\lambda - \lambda_\alpha^*}{\lambda + \lambda_\alpha^*} \|\hat{\theta}_2(\lambda)\|_p \leq \|\gamma_2\|_p \leq C$ for all $\lambda > \lambda_\alpha^*$. Thus, the supremum of this quantity over λ in this set is also no greater than C on this event. \square

We now present a feasible version of this CI when the error distribution is unknown and possibly heteroskedastic in the case where $p = 1$. Let $\tilde{x}_{ij} = (M_{X_1}' X_2)_{ij}$. Since $q = \infty$ in this case, we need to choose $\hat{\lambda}_\alpha^*$ such that

$$2\|X_2 M_{X_1}' \varepsilon\|_\infty/n = \max_{1 \leq j \leq k_2} \left| \sum_{i=1}^n 2\tilde{x}_{ij} \varepsilon_i/n \right| \leq \hat{\lambda}_\alpha^*$$

with probability at least $1 - \alpha$ asymptotically. Let $\hat{V}_j = \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 \hat{\varepsilon}_i^2$, where $\hat{\varepsilon}_i$ is the residual from an initial regularized regression with λ chosen as in Theorem B.1 for some slowly increasing K_n . This leads to the moderate deviations critical value $\hat{\lambda}_\alpha^*$, which sets

$$\alpha = \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}_\alpha^*/\hat{V}_j^{1/2}). \quad (23)$$

Remark B.1. The analysis in Theorem B.1 of the regularized regression estimator (19) relies on choosing a penalty parameter that is greater than $2\|X_2 M'_{X_1} \varepsilon\|_\infty/n$ with high probability, which is precisely the goal of the critical value $\hat{\lambda}_\alpha^*$ given in (23). This suggests an iterative procedure in which one uses $\hat{\lambda}_\alpha^*$ (perhaps with some sequence α_n converging slowly to zero) as a data-driven penalty parameter in the regression (19) after using some initial penalty choice satisfying the conditions of Theorem B.1 to form the residuals used to compute $\hat{\lambda}_\alpha^*$.

The penalty choice $\hat{\lambda}_\alpha^*$ is related to data-driven choices of the lasso penalty in the case with unknown error distribution. Belloni et al. (2012) use similar ideas to choose the penalty parameter in this setting under ℓ_0 constraints, although our implementation is somewhat different, since our parameter space constrains the penalty loadings we place on each parameter. While $\hat{\lambda}_\alpha^*$ does not take into account correlations between the moments, one could take into account these correlations using a bootstrap implementation, as suggested by Chernozhukov et al. (2013).

Theorem B.4. *Suppose that, for some $\eta > 0$, the conditions of Theorem B.1 hold with $p = 1$, and that $\frac{1}{n} \sum_{i=1}^n \tilde{x}_{ij}^2 \geq \eta$ for $j = 1, \dots, k$ for all n , where $\tilde{x}_{ij} = (M_{X_1} X_2)_{ij}$. Let $\hat{\lambda}_\alpha^*$ be given in (22) with \hat{V}_j formed using residuals $\hat{\varepsilon}$ from the regularized regression (19) with penalty λ chosen as in Theorem B.1 for some $K_n \rightarrow \infty$ with $K_n(k_1/n + (C_n + 1)\sqrt{\log k_2}/\sqrt{n}) \cdot (\log k_2)^2 \rightarrow 0$. Then, $\limsup_n \sup_{\beta, \gamma: \|\gamma_2\|_1 \leq C_n} \sup_{Q \in \mathcal{Q}_n} P_{\theta, Q} \left(\max_{1 \leq j \leq k_2} |\sum_{i=1}^n 2\tilde{x}_{ij} \varepsilon_i/n| > \hat{\lambda}_\alpha^* \right) \leq \alpha$. In particular, letting \hat{C} be given in (22) with λ_α^* given by $\hat{\lambda}_\alpha^*$, we have*

$$\liminf_n \inf_{\beta, \gamma: \|\gamma_2\|_1 \leq C_n} \inf_{Q \in \mathcal{Q}_n} P_{\theta, Q} \left(C_n \in [\hat{C}, \infty) \right) \geq 1 - \alpha.$$

