Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework

HHS Public Access

Author manuscript

Ann Stat. Author manuscript; available in PMC 2019 April 04.

Published in final edited form as:

Ann Stat. 2018 December ; 46(6B): 3643–3675. doi:10.1214/17-AOS1671.

UNIFORMLY VALID POST-REGULARIZATION CONFIDENCE REGIONS FOR MANY FUNCTIONAL PARAMETERS IN Z-ESTIMATION FRAMEWORK

Alexandre Belloni[*], **Victor Chernozhukov**[†], **Denis Chetverikov**[‡], and **Ying Wei**[§]

[*]Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, North Carolina 27708, USA, abn5@duke.edu

[†]Department of Economics and Operations Research Center, MIT, 50 Memorial Drive, Cambridge, Massachusetts 02142, USA, vchern@mit.edu

[‡]Department of Economics, UCLA, Bunche Hall, Rm 8283, 315 Portola Plaza, Los Angeles, California 90095, USA, chetverikov@econ.ucla.edu

[§]Department of Biostatistics, Columbia University, 722 West 168th St, Rm 633, New York, New York 10032, USA, yw2148@cumc.columbia.edu

Abstract

In this paper, we develop procedures to construct simultaneous confidence bands for $\tilde{p}$ potentially infinite-dimensional parameters after model selection for general moment condition models where $\tilde{p}$ is potentially much larger than the sample size of available data, n. This allows us to cover settings with functional response data where each of the $\tilde{p}$ parameters is a function. The procedure is based on the construction of score functions that satisfy Neyman orthogonality condition approximately. The proposed simultaneous confidence bands rely on uniform central limit theorems for high-dimensional vectors (and not on Donsker arguments as we allow for $\tilde{p} \gg n$). To construct the bands, we employ a multiplier bootstrap procedure which is computationally efficient as it only involves resampling the estimated score functions (and does not require resolving the high-dimensional optimization problems). We formally apply the general theory to inference on regression coefficient process in the distribution regression model with a logistic link, where two implementations are analyzed in detail. Simulations and an application to real data are provided to help illustrate the applicability of the results.

Keywords

Inference after model selection; moment condition models with a continuum of target parameters; Lasso and Post-Lasso with functional response data

SUPPLEMENTARY MATERIAL

Supplement to "Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework" (DOI: 10.1214/17-AOS1671SUPP; .pdf). The supplemental material contains additional proofs omitted in the main text, a discussion of the double selection method, a set of new results for ℓ_1-penalized M-estimators with functional data, additional simulation results, and an empirical application.

1. Introduction.

High-dimensional models have become increasingly popular in the last two decades. Much research has been conducted on estimation of these models. However, inference about parameters in these models is much less understood, although the literature on inference is growing quickly; see the list of references below. In this paper, we construct simultaneous confidence bands for many functional parameters parameters in a very general framework of moment condition models, where each parameter itself can be an infinite-dimensional object, and the number of such parameters can be much larger than the sample size of available data. Our paper builds upon [11], where simultaneous confidence bands have been constructed for many scalar parameters in a high-dimensional sparse z-estimation framework.

As a substantive application, we apply our general results to provide simultaneous confidence bands for parameters in a logistic regression model with functional response data

$$\mathrm{E}_P\left[Y_u \middle| D, X\right] = \Lambda\left(D'\theta_u + X'\beta_u\right), \quad u \in \mathcal{U}, \quad (1.1)$$

where $D = \left(D_1, \ldots, D_{\tilde{p}}\right)'$ is a $\tilde{p}$-vector of covariates whose effects are of interest, $X = (X_1, \ldots, X_p)'$ is a p-vector of controls, $\Lambda : \mathbb{R} \to \mathbb{R}$ is the logistic link function, $\mathcal{U} = [0, 1]$ is a set of indices and for each $u \in \mathcal{U}$, $Y_u = 1\left\{Y \leq (1-u)\underline{y} + u\bar{y}\right\}$ for some constants $\underline{y} \leq \bar{y}$ and the response variable Y, $\theta_{u1} = \left(\theta_{u1}, \ldots, \theta_{u\tilde{p}}\right)'$ is a vector of target parameters and $\beta_u = (\beta_{u1}, \ldots, \beta_{up})'$ is a vector of nuisance parameters. Here, both $\tilde{p}$ and p are allowed to be potentially much larger than the sample size n, and we have $\tilde{p}$ functional target parameters $\left(\theta_{uj}\right)_{u \in \mathcal{U}}$ and p functional nuisance parameters $\left(\beta_{uj}\right)_{u \in \mathcal{U}}$. This example is important because it demonstrates that our methods can be used for inference about the whole distribution of the response variable Y given D and X in a high-dimensional setting, and not only about some particular features of it such as mean or median. This model is called a distribution regression model in [22] and a conditional transformation model in [26], who argue that the model provides a rich class of models for conditional distributions, and offers a useful generalization of traditional proportional hazard models as well as a useful alternative to quantile regression. We develop inference methods to construct simultaneous confidence bands for many functional parameters of this model in Section 3.

Toward this goal, our contributions include to effectively estimate a continuum of high-dimensional nuisance parameters, allow for approximately sparse models, control sparse eigenvalues of a continuum of random matrices, establish an approximate linearization for a collection of “orthogonalized” (or “de-biased”) estimators and establish the validity of a multiplier bootstrap for the construction of confidence bands for the many functional parameters of interest based on these estimators. In particular, these contributions build upon but go much beyond [11] (Corollary 4), which considers the special case of many scalar parameters in a z-estimation framework, and beyond [19] (Theorem 5.1), where simultaneous confidence bands are constructed via multiplier bootstrap for any large

collection of approximately linear scalar estimator $\sqrt{n}\left(\hat{\theta}_j - \theta_j\right) = n^{-1}\sum_{j=1}^{p}\psi_j + r_{nj}$, $j = 1, \ldots, p$, where the ℓ_∞-norm of r_n is $o_P(1/\sqrt{\log p})$.

Our general results refer to the problem of estimating the set of parameters $\left(\theta_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ in the moment condition model,

$$E_P\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right] = 0, \quad u \in \mathcal{U}, j \in [\tilde{p}], \quad (1.2)$$

where W is a random element that takes values in a measurable space $\left(\mathcal{W}, \mathcal{A}_{\mathcal{W}}\right)$ according to a probability measure P, $\mathcal{U} \subset \mathbb{R}^{d_u}$ and $[\tilde{p}] := \{1, \ldots, \tilde{p}\}$ are sets of indices, and for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, ψ_{uj} is a known score function, θ_{uj} is a scalar parameter of interest and η_{uj} is a potentially high-dimensional (or infinite-dimensional) nuisance parameter. Assuming that a random sample of size n, $\left(W_i\right)_{i=1}^{n}$, from the distribution of W is available together with suitable estimators $\hat{\eta}_{uj}$ of η_{uj}, we aim to construct simultaneous confidence bands for $\left(\theta_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ that are valid uniformly over a large class of probability measures P, say $\mathcal{P}_n$. Specifically, for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, we construct an appropriate estimator $\check{\theta}_{uj}$ of θ_{uj} along with an estimator of the standard deviation of $\sqrt{n}\left(\check{\theta}_{uj} - \theta_{uj}\right)$, $\hat{\sigma}_{uj}$, such that

$$P_P\left(\check{\theta}_{uj} - \frac{c_\alpha \hat{\sigma}_{uj}}{\sqrt{n}} \leq \theta_{uj} \leq \check{\theta}_{uj} + \frac{c_\alpha \hat{\sigma}_{uj}}{\sqrt{n}}, \forall u \in \mathcal{U}, j \in [\tilde{p}]\right) \to 1 - \alpha, \quad (1.3)$$

uniformly over $P \in \mathcal{P}_n$, where $\alpha \in (0, 1)$ and c_α is an appropriate critical value, which we choose to construct using a multiplier bootstrap method. The left- and the right-hand sides of the inequalities inside the probability statement (1.3) then can be used as bounds in simultaneous confidence bands for θ_{uj}'s. In this paper, we are particularly interested in the case when $\tilde{p}$ is potentially much larger than n and $\mathcal{U}$ is an uncountable subset of $\mathbb{R}^{d_u}$, so that for each $j \in [\tilde{p}]$, $\left(\theta_{uj}\right)_{u \in \mathcal{U}}$ is an infinite-dimensional (i.e., functional) parameter.

In the presence of high-dimensional nuisance parameters, construction of valid confidence bands is delicate. Dealing with high-dimensional parameters requires relying upon regularization that leads to lack of asymptotic linearization of the estimators of target parameters since regularized estimators of nuisance parameters suffer from a substantial bias and this bias spreads into the estimators of the target parameters. This lack of asymptotic linearization in turn typically translates into severe distortions in coverage probability of the confidence bands constructed by traditional techniques that are based on perfect model selection; see [30–32, 40]. To deal with this problem, we assume that the score functions ψ_{uj} are constructed to satisfy a near-orthogonality condition that makes them immune to first-order changes in the value of the nuisance parameter, namely

$$\partial_r \left\{ \mathrm{E}_P \left[\psi_{uj} \left(W, \theta_{uj}, \eta_{uj} + r\tilde{\eta} \right) \right] \right\} \Big|_{r=0} \approx 0, \quad u \in \mathcal{U}, j \in [\tilde{p}], \quad (1.4)$$

for all $\tilde{\eta}$ in an appropriate set where ∂_r denotes the derivative with respect to r. We shall often refer to this condition as *Neyman orthogonality*, since in lowdimensional parametric settings the orthogonality property originates in the work of Neyman on the $C(\alpha)$ test in the 50s. In Section 2 below, we describe a few general methods for constructing the score functions ψ_{uj} obeying the Neyman orthogonality condition.

The Neyman orthogonality condition (1.4) is important because it helps to make sure that the bias from the estimators of the high-dimensional nuisance parameters does not spread into the estimators of the target parameters. In particular, under (1.4), it follows that

$$\mathrm{E}_{P,W} \left[\psi_{uj} \left(W, \theta_{uj}, \hat{\eta}_{uj} \right) \right] \approx 0, \quad u \in \mathcal{U}, j \in [\tilde{p}],$$

at least up to the first order, where the index W in $\mathrm{E}_{P,W}[\cdot]$ means that the expectation is taken over W only. This makes the estimators of the target parameters θ_{uj} immune to the bias in the estimators $\hat{\eta}_{uj}$, which in turn improves their statistical properties and opens up the possibilities for valid inference.

As the framework (1.2) covers a broad variety of applications, it is instructive to revisit the logistic regression model with functional response data (1.1). To construct score functions ψ_{uj} that satisfy both the moment conditions (1.2) and the Neyman orthogonality condition (1.4) in this example, for $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, define a $(\tilde{p} + p - 1)$-vector of additional nuisance parameters

$$\gamma_u^j = \arg \min_{\gamma \in \mathbb{R}^{\tilde{p}+p-1}} E_P \left[f_u^2 \left\{ D_j - X^j \gamma \right\}^2 \right], \quad (1.5)$$

where $X^j = \left(D'_{[\tilde{p}]\setminus j}, X' \right), D_{[\tilde{p}]\setminus j} = \left(D_1, \ldots, D_{j-1}, D_{j+1}, \ldots, D_{\tilde{p}} \right)'$, and

$$f_u^2 = f_u^2(D, X) = \mathrm{Var}_P \left(Y_u \middle| D, X \right). \quad (1.6)$$

Then, denoting $W = (Y, D, X)$ and splitting θ_u into θ_{uj} and $\theta_{u[\tilde{p}]\setminus j} = \left(\theta_{u1}, \ldots, \theta_{uj-1}, \theta_{uj+1}, \ldots, \theta'_{u[\tilde{p}]} \right)'$, we set

$$\psi_{uj} \left(W, \theta_{uj}, \eta_{uj} \right) = \left\{ Y_u - \Lambda \left(D_j \theta_{uj} + X^j \beta_u^j \right) \right\} \left(D_j - X^j \gamma_u^j \right),$$

where $\eta_{uj} = \left(\beta_u^j, \gamma_u^j\right)$ and $\beta_u^j = \left(\theta'_{u[\tilde{p}]\setminus j}, \beta'_u\right)'$. It is straightforward to see that these score functions ψ_{uj} satisfy the moment conditions (1.2) and to see that they also satisfy the Neyman orthogonality condition (1.4), observe that

$$\partial_\beta \left\{ \mathrm{E}_P\left[\psi_{uj}\left(W, \theta_{uj}, \beta, \gamma_u^j\right)\right]\right\}\Big|_{\beta = \beta_u^j} = -\mathrm{E}_P\left[f_u^2\left\{D_j - X^j\gamma_u^j\right\}\left(X^j\right)'\right] = 0,$$

$$\partial_\gamma \left\{ \mathrm{E}_P\left[\psi_{uj}\left(W, \theta_{uj}, \beta_u^j, \gamma\right)\right]\right\}\Big|_{\gamma = y_{ii}^j} = -\mathrm{E}_P\left[\left\{Y_u - \Lambda\left(D'\theta_u + X'\beta_u\right)\right\}\left(X^j\right)'\right] = 0,$$

where the first line by definition of f_u^2 and γ_u^j since $\mathrm{Var}_P(Y_u \mid D, X) = \Lambda'(D'\theta_u + X'\beta_u)$, and the second by (1.1). Because of this orthogonality condition, we can exploit the moment conditions (1.2) to construct regular, $\sqrt{n}$-consistent, estimators of θ_{uj} even if nonregular, regularized or post-regularized, estimators of $\eta_{uj} = \left(\beta_u^j, \gamma_u^j\right)$ are used to cope with high-dimensionality. Using these regular estimators of θ_{uj}, we then can construct valid confidence bands (1.3).

Our general approach to construct simultaneous confidence bands, which is developed in Section 2, can be described as follows. First, we construct the moment conditions (1.2) that satisfy the Neyman orthogonality condition (1.4), and use these moment conditions to construct estimators $\check{\theta}_{uj}$ of θ_{uj} for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Second, under appropriate regularity conditions, we establish a Bahadur representation for $\check{\theta}_{uj}$'s. Third, employing the Bahadur representation, we are able to derive a suitable Gaussian approximation for the distribution of $\check{\theta}_{uj}$'s. Importantly, the Gaussian approximation is possible even if both $\tilde{p}$ and the dimension of the index set $\mathcal{U}$, d_u, are allowed to grow with n, and $\tilde{p}$ asymptotically remains much larger than n. Finally, from the Gaussian approximation, we construct simultaneous confidence bands using a multiplier bootstrap method. Here, the Gaussian and bootstrap approximations are constructed by applying the results on highdimensional central limit and bootstrap theorems established in [16–21] by verifying the conditions there.

Although regularity conditions underlying our approach can be verified for many models defined by moment conditions, for illustration purposes, we explicitly verify these conditions for the logistic regression model with functional response data (1.1) in Section 3. We also note that the regularity conditions, in particular those related to the entropy of the nuisance parameter estimators, can be substantially relaxed if we use sample splitting, so that the nuisance parameters and parameters of interest are estimated on separate samples; see [15]. In addition, we examine the performance of the proposed procedures in a Monte Carlo simulation study and provide an example based on real data in Section 5. Moreover, in the Supplementary Material [5], we discuss the construction of simultaneous confidence bands based on a double-selection estimator. This estimator does not require to explicitly construct the score functions satisfying the Neyman orthogonality condition but nonetheless is first-order equivalent to the estimator based on such functions.

We also develop new results for ℓ_1-penalized M-estimators in Section 4 to handle functional data and criterion functions that depend on nuisance functions for which only estimates are available building on ideas in [3,4,12] (for brevity of the paper, generic results are deferred to Supplementary Material, and Section 4 only contains results that are relevant for the logistic regression model studied in Section 3). Specifically, we develop a method to select penalty parameters for these estimators and extend the existing theory to cover functional data to achieve rates of convergence and sparsity guarantees that hold uniformly over $u \in \mathcal{U}$. The ability to allow both for functional data and for nuisance functions is crucial in the implementation and in theoretical analysis of the methods proposed in this paper.

Orthogonality conditions like that in (1.4) have played an important role in statistics and econometrics. In low-dimensional settings, a similar condition was used by Neyman in [37] and [38] while in semiparametric models the orthogonality conditions were used in [1, 35, 36, 41] and [33]. In high-dimensional settings, [7] and [2] were the first to use the orthogonality condition (1.4) in a linear instrumental variables model with many instruments. Related ideas have also been used in the literature to construct confidence bands in high-dimensional linear models, generalized linear models and other nonlinear models; see [6, 8–11, 13, 27,28,43, 46, 47] and [39], where we can interpret each procedure as implicitly or explicitly constructing and solving an approximately Neyman-orthogonal estimating equation. We contribute to this quickly growing literature by providing procedures to construct *simultaneous* confidence bands for *many infinite-dimensional* parameters identified by moment conditions.

Throughout the paper, we use the standard notation from the empirical process theory. In particular, we use $\mathbb{E}_n$ to denote the expectation with respect to the empirical measure associated with the data $(W_i)_{i=1}^n$, and we use $\mathbb{G}_n$ to denote the empirical process $\sqrt{n}(\mathbb{E}_n - \mathrm{E}_P)$. More details about the notation are given in the Supplementary Material.

2. Confidence regions for function-valued parameters based on moment conditions.

2.1. Generic construction of confidence regions.

In this section, we state our results under high-level conditions. In the next section, we will apply these results to construct simultaneous confidence bands for many infinite-dimensional parameters in the logistic regression model with functional response data.

Recall that we are interested in constructing simultaneous confidence bands for a set of target parameters $\left(\theta_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ where for each $u \in \mathcal{U} \subset \mathbb{R}^{d_u}$ and $j \in [\tilde{p}] = \{1, \ldots, \tilde{p}\}$, the parameter θ_{uj} satisfies the moment condition (1.2) with η_{uj} being a potentially high-dimensional (or infinite-dimensional) nuisance parameter. Assume that $\theta_{uj} \in \Theta_{uj}$, a finite or infinite interval in $\mathbb{R}$, and that $\eta_{uj} \in T_{uj}$, a convex set in a normed space equipped with a norm $\|\cdot\|_e$. We allow $\mathcal{U}$ to be a possibly uncountable set of indices, and $\tilde{p}$ to be potentially large.

We assume that a random sample $(W_i)_{i=1}^n$ from the distribution of W is available for constructing the confidence bands. We also assume that for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, the nuisance parameter η_{uj} can be estimated by $\hat{\eta}_{uj}$ using the same data $(W_i)_{i=1}^n$. In the next section, we discuss examples where $\hat{\eta}_{uj}$'s are based on Lasso or Post-Lasso methods (although other modern regularization and postregularization methods can be applied). Our confidence bands will be based on the estimators $\check{\theta}_{uj}$ of θ_{uj} that are for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ defined as approximate ϵ_n-solutions in Θ_{uj} to sample analogs of the moment conditions (1.2), that is,

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left\{ \left| \mathbb{E}_n\left[\psi_{uj}\left(W, \check{\theta}_{uj}, \hat{\eta}_{uj}\right)\right] \right| - \inf_{\theta \in \Theta_{uj}} \left| \mathbb{E}_n\left[\psi_{uj}\left(W, \theta, \hat{\eta}_{uj}\right)\right] \right| \right\} \leq \epsilon_n, \quad (2.1)$$

where $\epsilon_n = o(\delta_n n^{-1/2})$ for all $n \geq 1$ and some sequence $(\delta_n)_{n \geq 1}$ of positive constants converging to zero.

To motivate the construction of the confidence bands based on the estimators $\check{\theta}_{uj}$, we first study distributional properties of these estimators. To do that, we will employ the following regularity conditions. Let C_0 be a strictly positive (and finite) constant, and for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, let $\mathcal{T}_{uj}$ be some subset of T_{uj}, whose properties are specified below in assumptions. In particular, we will choose the sets $\mathcal{T}_{uj}$ so that, on the one hand, their complexity does not grow too fast with n but, on the other hand, for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, the estimator $\hat{\eta}_{uj}$ takes values in $\mathcal{T}_{uj}$ with high probability. As discussed before, we rely on the following nearorthogonality condition.

DEFINITION 2.1 (Near-orthogonality condition). For each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, we say that ψ_{uj} obeys the near-orthogonality condition with respect to $\mathcal{T}_{uj} \subset T_{uj}$ if the following conditions hold: The Gateaux derivative map

$$\mathrm{D}_{u,j,\bar{r}}\left[\eta - \eta_{uj}\right] := \partial_r \left\{ \mathrm{E}_P\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj} + r\left(\eta - \eta_{uj}\right)\right)\right] \right\}\Big|_{r=\bar{r}}$$

exists for all $\bar{r} \in [0, 1)$ and $\eta \in \mathcal{T}_{uj}$ and (nearly) vanishes at $\bar{r} = 0$, namely,

$$\left| \mathrm{D}_{u,j,0}\left[\eta - \eta_{uj}\right] \right| \leq C_0 \delta_n n^{-1/2}, \quad (2.2)$$

for all $\eta \in \mathcal{T}_{uj}$.