Proof. Let $\tilde{V}_j = \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 \varepsilon_i^2$ and let $V_{Q,j} = \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 E_Q \varepsilon_i^2$. Note that

$$\begin{aligned} |\hat{V}_j - \tilde{V}_j| &= \left| \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 (\hat{\varepsilon}_i^2 - \varepsilon_i^2) \right| = \left| \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 (\hat{\varepsilon}_i + \varepsilon_i)(\hat{\varepsilon}_i - \varepsilon_i) \right| \\ &\leq (2K_X/n)^2 \|\hat{\varepsilon} + \varepsilon\|_2 \|\hat{\varepsilon} - \varepsilon\|_2 \leq (2K_X/n)^2 (2\|\varepsilon\|_2 + \|\hat{\varepsilon} - \varepsilon\|_2) \|\hat{\varepsilon} - \varepsilon\|_2. \end{aligned}$$

On the event that $2\|\varepsilon\|_2 \leq \sqrt{n}\tilde{K}$ and

$$\|\hat{\varepsilon} - \varepsilon\|_2 = \|X(\hat{\theta} - \theta)\|_2 \leq \sqrt{nK_n} \cdot (k_1/n + 2C_n\sqrt{\log n}/\sqrt{n})^{1/2},$$

which holds with probability approaching one uniformly over $Q \in \mathcal{Q}_n$ when \tilde{K} is large enough, this is bounded by $(2K_X/n)^2(\tilde{K}\sqrt{n} + \sqrt{nK_n}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2}) \cdot \sqrt{nK_n}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2}$. Since $V_{Q,j} \geq \tilde{\eta}/n$ uniformly over j and over n for some $\tilde{\eta} > 0$, this implies that, on this event, $\max_{1 \leq j \leq k_2} |\hat{V}_j - \tilde{V}_j|/V_{Q,j}$ is bounded by

$$4\tilde{\eta}^{-1}(K_X^2/n)(\tilde{K}\sqrt{n} + \sqrt{nK_n}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2}) \cdot \sqrt{nK_n}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2},$$

which in turn is bounded by a constant times $K_n^{1/2}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2}$ so long as this quantity converges to zero.

In addition, note that $(\tilde{V}_j - V_{Q,j})/V_{Q,j} = \sum_{i=1}^n \tilde{a}_{ij}(\varepsilon_i - E_Q \varepsilon_i)/n$, where $\tilde{a}_{ij} = \tilde{x}_{ij}^2/(nV_{j,Q}) \leq K_X^2\tilde{\eta}^{-1}$ and $\tilde{\eta}$ is a lower bound for $nV_{Q,j}$. Using this bound on \tilde{a}_{ij} and the tail bound on ε_i , it follows from Bernstein's inequality for sub-exponential random variables that, for $\delta < 1$, $P_Q(|\tilde{V}_j - V_{Q,j}|/V_{Q,j} \geq \delta)$ is bounded from above by $2\exp(-c_n\delta^2)$ for some constant c that depends only on K_X , $\tilde{\eta}$ and η . Thus, for any sequence δ_n , we have $P_Q(\max_{1 \leq j \leq k_2} |\tilde{V}_j - V_{Q,j}|/V_{Q,j} \geq \delta) \leq 2k_2\exp(-c_n\delta_n^2)$, which converges to zero so long as δ_n is bounded from below by a large enough constant times $\sqrt{\log k_2}/\sqrt{n}$.