At the end of this section, we describe several methods to obtain score functions ψ_{uj} that obey the near-orthogonality condition. Together these methods cover a wide variety of applications.

Let ω and c_0 be some strictly positive (and finite) constants, and let $n_0 \geq 3$ be some positive integer. Also, let $(B_{1n})_{n\geq 1}$ and $(B_{2n})_{n\geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $B_{1n} \geq 1$ for all $n \geq 1$. In addition, denote

$$\begin{aligned} \mathcal{S}_n &:= \mathrm{E}_P\left[\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}\left|\sqrt{n}\mathbb{E}_n\left[\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right]\right|\right], \\ J_{uj} &:= \partial_\theta\left\{\mathrm{E}_P\left[\psi_{uj}\left(W,\theta,\eta_{uj}\right)\right]\right\}\Big|_{\theta=\theta_{uj}}. \end{aligned} \tag{2.3}$$

The quantity $\mathcal{S}_n$ measures how rich the process $\left\{\psi_{uj}\left(\cdot,\theta_{uj},\eta_{uj}\right): u\in\mathcal{U}, j\in[\tilde{p}]\right\}$ is. The quantity J_{uj} measures the degree of identifiability of θ_{uj} by the moment condition (1.2). In many applications, it is bounded in absolute value from above and away from zero. Finally, let $\mathcal{P}_n$ be a set of probability measures P of possible distributions of W on the measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$.

We collect our main conditions on the score functions ψ_{uj} and the true values of the target parameters θ_{uj} in the following assumption.

Assumption 2.1 (Moment condition problem). For all $n \geq n_0$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}$, and $j \in [\tilde{p}]$, the following conditions hold: (i) The true parameter value θ_{uj} obeys (1.2), and Θ_{uj} contains a ball of radius $C_0 n^{-1/2}\mathcal{S}_n \log n$ centered at θ_{uj}. (ii) The map $(\theta,\eta) \mapsto \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]$ is twice continuously Gateaux-differentiable on $\Theta_{uj}\times\mathcal{T}_{uj}$. (iii) The score function ψ_{uj} obeys the nearorthogonality condition given in Definition 2.1 for the set $\mathcal{T}_{uj} \subset T_{uj}$. (iv) For all $\theta \in \Theta_{uj}$, $|\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]| \geq 2^{-1}|J_{uj}(\theta - \theta_{uj})| \wedge c_0$, where J_{uj} satisfies $c_0 \leq |J_{uj}| \leq C_0$. (v) For all $r \in [0, 1)$, $\theta \in \Theta_{uj}$, and $\eta \in T_{uj}$:

a. $$\mathrm{E}_P\left[\left(\psi_{uj}(W,\theta,\eta) - \psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right)^2\right] \leq C_0\left(\left|\theta-\theta_{uj}\right| \vee \left\|\eta-\eta_{uj}\right\|_e\right)^\omega,$$

b. $$\left|\partial_r \mathrm{E}_P\left[\psi_{uj}\left(W,\theta,\eta_{uj}+r\left(\eta-\eta_{uj}\right)\right)\right]\right| \leq B_{1n}\left\|\eta-\eta_{uj}\right\|_e,$$

c. $$\left|\partial_r^2 \mathrm{E}_P\left[\psi_{uj}\left(W,\theta_{uj}+r\left(\theta-\theta_{uj}\right),\eta_{uj}+r\left(\eta-\eta_{uj}\right)\right)\right]\right| \leq B_{2n}\left(\left|\theta-\theta_{uj}\right|^2 \vee \left\|\eta-\eta_{uj}\right\|_e^2\right).$$

Assumption 2.1 is mild and standard in moment condition problems. Assumption 2.1 (i) requires θ_{uj} to be sufficiently separated from the boundary of Θ_{uj}. Assumption 2.1(ii) requires that the functions $(\theta,\eta) \mapsto \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]$ are smooth. It is a mild condition because it does not require smoothness of the functions $(\theta, \eta) \mapsto \psi_{uj}(W,\theta,\eta)$. Assumption 2.1(iii) is our key condition and is discussed above. Assumption 2.1(iv) implies sufficient identifiability of θ_{uj}. In particular, it implies that the equation $\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})] = 0$ has only one solution $\theta = \theta_{uj}$. If this equation has multiple solutions, Assumption 2.1(iv) implies that the set Θ_{uj} is restricted enough so that there is only one solution in Θ_{uj}. Assumption 2.1(v-a) means that the functions $(\theta, \eta) \mapsto \psi_{uj}(W,\theta,\eta)$ mapping $\Theta_{uj} \times T_{uj}$ into $L^2(P)$ are Lipschitz-continuous at $(\theta, \eta) = (\theta_{uj},\eta_{uj})$ with Lipschitz order $\omega/2$. In most applications, we can set ω

= 2. Assumptions 2.1(v-b,v-c) impose smoothness bounds on the functions $(\theta, \eta) \mapsto \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]$.

Next, we state our conditions related to the estimators $\hat{\eta}_{uj}$. Let $(\Delta_n)_{n\geq 1}$ and $(\tau_n)_{n\geq 1}$ be some sequences of positive constants converging to zero. Also, let $(a_n)_{n\geq 1}$, $(v_n)_{n\geq 1}$, and $(K_n)_{n\geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $a_n \geq n \vee K_n$ and $v_n \geq 1$ for all $n \geq 1$. Finally, let $q \geq 2$ be some constant.

Assumption 2.2 (Estimation of nuisance parameters). For all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold: (i) With probability at least $1 - \Delta_n$, we have $\hat{\eta}_{uj} \in \mathcal{T}_{uj}$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. (ii) For all $u \in \mathcal{U}$, $j \in [\tilde{p}]$, and $\eta \in \mathcal{T}_{uj}$, $\|\eta - \eta_{uj}\|_e \leq \tau_n$. (iii) For all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, we have $\eta_{uj} \in \mathcal{T}_{uj}$. (iv) The function class $\mathcal{F}_1 = \left\{\psi_{uj}(\cdot, \theta, \eta) : u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}, \eta \in \mathcal{T}_{uj}\right\}$ is suitably measurable and its uniform entropy numbers obey

$$\sup_Q \log N\left(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}\right) \leq v_n \log(a_n/\epsilon) \quad \text{for all } 0 < \epsilon \leq 1, \qquad (2.4)$$

where F_1 is a measurable envelope for $\mathcal{F}_1$ that satisfies $\|F_1\|_{P,q} \leq K_n$. (v) For all $f \in \mathcal{F}_1$, we have $c_0 \leq \|f\|_{P,2} \leq C_0$. (vi) The complexity characteristics a_n and v_n satisfy:

a. $(v_n \log a_n/n)^{1/2} \leq C_0 \tau_n$,

b. $\left(B_{1n}\tau_n + \mathcal{S}_n \log n/\sqrt{n}\right)^{\omega/2}\left(v_n \log a_n\right)^{1/2} + n^{-1/2+1/q} v_n K_n \log a_n \leq C_0 \delta_n$,

c. $n^{1/2} B_{1n}^2 B_{2n} \tau_n^2 \leq C_0 \delta_n$.

Assumption 2.2 provides sufficient conditions for the estimation of the nuisance parameters $\left(\eta_{uj}\right)_{u\in\mathcal{U}, j\in[\tilde{p}]}$. Assumption 2.2 (i) requires that the set $\mathcal{T}_{uj}$ is large enough so that $\hat{\eta}_{uj} \in \mathcal{T}_{uj}$ with high probability. Assumptions 2.2 (i,ii) together require that the estimator $\hat{\eta}_{uj}$ converges to η_{uj} with the rate τ_n. This rate should be fast enough so that Assumptions 2.2(vi-b,vi-c) are satisfied. Assumption 2.2(iv) gives a bound on the complexity of the set $\mathcal{T}_{uj}$ expressed via uniform entropy numbers, and Assumptions 2.2(vi-a,vi-b) require that the set $\mathcal{T}_{uj}$ is small enough so that its complexity does not grow too fast. Assumption 2.2(v) requires that the functions $(\theta, \eta) \mapsto \psi_{uj}(W,\theta,\eta)$ are scaled properly. Suitable measurability of $\mathcal{F}_1$, required in Assumption 2.2(iv), is a mild condition that is satisfied in most practical cases; see the Supplementary Material and [25] for clarifications. Overall, Assumption 2.2 shows the trade-off in the choice of the sets $\mathcal{T}_{uj}$: setting $\mathcal{T}_{uj}$ large, on the one hand, makes it easy to satisfy Assumption 2.2(i) but, on the other hand, yields large values of a_n and v_n in Assumption 2.2(iv) making it difficult to satisfy Assumption 2.2(vi).

We stress that the class $\mathscr{F}_1$ does not need to be Donsker because its uniform entropy numbers are allowed to increase with n. This is important because allowing for non-Donsker classes is necessary to deal with high-dimensional nuisance parameters. Note also that our conditions are very different from the conditions imposed in various settings with nonparametrically estimated nuisance functions; see, for example, [44, 45] and [29].

In addition, we emphasize that the conditions stated in Assumption 2.2 are sufficient for our results for the general model (1.2) but can often be relaxed if the structure of the functions $\psi_{uj}(W,\theta,\eta)$ is known. For example, it is possible to relax Assumption 2.2(vi) if the functions $\psi_{uj}(W,\theta,\eta)$ are linear in θ, which happens in the linear regression model with θ being the coefficient on the covariate of interest; see [9]. Moreover, it is possible to relax the entropy condition (2.4) of Assumption 2.2 by relying upon sample splitting, where part of the data is used to estimate η_{uj}, and the other part is used to estimate θ_{uj} given an estimate $\hat{\eta}_{uj}$ of η_{uj}; see [2] and [15]. By swapping the role of two parts, and averaging the resulting two estimators, we do not incur any efficiency losses.

The following theorem is our first main result in this paper.

THEOREM 2.1 (Uniform Bahadur representation). *Under Assumptions* 2.1 *and* 2.2, *for an estimator* $\left(\check{\theta}_{uj}\right)_{u\in\mathcal{U},\, j\in[\tilde{p}]}$ *that obeys* (2.1), *we have*

$$\sqrt{n}\sigma_{uj}^{-1}\left(\check{\theta}_{uj}-\theta_{uj}\right)=\mathbb{G}_n\bar{\psi}_{uj}+O_P(\delta_n) \quad (2.5)$$

in $\ell^\infty(\mathcal{U}\times[\tilde{p}])$ *uniformly over* $P\in\mathscr{P}_n$, *where* $\bar{\psi}_{uj}(\cdot):=-\sigma_{uj}^{-1}J_{uj}^{-1}\psi_{uj}\left(\cdot,\theta_{uj},\eta_{uj}\right)$ *and* $\sigma_{uj}^2:=J_{uj}^{-2}\mathrm{E}_P\left[\psi_{uj}^2\left(W,\theta_{uj},\eta_{uj}\right)\right]$.

COMMENT 2.1 (On the proof ofTheorem 2.1). To prove this theorem, we use the following identity:

$$\sqrt{n}\mathrm{E}_{P,W}\left[\psi_{uj}\left(W,\check{\theta}_{uj},\hat{\eta}_{uj}\right)-\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right]=-\sqrt{n}\mathbb{E}_n\left[\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right] \quad (2.6)$$

$$+\sqrt{n}\mathbb{E}_n\left[\psi_{uj}\left(W,\check{\theta}_{uj},\hat{\eta}_{uj}\right)\right]+\mathbb{G}_n\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)-\mathbb{G}_n\psi_{uj}\left(W,\check{\theta}_{uj},\hat{\eta}_{uj}\right). \quad (2.7)$$

Here, the term on the right-hand side of (2.6) is the main term on the right-hand side of (2.5), up to a normalization $(\sigma_{uj}J_{uj})^{-1}$. Also, we show that the first term in (2.7) is $O_P(\delta_n)$ since $\check{\theta}_{uj}$ satisfies (2.1). Moreover, using arather standard theory of Z-estimators, we show that $\check{\theta}_{uj}-\theta_{uj}=O_P\left(B_{1n}\tau_n\right)$. This in turn allows us to show with the help of empirical process arguments that the difference of the last two terms in (2.7) is $O_P(\delta_n)$ as well. (In [15], we also point out that this difference is $O_P(\delta_n)$ under much weaker entropy conditions than

those in Assumption 2.2 if $\hat{\eta}_{uj}$ and $\check{\theta}_{uj}$ are obtained using separate samples.) Thus, it remains to show that the left-hand side of (2.6) is equal to the left-hand side of (2.5) up to an approximation error $O_P(\delta_n)$ and up to a normalization $(\sigma_{uj}J_{uj})^{-1}$. To do so, we use second-order Taylor's expansion of the function

$$f(r) = \sqrt{n}\mathrm{E}_{P,W}\left[\psi_{uj}\left(W, \theta_{uj} + r\left(\check{\theta}_{uj} - \theta_{uj}\right), \eta_{uj} + r\left(\hat{\eta}_{uj} - \eta_{uj}\right)\right)\right]$$

at $r = 1$ around $r = 0$. This gives

$$\begin{aligned}&\sqrt{n}\mathrm{E}_{P,W}\left[\psi_{uj}\left(W, \check{\theta}_{uj}, \hat{\eta}_{uj}\right) - \psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right]\\ &= f(1) - f(0)\\ &= \sqrt{n}J_{uj}\left(\check{\theta}_{uj} - \theta_{uj}\right) + \sqrt{n}D_{u,j,0}\left[\hat{\eta}_{uj} - \eta_{uj}\right] + \sqrt{n}f''(\bar{r})/2\end{aligned}$$

for some $\bar{r} \in (0, 1)$. Here, $\sqrt{n}f''(\bar{r}) = O_P(\delta_n)$ follows from Assumptions 2.1 and 2.2 and the key near-orthogonality condition also allows us to show that $\sqrt{n}D_{u,j,0}\left[\hat{\eta}_{uj} - \eta_{uj}\right] = O_P(\delta_n)$. Without this condition, the term $\sqrt{n}D_{u,j,0}\left[\hat{\eta}_{uj} - \eta_{uj}\right]$ would give first-order bias and lead to slower-than-$\sqrt{n}$ rate of convergence of the estimator $\check{\theta}_{uj}$. Finally, again using the empirical process arguments, we can show that all the bounds including the term $O_P(\delta_n)$ hold uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$.

Comment 2.2 (On uniformity in u in Theorem 2.1). When the functions $u \mapsto \sqrt{n}\sigma_{uj}^{-1}\left(\check{\theta}_{uj} - \theta_{uj}\right) - \mathbb{G}_n\overline{\psi}_{uj}$ are Lipschitz-continuous, one can use a simple discretization argument to conclude that the approximation in (2.5) holds uniformly over $(u, j) \in \mathcal{U} \times [\tilde{p}]$ as long as we can show that it holds for each $(u, j) \in \mathcal{U} \times [\tilde{p}]$. However, in many applications, including the distribution regression model discussed in Section 3, this function is actually not continuous, and the location of jumps depends on the data. Therefore, we have to rely on a more complicated argument to establish uniformity in u in the bound (2.5).

The uniform Bahadur representation derived in Theorem 2.1 is useful for the construction of simultaneous confidence bands for $\left(\theta_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ as in (1.3). For this purpose, we apply new high-dimensional central limit and bootstrap theorems that have been recently developed in a sequence of papers [16, 18–20] and [21]. To apply these theorems, we make use of the following regularity condition.

Let $\left(\bar{\delta}_n\right)_{n \geq 1}$ be a sequence of positive constants converging to zero. Also, let $(\varrho_n)_{n \geq 1}$, $\left(\bar{\varrho}_n\right)_{n \geq 1}$, $(A_n)_{n \geq 1}$, $\left(\bar{A}_n\right)_{n \geq 1}$, and $(L_n)_{n \geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $\varrho_n \geq 1$, $A_n \geq n$, and $\bar{A}_n \geq n$ for all $n \geq 1$. In addition, from now on, we assume that $q > 4$. Denote by $\widehat{\psi}_{uj}(\cdot) := -\hat{\sigma}_{uj}^{-1}\hat{J}_{uj}^{-1}\psi_{uj}\left(\cdot, \check{\theta}_{uj}, \hat{\eta}_{uj}\right)$ an estimator of $\widehat{\psi}_{uj}(\cdot)$, with $\hat{J}_{uj}$ and $\hat{\sigma}_{uj}$ being suitable estimators of J_{uj} and σ_{uj}.

ASSUMPTION 2.3 (Additional score regularity). For all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold: (i) The function class $\mathcal{F}_0 = \left\{\overline{\psi}_{uj}(\cdot) : u \in \mathcal{U}, j \in [\tilde{p}]\right\}$ is suitably measurable and its uniform entropy numbers obey

$$\sup_Q \log N\left(\epsilon \|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}\right) \leq \varrho_n \log\left(A_n/\epsilon\right) \quad \text{for all } 0 < \epsilon \leq 1,$$

where F_0 is a measurable envelope for $\mathcal{F}_0$ that satisfies $\|F_0\|_{P,q} \leq L_n$. (ii) For all $f \in \mathcal{F}_0$ and $k = 3,4$, we have $\mathrm{E}_P\left[|f(W)|^k\right] \leq C_0 L_n^{k-2}$. (iii) The function class $\widehat{\mathcal{F}}_0 = \left\{\overline{\psi}_{uj}(\cdot) - \widehat{\psi}_{uj}(\cdot) : u \in \mathcal{U}, j \in [\tilde{p}]\right\}$ satisfies with probability $1 - \Delta_n$: $\log N\left(\epsilon, \widehat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}\right) \leq \bar{\varrho}_n \log\left(\bar{A}_n/\epsilon\right)$ for all $0 < \epsilon \leq 1$ and $\|f\|_{\mathbb{P}_n,2} \leq \bar{\delta}_n$ for all $f \in \widehat{\mathcal{F}}_0$.

This assumption is technical, and its verification in applications is rather standard. For the Gaussian approximation result below, we actually only need the first and the second part of this assumption. The third part will be needed for establishing validity of the simultaneous confidence bands based on the multiplier bootstrap procedure. As a side note, observe that Assumption 2.3 allows to bound $\mathcal{S}_n$ defined in (2.3) and used in Assumptions 2.1 and 2.2; see Appendix G of the Supplementary Material.

Next, let $\left(\mathcal{N}_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ denote a tight zero-mean Gaussian process indexed by $\mathcal{U} \times [\tilde{p}]$ with covariance operator given by $\mathrm{E}_P\left[\overline{\psi}_{uj}(W)\overline{\psi}_{u'j'}(W)\right]$ for $u, u' \in \mathcal{U}$ and $j, j' \in [\tilde{p}]$. We have the following corollary of Theorem 2.1, which is our second main result in this paper.

COROLLARY 2.1 (Gaussian approximation). *Suppose that Assumptions* 2.1, 2.2 *and* 2.3(i,ii) *hold. In addition, suppose that thefollowing growth conditions are satisfied*: $\delta_n^2 \varrho_n \log A_n = o(1)$, $L_n^{2/7} \varrho_n \log A_n = o\left(n^{1/7}\right)$ *and* $L_n^{2/3} \varrho_n \log A_n = o\left(n^{1/3 - 2/(3q)}\right)$. *Then*

$$\sup_{t \in \mathbb{R}} \left| \mathrm{P}_P\left(\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left|\sqrt{n}\sigma_{uj}^{-1}\left(\check{\theta}_{uj} - \theta_{uj}\right)\right| \leq t\right) - \mathrm{P}_P\left(\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left|\mathcal{N}_{uj}\right| \leq t\right)\right| = o(1)$$

uniformly over $P \in \mathcal{P}_n$.

Based on Corollary 2.1, we are now able to construct simultaneous confidence bands for θ_{uj}'s as in (1.3). In particular, we will use the Gaussian multiplier bootstrap method employing the estimates $\hat{\psi}_{uj}$ of $\overline{\psi}_{uj}$. To describe the method, define the process

$$\widehat{\mathcal{G}} = \left(\widehat{\mathcal{G}}_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \widehat{\psi}_{uj}(W_i)\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}, \quad (2.8)$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables which are independent from the data $(W_i)_{i=1}^n$. Then the multiplier bootstrap critical value c_α is defined as the $(1-\alpha)$ quantile of the conditional distribution of $\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}\left|\widehat{\mathbb{G}}_{uj}\right|$ given the data $(W_i)_{i=1}^n$. To prove validity of this critical value for the construction of simultaneous confidence bands of the form (1.3), we will impose the following additional assumption. Let $(\varepsilon_n)_{n\geq 1}$ be a sequence of positive constants converging to zero.