This gives the rate of convergence for $\hat{V}_j/V_{Q,j}$ to one which, by continuous differentiability of $t \mapsto \sqrt{t}$ at $t = 1$, gives the same rates for $\sqrt{\hat{V}_j}/\sqrt{V_{Q,j}}$. In particular, letting c_n be given by a large enough constant times $K_n^{1/2}(k_1/n + (C_n + 1)\sqrt{\log k_2}/\sqrt{n})^{1/2}$, the event $\max_{1 \leq j \leq k_2} \left| \sqrt{\hat{V}_j}/\sqrt{V_{Q,j}} - 1 \right| \leq c_n$ holds with probability approaching one uniformly over $Q \in \mathcal{Q}_n$ and β, γ with $\|\gamma_2\| \leq C_n$. On this event, we have

$$\alpha = \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}_\alpha^*/\sqrt{\hat{V}_j}) \geq \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}_\alpha^*/(\sqrt{V_{Q_n,j}}(1 - c_n))).$$

Thus, letting $\lambda_{\alpha,n}$ solve $\alpha = \sum_{j=1}^{k_2} 2\Phi(-\lambda_{\alpha,n}/(\sqrt{V_{Q_n,j}}))$, we have $\hat{\lambda}_\alpha^*/(1 - c_n) = \lambda_{\tilde{\alpha},n}$ for some $\tilde{\alpha} \leq \alpha$, so that $\hat{\lambda}_\alpha^*/(1 - c_n) \geq \lambda_{\alpha,n}$. It follows that the non-coverage probability under any sequence of parameters with $\|\gamma_2\|_p \leq C_n$ and any sequence $Q_n \in \mathcal{Q}_n$ is bounded by a term that converges to zero plus

$$P_{Q_n} \left(\max_{1 \leq j \leq k_2} \left| \sum_{i=1}^n 2\tilde{x}_{ij}\varepsilon_i \right| > (1 - c_n)\lambda_{\alpha,n} \right) \leq \sum_{j=1}^{k_2} F_{n,j}(- (1 - c_n)\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}})$$

$$= \sum_{j=1}^{k_2} 2\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}}) \cdot A_{n,j} \cdot B_{n,j},$$

where $F_{n,j}(t) = P_{Q_n}(|\sum_{i=1}^n 2\tilde{x}_{ij}\varepsilon_i/\sqrt{V_{Q_{n,j}}}| > t)$, $A_{n,j} = \frac{\Phi(-(1-c_n)\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}})}{\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}})}$ and $B_{n,j} = \frac{F_{n,j}(-(1-c_n)\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}})}{2\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}})}$. Since $\sum_{j=1}^{k_2} 2\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}}) = \alpha$ by definition, it suffices to show that $\limsup_{n \rightarrow \infty} \max_{1 \leq j \leq k_2} \max\{A_{n,j}, B_{n,j}\} \leq 1$.

For $A_{n,j}$, we use the bound $\Phi(-s)/\Phi(-t) \leq [s^{-1}/(t^{-1} - t^{-3})] \exp((t^2 - s^2)/2)$ (this follows from the bound $(t^{-1} - t^{-3}) \exp(-t^2/2)/\sqrt{2\pi} \leq \Phi(-t) \leq t^{-1} \exp(-t^2/2)/\sqrt{2\pi}$ given in Lemma 2, Section 7.1 in [Feller \(1968\)](#)), which gives

$$A_{n,j} \leq \frac{(1 - c_n)^{-1}}{1 - (\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}})^{-2}} \exp([1 - (1 - c_n)^2] \lambda_{\alpha,n}^2 / (2V_{Q_{n,j}})).$$

Using standard calculations and the fact that $nV_{Q_{n,j}}$ is uniformly bounded from above and below, we have $(\log k_2)/K \leq \lambda_{\alpha,n}^2/V_{Q_{n,j}} \leq K \log k_2$ for some constant K . Thus, the right-hand side of the above display converges to 1 uniformly over n and $1 \leq j \leq k$ so long as $c_n \log k_2 \rightarrow 0$, which is guaranteed by the assumptions of the theorem.