Assumption 2.4 (Variation estimation). For all $n \geq n_0$ and $P \in \mathcal{P}_n$,

$$\mathrm{P}_P\left(\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}\left|\frac{\hat{\sigma}_{uj}}{\sigma_{uj}}-1\right|>\varepsilon_n\right)\leq\Delta_n.$$

The following corollary establishing validity of the multiplier bootstrap critical value c_α for the simultaneous confidence bands construction is our third main result in this paper.

Corollary 2.2 (Simultaneous confidence bands). Suppose that Assumptions 2.1–2.4 hold. In addition, suppose that the growth conditions of Corollary 2.1 hold. Finally, suppose that $\varepsilon_n \varrho_n \log A_n = o(1)$, and $\bar{\delta}_n^2 \bar{\varrho}_n \varrho_n (\log \bar{A}_n) \cdot (\log A_n) = o(1)$. Then (1.3) holds uniformly over $P \in \mathcal{P}_n$.

Comment 2.3 (Confidence bands based on other bootstrap schemes). Results in [24] suggest that the conditions of Corollary 2.2 can be somewhat relaxed if, instead of using the Gaussian weights in the multiplier bootstrap method, we use Mammen's weights as in [34] or if we use the empirical bootstrap instead of the multiplier bootstrap. Since the results in [24] apply only to high-dimensional random vectors and do not apply to infinite-dimensional random processes, we leave a formal discussion of the results under these alternative bootstrap schemes to future work.

2.2. Construction of score functions satisfying the orthogonality condition.

Here, we discuss several methods for generating orthogonal scores in a wide variety of settings, including the classical Neyman's construction. In what follows, since the argument applies to each u and j, it is convenient to omit the indices u and j and also to use the subscript 0 to indicate the true values of the parameters. For simplicity, we also focus the discussion on the exactly orthogonal case. With these simplifications, we can restate the orthogonality condition as follows: we say that the score ψ obeys the Neyman orthogonality condition with respect to $\eta_0 \in \mathcal{T}$ if the following conditions hold: The Gateaux derivative map

$$\mathrm{D}_{\bar{r}}[\eta-\eta_0] := \partial_r\left\{\mathrm{E}_P[\psi(W,\theta_0,\eta_0+r(\eta-\eta_0))]\right\}\Big|_{r=\bar{r}}$$

exists for all $\bar{r} \in [0,1)$ and $\eta \in \mathcal{T}$ and vanishes at $\bar{r} = 0$, namely,

$$\mathrm{D}_{u,j,0}[\eta - \eta_0] = 0, \quad (2.9)$$

for all $\eta \in \mathcal{T}$.

(1) *Orthogonal scores for likelihood problems with finite-dimensional nuisance parameters*. In likelihood settings with finite-dimensional parameters, the construction of orthogonal equations was proposed by Neyman [37] who used them in construction of his celebrated $C(\alpha)$-statistic.[1]

To describe the construction, suppose that the log-likelihood function associated to observation W is $(\theta, \beta) \mapsto \ell(W, \theta, \beta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ is the target parameter and $\beta \in T \subset \mathbb{R}^{p_0}$ is the nuisance parameter. Under regularity conditions, the true parameter values θ_0 and β_0 obey

$$\mathrm{E}_P\left[\partial_\theta \ell(W, \theta_0, \beta_0)\right] = 0, \quad \mathrm{E}_P\left[\partial_\beta \ell(W, \theta_0, \beta_0)\right] = 0. \quad (2.10)$$

Now consider the new score function

$$\psi(W, \theta, \eta) = \partial_\theta \ell(W, \theta, \beta) - \mu \partial_\beta \ell(W, \theta, \beta), \quad (2.11)$$

where the nuisance parameter is

$$\eta = (\beta', \ \mathrm{vec}\,(\mu)')' \in T \times \mathcal{D} \subset \mathbb{R}^p, \quad p = p_0 + dp_0,$$

μ is the $d \times p_0$ *orthogonalization* parameter matrix whose true value μ_0 solves the equation

$$J_{\theta\beta} - \mu J_{\beta\beta} = 0 \quad \left(\text{i.e., } \mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}\right),$$

And

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta', \beta')} \mathrm{E}_P\left[\partial_{(\theta', \beta')'} \ell(W, \theta, \beta)\right]\Big|_{\theta = \theta_0; \beta = \beta_0}.$$

[1]Note that the $C(\alpha)$-statistic, or the orthogonal score statistic, had been explicitly used for testing (and also for setting up estimation) in high-dimensional sparse models in [11] and in [39], where it is referred to as the decorrelated score statistic. The discussion of Neyman's construction here draws on [23]. Note also that our results cover other types of orthogonal score statistics as well, which allows us to cover much broader classes of models; see, for example, the discussion of conditional moment models with infinite-dimensional nuisance parameters below.

Provided that μ_0 is well defined, we have by (2.10) that $E_P[\psi(W, \theta_0, \eta_0)] = 0$, where $\eta_0 = (\beta_0', vec(\mu_0)')'$. Moreover, it is trivial to verify that under standard regularity conditions the score function ψ obeys the near orthogonality condition (2.2) exactly (i.e., with $C_0 = 0$), that is,

$$\partial_\eta E_P[\psi(W, \theta_0, \eta)]\big|_{\eta = \eta_0} = 0.$$

Note that in this example, μ_0 not only creates the necessary orthogonality but also creates the *efficient score* for inference on the main parameter θ, as emphasized by Neyman.

(2) *Orthogonal scores for likelihood problems with infinite-dimensional nuisance parameters*. The Neyman's construction can be extended to semi-parametric models, where the nuisance parameter β is a function. In this case, the original score functions $(\theta, \beta) \mapsto \partial_\theta \ell(W, \theta, \beta)$ corresponding to the log-likelihood function $(\theta, \beta) \mapsto \ell(W, \theta, \beta)$ associated to observation W can be transformed into efficient score functions ψ that obey the exact orthogonality condition (2.9) by projecting the original score functions onto the orthocomplement of the tangent space induced by the nuisance parameter β; see Chapter 25 of [44] for a detailed description of this construction. Note that the projection may create additional nuisance parameters, so that the new nuisance parameter η could be of larger dimension than β. Other relevant references include [9, 29, 45] and [11]. The approach is related to Neyman's construction in the sense that the score ψ arising in this model is actually the Neyman's score arising in a one-dimensional least favorable parametric subfamily, [42]; see Chapter 25 of [44] for details.

(3) *Orthogonal scores for conditional moment problems with infinite-dimensional nuisance parameters*. Next, consider a conditional moment restrictions framework studied by Chamberlain [14]. To define the framework, let W, D and V be random vectors in $\mathbb{R}^{d_W}$, $\mathbb{R}^{d_D}$ and $\mathbb{R}^{d_V}$, respectively, with D and V being subvectors of W, so that $d_D + d_V \leq d_W$. Also, let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be a finite-dimensional parameter whose true value θ_0 is of interest, and let $h: \mathbb{R}^{d_V} \to \mathbb{R}^{d_h}$ be a vectorvalued functional nuisance parameter, with the true value being $h_0: \mathbb{R}^{d_V} \to \mathbb{R}^{d_h}$. The conditional moment restrictions framework assumes that θ_0 and h_0 satisfy the following equation:

$$E_P[m(W, \theta_0, h_0(V))|D, V] = 0, \quad (2.12)$$

where $m: \mathbb{R}^{d_W} \times \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_h} \to \mathbb{R}^{d_m}$ is some known function. This framework is of interest because it covers an extremely rich variety of models, without having to explicitly rely on the likelihood formulation. For example, it covers the partial linear model

$$Y = D\theta_0 + h_0(V) + U, \quad \mathrm{E}_P[U|D, V] = 0, \quad (2.13)$$

where Y is a scalar dependent random variable, D is a scalar independent treatment random variable, V is a vector of control random variables and U is a scalar unobservable noise random variable. Indeed, (2.13) implies (2.12) by setting $W = (Y, D, V')'$ and $m(W, \theta, h) = Y - D\theta - h$.

Here, we would like to build a (generalized) score function $(\theta, \eta) \mapsto \psi(W, \theta, \eta)$ for estimating θ_0, the true value of parameter θ, where η is a new nuisance parameter with true value η_0, that obeys the near orthogonality condition (2.2). There are many ways to do so but one particularly useful way is the following. Consider the functional parameters $\Sigma: \mathbb{R}^{d_D + d_V} \to \mathbb{R}^{d_m \times d_m}$ and $\varphi: \mathbb{R}^{d_D + d_V} \to \mathbb{R}^{d_\theta \times d_m}$ whose true values are given by

$$\Sigma_0(D, V) = \mathrm{E}_P\left[m\left(W, \theta_0, h_0(V)\right)m\left(W, \theta_0, h_0(V)\right)'\middle|D, V\right],$$
$$\varphi_0(D, V) = \left(A_0(D, V) - \Gamma_0(D, V)G_0(V)\right)',$$

where

$$A_0(D, V) = \partial_{\theta'} \mathrm{E}_P\left[m\left(W, \theta, h_0(V)\right)\middle|D, V\right]\Big|_{\theta = \theta_0},$$
$$\Gamma_0(D, V) = \partial_{h'} \mathrm{E}_P\left[m\left(W, \theta_0, h\right)\middle|D, V\right]\Big|_{h = h_0(V)},$$
$$G_0(V) = \left(\mathrm{E}_P\left[\Gamma_0(D, V)'\Sigma_0(D, V)^{-1}\Gamma_0(D, V)\middle|V\right]\right)^{-1} \times \mathrm{E}_P\left[\Gamma_0(D, V)'\Sigma_0(D, V)^{-1}A_0(D, V)\middle|V\right].$$

Then set $\eta = (h, \varphi, \Sigma)$ and $\eta_0 = (h_0, \varphi_0, \Sigma_0)$, and define the score function:

$$\psi(W, \theta, \eta) = \underbrace{\varphi(D, V)}_{\text{"instrument"}} \underbrace{\Sigma(D, V)^{-1}}_{\text{weight}} \underbrace{m(W, \theta, h(V))}_{\text{residual}}.$$

It is rather straightforward to verify that under mild regularity conditions, the score function ψ satisfies the moment condition, $\mathrm{E}_P[\psi(W, \theta_0, \eta_0)] = 0$, and in addition, the orthogonality condition:

$$\partial_\eta \mathrm{E}_P\left[\psi\left(W, \theta_0, \eta\right)\right]\Big|_{\eta = \eta_0} = 0.$$

Note that this construction gives the efficient score function ψ that yields an estimator of θ_0 achieving the semiparametric efficiency bound, as calculated by Chamberlain [14].