For $B_{n,j}$, we use a moderate deviations bound as in [Feller \(1971, Chapter 16.7\)](#). In particular, the bound $|F_{n,j}(t)/(2\Phi(t)) - 1| \leq \tilde{K}t^3/\sqrt{n}$ holds for all $1 \leq t < \bar{t}_n$, where \bar{t}_n is any sequence with $\bar{t}_n/n^{1/6} \rightarrow 0$, and \tilde{K} depends only on \bar{t}_n and the moment conditions and tail bounds on ε_i ([Armstrong and Chan, 2016, Lemma B.5](#)). Using the fact that $\lambda_{\alpha,n}/\sqrt{V_{Q_{n,j}}}$ is bounded by a constant times $\sqrt{\log k_2}$, it follows that $\limsup_{n \rightarrow \infty} \max_{1 \leq j \leq k_2} B_{n,j} \leq 1$ so long as $(\log k_2)^{3/2}/\sqrt{n} \rightarrow 0$, which is guaranteed by the conditions of the theorem. \square

References

- Armstrong, T. B. and Chan, H. P. (2016). Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics*, 194(1):24–43.
- Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.
- Armstrong, T. B. and Kolesár, M. (2016). Optimal inference in a class of regression models. arXiv:1511.06028v2.
- Armstrong, T. B. and Kolesár, M. (2020a). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. arXiv: 1712.04594.

- Armstrong, T. B. and Kolesár, M. (2020b). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, forthcoming.
- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.
- Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 1(4):791–896.
- Bühlmann, P. and van de Geer, S. A. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin, Heidelberg.
- Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646.
- Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.
- Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.
- Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829.
- Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. J. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, New York, NY, third edition.
- Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, NY.

- Goldberg, M. (2017). Continuity of seminorms on finite-dimensional vector spaces. *Linear Algebra and its Applications*, 515:175–179.
- Heckman, N. E. (1988). Minimax estimates in a semiparametric model. *Journal of the American Statistical Association*, 83(404):1090–1096.
- Ibragimov, I. A. and Khas'minskii, R. Z. (1985). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32.
- Imbens, G. and Wager, S. (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278.
- Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909.
- Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *Annals of Statistics*, 46(6A):2593–2622.
- Kolesár, M. and Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304.
- Kwon, K. and Kwon, S. (2020). Inference in regression discontinuity designs under monotonicity. arXiv: 2011.14216.
- Li, C. M. and Müller, U. K. (2020). Linear regression with many controls of limited explanatory power. *Quantitative Economics*, forthcoming.
- Li, K.-C. (1982). Minimality of the method of regularization of stochastic processes. *The Annals of Statistics*, 10(3):937–942.
- Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23(3):824–835.
- Muralidharan, K., Romero, M., and Wüthrich, K. (2020). Factorial designs, model selection, and (incorrect) inference in randomized experiments. Working Paper 26562, National Bureau of Economic Research.
- Noack, C. and Rothe, C. (2020). Bias-aware inference in fuzzy regression discontinuity designs. arXiv: 1906.04631.
- Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

- Rambachan, A. and Roth, J. (2019). An honest approach to parallel trends. Unpublished manuscript, Harvard University.
- Robinson, P. M. (1988). Root- N -consistent semiparametric regression. *Econometrica*, 56(4):931–954.
- Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030.
- Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40(2):812–831.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- van de Geer, S. A. (2000). *Empirical Processes in M -Estimation*. Cambridge University Press, Cambridge, UK.
- van de Geer, S. A., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.
- Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, UK, first edition.
- von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Annals of Mathematical Statistics*, 36(3):808–818.
- von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.
- Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia, PA.
- Yosida, K. (1995). *Functional Analysis*. Springer-Verlag, Berlin, reprint of 6th edition.
- Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.
- Zhang, L. (2013). Nearly optimal minimax estimator for high-dimensional sparse linear regression. *The Annals of Statistics*, 41(4):2149–2175.