3. Application to logistic regression model with functional response data.

In this section, we apply our main results to a logistic regression model with functional response data described in the Introduction.

3.1. Model.

We consider a response variable $Y \in \mathbb{R}$ that induces a functional response $(Y_u)_{u \in \mathcal{U}}$ by $Y_u = 1\{Y \leq (1-u)\underline{y} + u\bar{y}\}$ for a set of indices $\mathcal{U} = [0,1]$ and some constants $\underline{y} \leq \bar{y}$. We are interested in the dependence of this functional response on a $\tilde{p}$-vector of covariates, $D = (D_1, \ldots, D_{\tilde{p}})' \in \mathbb{R}^{\tilde{p}}$, controlling for a p-vector of additional covariates $X = (X_1, \ldots, X_p)' \in \mathbb{R}^p$. We allow both $\tilde{p}$ and p to be (much) larger than the sample size of available data, n.

For each $u \in \mathcal{U}$, we assume that Y_u satisfies the generalized linear model with the logistic link function

$$\mathrm{E}_P[Y_u | D, X] = \Lambda(D'\theta_u + X'\beta_u) + r_u, \quad (3.1)$$

where $\theta_{u1} = (\theta_{u1}, \ldots, \theta_{u\tilde{p}})'$ is a vector of parameters that are of interest, $\beta_u = (\beta u_1, \ldots, \beta_{up})'$ is a vector of nuisance parameters, $r_u = r_u(D, X)$ is an approximation error, $\Lambda: \mathbb{R} \to \mathbb{R}$ is the logistic link function defined by

$$\Lambda(t) = \frac{\exp(t)}{1 + \exp(t)}, \quad t \in \mathbb{R},$$

and $P \in \mathcal{P}_n$ is the distribution of the triple $W = (Y, D, X)$. As in the previous section, we construct simultaneous confidence bands for the parameters $(\theta_{uj})_{u \in \mathcal{U}, j \in [\tilde{p}]}$ based on a random sample $(W_i)_{i=1}^n = (Y_i, D_i, X_i)_{i=1}^n$ from the distribution of W = (Y, D, X).

3.2. Orthogonal score functions.

Before setting up score functions that satisfy both the moment conditions (1.2) and the orthogonality condition (1.4), observe that "naive" score functions that follow directly from the model (3.1),

$$m_{uj}(W, \theta_{uj}, \theta_{u[\tilde{p}]\setminus j}, \beta_u, r_u) = \{Y_u - \Lambda(D_j\theta + X^j(\theta'_{u[\tilde{p}]\setminus j}, \beta'_u)') - r_u\} D_j,$$

where $X^j = (D'_{[\tilde{p}]\setminus j}, X')$, satisfy the moment conditions $\mathrm{E}_P[m_{uj}(W, \theta_{uj})] = 0$ but violate the orthogonality condition (1.4) [with m_{uj} replacing ψ_{uj} and $\eta_{uj} = (\theta_{u[\tilde{p}]\setminus j}, \beta_u, r_u)$]. To satisfy the orthogonality condition (1.4), we proceed using an approach from Section 2.2 as in the Introduction. Specifically, for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, define the $(\tilde{p} + p - 1)$-vector of additional nuisance parameters γ_u^j by (1.5) where $f_u^2 = f_u^2(D, X)$ is defined in (1.6). Thus, by the first-order condition of (1.5), the nuisance parameters γ_u^j satisfy

$$f_u D_j = f_u X^j \gamma_u^j + v_u^j, \quad \mathrm{E}_P\left[f_u X^j v_u^j\right] = 0. \quad (3.2)$$

Also, denote $\beta_u^j = \left(\theta'_{u[\tilde{p}]\setminus j}, \beta'_u\right)'$. Then we set

$$\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right) = \left\{Y_u - \Lambda\left(D_j\theta_{uj} + X^j\beta_u^j\right) - r_u\right\}\left(D_j - X^j\gamma_u^j\right),$$

where the nuisance parameter is $\eta_{uj} = \left(r_u, \beta_u^j, \gamma_u^j\right)$. As we formally demonstrate in the proof of Theorem 3.1 below, this function satisfies the near-orthogonality condition (1.4).

3.3. Estimation using orthogonal score functions.

Next, we discuss estimation of η_{uj}'s and θ_{uj}'s. First, we assume that the approximation error $r_u = r_u(D, X)$ is asymptotically negligible, so that it can be estimated by $\mathcal{O} = \mathcal{O}(D, X)$, the identically zero function of D and X. Second, for γ_u^j, we consider an estimator $\tilde{\gamma}_u^j$ defined as a post-regularization weighted least squares estimator corresponding to the problem (1.5). Third, for β_u^j, we consider a plug-in estimator $\hat{\beta}_u^j = \left(\tilde{\theta}'_{u[\tilde{p}]\setminus j}, \tilde{\beta}'_u\right)$, where $\tilde{\theta}_u$ and $\tilde{\beta}_u$ are suitable estimators of θ_u and β_u. In particular, we assume that $\tilde{\theta}_u$ and $\tilde{\beta}_u$ are post-regularization maximum likelihood estimators corresponding to the log-likelihood function $(\theta, \beta) \mapsto -M_u(W, \theta, \beta)$ where

$$M_u(W, \theta, \beta) = -\left(1\{Y_u = 1\}\log\Lambda(D'\theta + X'\beta) + 1\{Y_u = 0\}\log\left(1 - \Lambda(D'\theta + X'\beta)\right)\right). \quad (3.3)$$

The details of the estimators $\tilde{\theta}_u$, $\tilde{\beta}_u$ and $\tilde{\gamma}_u^j$ are given in Algorithm 1 below. The results in this paper can also be easily extended to the case where $\tilde{\theta}_u$, $\tilde{\beta}_u$ and $\tilde{\gamma}_u^j$ are replaced by penalized maximum likelihood estimators $\hat{\theta}_u$ and $\hat{\beta}_u$ and penalized weighted least squares estimator $\hat{\gamma}_u^j$, respectively.

Then our estimator of η_{uj} is $\hat{\eta}_{uj} = \left(\mathcal{O}, \hat{\beta}_u^j, \tilde{\gamma}_u^j\right)$. Substituting this estimator into the score function ψ_{uj} gives

$$\psi_{uj}\left(W, \theta_{uj}, \hat{\eta}_{uj}\right) = \left\{Y_u - \Lambda\left(D_j\theta_{uj} + X^j\hat{\beta}_u^j\right)\right\}\left(D_j - X^j\tilde{\gamma}_u^j\right), \quad (3.4)$$

which, using the sample analog (2.1) of the moment conditions (1.2), gives the following estimator of θ_{uj}:

$$\check{\theta}_{uj} \in \arg \inf_{\theta \in \Theta_{uj}} \left| \mathbb{E}_n \left[\psi_{uj} \left(W, \theta, \hat{\eta}_{uj} \right) \right] \right|. \quad (3.5)$$

The algorithm is summarized as follows.

ALGORITHM 1. For each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$:

Step 1. Run post-ℓ_1-penalized logistic estimator (4.2) of Y_u on D and X to compute $(\tilde{\theta}_u, \tilde{\beta}_u)$.

Step 2. Define the weights $\hat{f}_u^2 = \hat{f}_u^2(D, X) = \Lambda'\left(D'\tilde{\theta}_u + X'\tilde{\beta}_u\right)$.

Step 3. Run the Post-Lasso estimator (4.5) of $\widehat{f_u D_j}$ to compute $\tilde{\gamma}_u^j$.

Step 4. Compute $\hat{\beta}_u^j = \left(\tilde{\theta}'_{u[\tilde{p}]\setminus j}, \tilde{\beta}'_u\right)'$.

Step 5. Solve (3.5) with $\psi_{uj}\left(W, \theta, \hat{\eta}_{uj}\right)$ defined in (3.4) to compute $\check{\theta}_{uj}$.

3.4. Regularity conditions.

Next, we specify our regularity conditions. For all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, denote $Z_u^j = D_j - X^j \gamma_u^j$. Also, denote $a_n = p \vee \tilde{p} \vee n$. Let q, c_1 and C_1 be some strictly positive (and finite) constants where $q > 4$. Moreover, let $(\delta_n)_{n \geq 1}$ and $(\Delta_n)_{n \geq 1}$ be some sequences of positive constants converging to zero. Finally, let $(M_{n,1})_{n \geq 1}$ and $(M_{n,2})_{n \geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $M_{n,1} \geq 1$ and $M_{n,2} \geq 1$ for all n.

ASSUMPTION 3.1 (Parameters). For all $u \in \mathcal{U}$, we have $\left\|\theta_u\right\| + \left\|\beta_u\right\| + \max_{j \in [\tilde{p}]} \left\|\gamma_u^j\right\| \leq C_1$ and $\max_{j \in [\tilde{p}]} \sup_{\theta \in \Theta_{uj}} |\theta| \leq C_1$. In addition, for all $u_1, u_2 \in \mathcal{U}$, we have $\left(\left\|\theta_{u_2} - \theta_{u_1}\right\| + \left\|\beta_{u_2} - \beta_{u_1}\right\| \right) \leq C_1 |u_2 - u_1|$. Finally, for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, Θ_{uj} contains a ball of radius $(\log \log n)(\log a_n)^{3/2}/n^{1/2}$ centered at θ_{uj}.

ASSUMPTION 3.2 (Sparsity). There exist $s = s_n$ and $\bar{\gamma}_u^j$, $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, such that for all $u \in \mathcal{U}$, $\left\|\beta_u\right\|_0 + \left\|\theta_u\right\|_0 + \max_{j \in [\tilde{p}]} \left\|\bar{\gamma}_u^j\right\|_0 \leq s_n$ and $\max_{j \in [\tilde{p}]} \left(\left\|\bar{\gamma}_u^j - \gamma_u^j\right\| + s_n^{-1/2} \left\|\bar{\gamma}_u^j - \gamma_u^j\right\|_1 \right) \leq C_1 \left(s_n \log a_n / n\right)^{1/2}$.

ASSUMPTION 3.3 (Distribution of Y). The conditional pdf of Y given (D, X) is bounded by C_1.

Assumptions 3.1–3.3 are mild and standard in the literature. In particular, Assumption 3.1 requires the parameter spaces Θ_{uj} to be bounded, and also requires that for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, the parameter θ_{uj} to be sufficiently separated from the boundaries of the parameter space Θ_{uj}. Assumption 3.2 requires approximate sparsity of the model (3.1). Note that in

Assumption 3.2, given that $\bar{\gamma}_u^j$'s exist, we can and will assume without loss of generality that $\bar{\gamma}_u^j = \gamma_{uT}^j$ for some $T \subset \{1, \ldots, p + \tilde{p} - 1\}$ with $|T| \leq s_n$, where $T = T_u^j$ is allowed to depend on u and j. Here, the $(p + \tilde{p} - 1)$-vector γ_{uT}^j is defined from γ_u^j by keeping all components of γ_u^j that are in T and setting all other components to be zero. Assumption 3.3 can be relaxed at the expense of more technicalities.

Assumption 3.4 (Covariates). For all $u \in \mathcal{U}$, the following inequalities hold: (i) $\inf_{\|\xi\|=1} \mathrm{E}_P\left[f_u^2\{(D', X')\xi\}^2\right] \geq c_1$, (ii) $\min_{j,k}\left(\mathrm{E}_P\left[\left|f_u^2 Z_u^j X_k^j\right|^2\right] \wedge \mathrm{E}_P\left[\left|f_u^2 D_j X_k^j\right|^2\right]\right) \geq c_1$, and (iii) $\max_{j,k} \mathrm{E}_P\left[\left|Z_u^j X_k^j\right|^3\right]^{1/3} \log^{1/2} a_n \leq \delta_n n^{1/6}$. In addition, we have that (iv) $\sup_{\|\xi\|=1} \mathrm{E}_P\left[\{(D', X')\xi\}^4\right] \leq C_1$, (v) $M_{n,1} \geq \mathrm{E}_P\left[\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left|Z_u^j\right|^{2q}\right]^{1/(2q)}$ (vi) $M_{n,1}^2 s_n \log a_n \leq \delta_n n^{1/2 - 1/q}$, (vii) $M_{n,2} \geq \left\{\mathrm{E}_P\left[\left(\|D\|_\infty \vee \|X\|_\infty\right)^{2q}\right]\right\}^{1/(2q)}$, (viii) $M_{n,2}^2 s_n \log^{1/2} a_n \leq \delta_n n^{1/2 - 1/q}$ and (ix) $M_{n,1}^2 M_{n,2}^4 s_n \leq \delta_n n^{1 - 3/q}$.

This assumption requires that there is no multicollinearity between covariates in vectors D and X. In addition, it requires that the constants $\underline{y}$ and $\bar{y}$ are chosen so that the probabilities of $Y < \bar{y}$ and $Y > \underline{y}$ are both nonvanishing since otherwise we would have $\mathrm{E}\left[f_u^2\right] = \mathrm{E}\left[\mathrm{Var}_P\left(Y_u \mid D, X\right)\right]$ vanishing either for $u = 0$ or $u = 1$ violating Assumption 3.4(i). Intuitively, sending $\underline{y}$ and $\bar{y}$ to the left and to the right tails of the distribution of Y, respectively, would blow up the variance of the estimators $\check{\theta}_{uj}$, given by σ_{uj}^2 in Theorem 2.1, and leading eventually to the estimators with slower-than-$\sqrt{n}$ rate of convergence. Although our results could be extended to allow for the case where $\underline{y}$ and $\bar{y}$ are sent to the tails of the distribution of Y slowly, we skip this extension for the sake of clarity. Moreover, Assumption 3.4 imposes constraints on various moments of covariates. Since these constraints might be difficult to grasp, at the end of this section, in Corollary 3.3, we provide an example for which these constraints simplify into easily interpretable conditions.

Assumption 3.5 (Approximation error). For all $u \in \mathcal{U}$, we have (i) $\sup_{\|\xi\|=1} \mathrm{E}_P\left[r_u^2\{(D', X')\xi\}^2\right] \leq C_1 \mathrm{E}_P\left[r_u^2\right]$, (ii) $\mathrm{E}_P\left[r_u^2\right] \leq C_1 s_n \log a_n / n$, (iii) $\max_{j \in [\tilde{p}]} \left|E_P\left[r_u Z_u^j\right]\right| \leq \delta_n n^{-1/2}$, and (iv) $\left|r_u(D, X)\right| \leq f_u^2(D, X)/4$ almost surely. In addition, with probability $1 - \overline{\Delta}_n$, (v) $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left(\mathbb{E}_n\left[\left(r_u Z_u^j / f_u\right)^2\right] + \mathbb{E}_n\left[r_u^2 / f_u^6\right]\right) \leq C_1 s_n \log a_n / n$.

This assumption requires the approximation error $r_u = r_u(D, X)$ to be sufficiently small. Under Assumption 3.4, the first condition of Assumption 3.5 holds if the approximation error is such that $r_u^2 \leq C \mathrm{E}_P\left[r_u^2\right]$ almost surely for some constant C.

3.5. Formal results.

Under specified assumptions, our estimators $\check{\theta}_{uj}$ satisfy the following uniform Bahadur representation theorem.

THEOREM 3. 1 (Uniform Bahadur representation for logistic model). *Suppose that Assumptions* 3.1–3.5 *hold for all* $P \in \mathcal{P}_n$. *In addition, suppose that the following growth condition holds*: $\delta_n^2 \log a_n = o(1)$. *Then for the estimators* $\check{\theta}_{uj}$ *satisfying* (3.5), *we have*

$$\sqrt{n}\sigma_{uj}^{-1}\left(\check{\theta}_{uj} - \theta_{uj}\right) = \mathbb{G}_n \overline{\psi}_{uj} + O_P(\delta_n) \quad (3.6)$$

in $\ell^\infty(\mathcal{U} \times [\tilde{p}])$, *uniformly over* $P \in \mathcal{P}_n$, *where* $\overline{\psi}_{uj}(W) := -\sigma_{uj}^{-1} J_{uj}^{-1} \psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)$, $\sigma_{uj}^2 := \mathrm{E}_P\left[J_{uj}^{-2}\psi_{uj}^2\left(W, \theta_{uj}, \eta_{uj}\right)\right]$, *and* J_{uj} *is defined in* (2.3).

This theorem allows us to establish a Gaussian approximation result for the supremum of the process $\left\{\sqrt{n}\sigma_{uj}^{-1}\left(\check{\theta}_{uj} - \theta_{uj}\right) : u \in \mathcal{U}, j \in [\tilde{p}]\right\}$:

COROLLARY 3.1 (Gaussian approximation for logistic model). *Suppose that Assumptions* 3.1–3.5 *hold for all* $P \in \mathcal{P}_n$. *In addition, suppose that the following growth conditions hold*: $\delta_n^2 \log a_n = o(1)$, $M_{n,1}^{2/7} \log a_n = o\left(n^{1/7}\right)$ *and* $M_{n,1}^{2/3} \log a_n = o\left(n^{1/3 - 2/(3q)}\right)$. *Then*

$$\sup_{t \in \mathbb{R}} \left| \mathrm{P}_P\left(\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| \sqrt{n}\sigma_{uj}^{-1}\left(\check{\theta}_{uj} - \theta_{uj}\right) \right| \le t \right) - \mathrm{P}_P\left(\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| \mathcal{N}_{uj} \right| \le t \right) \right| = o(1)$$

uniformly over $P \in \mathcal{P}_n$, *where* $\left(\mathcal{N}_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ *is a tight zero-mean Gaussian process indexed by* $\mathcal{U} \times [\tilde{p}]$ *with the covariance given by* $\mathrm{E}_P\left[\overline{\psi}_{uj}(W)\overline{\psi}_{u'j'}(W)\right]$ *for* $u, u' \in \mathcal{U}$ *and* $j, j' \in [\tilde{p}]$.

Based on this corollary, we are now able to construct simultaneous confidence bands for the parameters θ_{uj}. Observe that

$$J_{uj} = -\mathrm{E}_P\left[\Lambda'\left(D_j\theta_{uj} + X^j\beta_u^j\right)D_j\left(D_j - X^j\gamma_u^j\right)\right], \quad u \in \mathcal{U}, j \in [\tilde{p}],$$

and so it can be estimated by

$$\hat{J}_{uj} = -\mathbb{E}_n\left[\Lambda'\left(D_j\tilde{\theta}_{uj} + X^j\hat{\beta}_u^j\right)D_j\left(D_j - X^j\tilde{\gamma}_u^j\right)\right], \quad u \in \mathcal{U}, j \in [\tilde{p}].$$

In addition, $\sigma_{uj}^2 = \mathrm{E}_P\left[J_{uj}^{-2}\psi_{uj}^2\left(W, \theta_{uj}, \eta_{uj}\right)\right]$, and so it can be estimated by

$$\hat{\sigma}_{uj}^2 = \mathbb{E}_n\left[\hat{J}_{uj}^{-2}\psi_{uj}^2\left(W, \tilde{\theta}_{uj}, \hat{\eta}_{uj}\right)\right], \quad u \in \mathcal{U}, j \in [\tilde{p}].$$

Moreover, as in Section 2, define $\hat{\psi}_{uj}(W) = -\hat{\sigma}_{uj}^{-1}\hat{J}_{uj}^{-1}\psi_{uj}\left(W, \check{\theta}_{uj}, \hat{\eta}_{uj}\right)$, and let c_α be the $(1-\alpha)$ quantile of the conditional distribution of $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left|\widehat{\mathcal{G}}_{uj}\right|$ given the data $(W_i)_{i=1}^n$ where the process $\widehat{\mathcal{G}} = \left(\widehat{\mathcal{G}}_{uj}\right)_{u \in \mathcal{U}, j \in [\tilde{p}]}$ is defined in (2.8). Then we have the following.

Corollary 3.2 (Simultaneous confidence bands for logistic model). *Suppose that Assumptions* 3.1–3.5 *hold for all* $P \in \mathcal{P}_n$. *In addition, suppose that the following growth conditions hold:* $\delta_n^2 \log a_n = o(1)$, $M_{n,1}^{2/7} \log a_n = o\left(n^{1/7}\right)$, $M_{n,1}^{2/3} \log a_n = o\left(n^{1/3 - 2/(3q)}\right)$ *and* $s_n \log^3 a_n = o(n)$. *Then* (1.3) *holds uniformly over* $P \in \mathcal{P}_n$.

To conclude this section, we provide an example for which conditions of Corollary 3.2 are easy to interpret. Recall that $a_n = n \vee p \vee \tilde{p}$.

Corollary 3.3 (Uniform confidence bands for logistic regression model under simple conditions). *Suppose that Assumptions* 3.1–3.3, 3.4(i,ii,iv) *and* 3.5(i,ii,iv,v) *hold for* $q > 4$ *for all* $P \in \mathcal{P}_n$. *In addition, suppose that* $\left\{\mathrm{E}_P\left[\left(\|D\|_\infty \vee \|X\|_\infty\right)^{2q}\right]\right\}^{1/(2q)} \leq C_1$ *and* $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left\|\gamma_u^j\right\|_1 \leq C_1$. *Moreover, suppose that the following growth conditions hold:* $\log^7 a_n / n = o(1)$, $s_n^2 \log^3 a_n / n^{1-2/q} = o(1)$ *and* $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left|\mathrm{E}_P\left[r_u Z_u^J\right]\right| = o\left(\left(n \log a_n\right)^{-1/2}\right)$. *Then* (1.3) *holds uniformly over* $P \in \mathcal{P}_n$.

Comment 3.1 (Estimation of variance). When constructing the confidence bands based on (1.3), we find in simulations that it is beneficial to replace the estimators $\hat{\sigma}_{uj}^2$ of σ_{uj}^2 by $\max\left\{\hat{\sigma}_{uj}^2, \hat{\Sigma}_{uj}^2\right\}$ where $\hat{\Sigma}_{uj}^2 = \mathbb{E}_n\left[\hat{f}_u^2\left(D - X^j\tilde{\gamma}_u^j\right)^2\right]$ is an alternative consistent estimator of σ_{uj}^2.

Comment 3.2 (Alternative implementations, double selection). We note that the theory developed here is applicable for different estimators that construct the new score function with the desired orthogonality condition implicitly. For example, the double selection idea yields an implementation of an estimator that is first-order equivalent to the estimator based on the score function. The algorithm yielding the double selection estimator is as follows.

ALGORITHM 2. For each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$:

Step 1′. Run post-ℓ_1-penalized logistic estimator (4.2) of Y_u on D and X to compute $\left(\tilde{\theta}_u, \tilde{\beta}_u\right)$.

Step 2′. Define the weights $\hat{f}_u^2 = \hat{f}_u^2(D, X) = \Lambda'\left(D_i'\tilde{\theta}_u + X_i'\tilde{\beta}_u\right)$.

Step 3′. Run the Lasso estimator (4.4) of $\widehat{f_u D_j}$ on $\widehat{f_u X}$ to compute $\hat{\gamma}_u^j$.

Step 4′. Run logistic regression of Y_u on D_j and all the selected variables in Steps $1'$ and $3'$ to compute $\check{\theta}_{uj}$.

As mentioned by a referee, it is surprising that the double selection procedure has uniform validity. The use of the additional variables selected in Step $3'$, through the first-order conditions of the optimization problem, induces the necessary nearorthogonality condition. We refer to the Supplementary Material for a more detailed discussion.

COMMENT 3.3 (Alternative implementations, one-step correction). Another implementation for which the theory developed here applies is to replace Step 5 in Algorithm 1 with a one-step procedure. This relates to the debiasing procedure proposed in [43] to the case when the set $\mathcal{U}$ is a singleton. In this case, instead of minimizing the criterion (3.5) in Step 5, the method makes a full Newton step from the initial estimate,

Step $5''$. Compute $\bar{\theta}_{uj} = \hat{\theta}_{uj} - \hat{J}_{uj}^{-1}\mathbb{E}_n\left[\psi_{uj}\left(W, \hat{\theta}_{uj}, \hat{\eta}_{uj}\right)\right]$.

The theory developed here directly apply to those estimators as well.

COMMENT 3.4 (Extension to other approximately sparse generalized linear models). Inspecting the proofs of Theorem 3.1 and Corollaries 3.1–3.3 reveal that these results can be extended with minor modifications to cover other approximately sparse generalized linear models. For example, the results can be extended to cover the model (3.1) where we use the probit link function instead of the logit link function Λ.

4. ℓ_1-Penalized M-estimators: Nuisance functions and functional data.

In this section, we define the estimators $\tilde{\theta}_u$, $\tilde{\beta}_u$ and $\tilde{\gamma}_u^j$, which were used in the previous section, and study their properties. We consider the same setting as that in the previous section. The results in this section rely upon a set of new results for ℓ_1-penalized M-estimators with functional data presented in Appendix M of the Supplementary Material.

4.1. ℓ_1-Penalized logistic regression for functional response data: Asymptotic properties.

Here, we consider the generalized linear model with the logistic link function and functional response data (3.1). As explained in the previous section, we assume that $\tilde{\theta}_u$ and $\tilde{\beta}_u$ are post-regularization maximum likelihood estimators of θ_u and β_u corresponding to the log-likelihood function $M_u(W, \theta, \beta) = M_u(Y_u, D, X, \theta, \beta)$ defined in (3.3). To define these estimators, let $\hat{\theta}_u$ and $\hat{\beta}_u$ be ℓ_1-penalized maximum likelihood (logistic regression) estimators

$$\left(\hat{\theta}_u, \hat{\beta}_u\right) \in \operatorname*{argmin}_{\theta,\beta}\left(\mathbb{E}_n\left[M_u\left(Y_u, D, X, \theta, \beta\right)\right] + \frac{\lambda}{n}\left\|\hat{\Psi}_u(\theta', \beta')'\right\|_1\right), \quad (4.1)$$

where λ is a penalty level and $\hat{\Psi}_u$ a diagonal matrix of penalty loadings. We choose parameters λ and $\hat{\Psi}_u$ according to Algorithm 3 described below. Using the ℓ_1 - penalized estimators $\hat{\theta}_u$ and $\hat{\beta}_u$, we then define post-regularization estimators $\tilde{\theta}_u$ and $\tilde{\beta}_u$ by

$$\left(\widetilde{\theta}_u, \widetilde{\beta}_u\right) \in \operatorname{argmin}_{\theta} \mathbb{E}_n\left[M_u\left(Y_u, D, X, \theta, \beta\right)\right]: \; supp\,(\theta, \beta) \subseteq \; supp\left(\hat{\theta}_u, \hat{\beta}_u\right). \quad (4.2)$$

We derive the rate of convergence and sparsity properties of $\widetilde{\theta}_u$ and $\widetilde{\beta}_u$ as well as of $\hat{\theta}_u$ and $\hat{\beta}_u$ in Theorem 4.1 below. Recall that $a_n = n \vee p \vee \tilde{p}$.

Algorithm 3 (Penalty level and loadings for logistic regression). Choose $\gamma \in [1/n, 1/\log n]$ and $c > 1$ (in practice, we set $c = 1.1$ and $\gamma = 0.1/\log n$). Define $\lambda = c\sqrt{n}\Phi^{-1}\left(1 - \gamma/\left(2(p+\tilde{p})N_n\right)\right)$ with $N_n = n$. To select $\widehat{\Psi}_u$, choose a constant $\overline{m} \geq 0$ as an upper bound on the number of loops and proceed as follows: (0) Let $\widetilde{X} = (D', X')'$, $m = 0$ and initialize $\hat{l}_{uk,0} = \frac{1}{2}\left\{\mathbb{E}_n\left[\widetilde{X}_k^2\right]\right\}^{1/2}$ for $k \in [p + \tilde{p}]$. (1) Compute $\left(\hat{\theta}_u, \hat{\beta}_u\right)$ and $\left(\widetilde{\theta}_u, \widetilde{\beta}_u\right)$ based on $\widehat{\Psi}_{uu} = diag\left(\left\{\hat{l}_{uk,m}, k \in [p+\tilde{p}]\right\}\right)$. (2) Set $\hat{l}_{uk,m+1} := \left\{\mathbb{E}_n\left[\widetilde{X}_k^2\left(Y_u - \Lambda\left(D'\widetilde{\theta}_u + X'\widetilde{\beta}_u\right)\right)^2\right]\right\}^{1/2}$. (3) If $m \geq \overline{m}$, report the current value of $\widehat{\Psi}_{uu}$ and stop; otherwise set $m \leftarrow m + 1$ and go to step (1).

Theorem 4.1 (Rates and sparsity for functional response under logistic link). *Suppose that Assumptions* 3.1–3.5 *hold for all* $P \in \mathcal{P}_n$. *In addition, suppose that the penalty level* λ *and the matrices of penalty loadings* $\widehat{\Psi}_{uu}$ *are chosen according to Algorithm* 3. *Moreover, suppose that the following growth condition holds:* $\delta_n^2 \log a_n = o(1)$. *Then there exists a constant* $\bar{C}$ *such that uniformly over all* $P \in \mathcal{P}_n$ *with probability* $1 - o(1)$,

$$\sup_{u \in \mathcal{U}}\left(\left\|\hat{\theta}_u - \theta_u\right\| + \left\|\hat{\beta}_u - \beta_u\right\|\right) \leq \bar{C}\sqrt{\frac{s_n \log a_n}{n}},$$

$$\sup_{u \in \mathcal{U}}\left(\left\|\hat{\theta}_u - \theta_u\right\|_1 + \left\|\hat{\beta}_u - \beta_u\right\|_1\right) \leq \bar{C}\sqrt{\frac{s_n^2 \log a_n}{n}},$$

and the estimators $\hat{\theta}_u$ and $\hat{\beta}_u$ *are uniformly sparse*: $\sup_{u \in \mathcal{U}}\left\|\hat{\theta}_u\right\|_0 + \left\|\hat{\beta}_u\right\|_0 \leq \bar{C}s_n$. *Also, uniformly overall* $P \in \mathcal{P}_n$, *with probability* $1 - o(1)$,

$$\sup_{u \in \mathcal{U}}\left(\left\|\widetilde{\theta}_u - \theta_u\right\| + \left\|\widetilde{\beta}_u - \beta_u\right\|\right) \leq \bar{C}\sqrt{\frac{s_n \log a_n}{n}},$$

$$\sup_{u \in \mathcal{U}}\left(\left\|\widetilde{\theta}_u - \theta_u\right\|_1 + \left\|\widetilde{\beta}_u - \beta_u\right\|_1\right) \leq \bar{C}\sqrt{\frac{s_n^2 \log a_n}{n}}.$$

4.2. Lasso with estimated weights: Asymptotic properties.

Here, we consider the weighted linear model (3.2) for $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Using the parameter $\bar{\gamma}_u^j$ appearing in Assumption 3.2, it will be convenient to rewrite this model as

$$f_u D_j = f_u X^j \bar{\gamma}_u^j + f_u \bar{r}_{uj} + v_u^j, \quad \mathrm{E}_P\left[f_u X^j v_u^j\right] = 0, \quad (4.3)$$

where $\bar{r}_{uj} = X^j\left(\gamma_u^j - \bar{\gamma}_u^j\right)$ is an approximation error, which is asymptotically negligible under Assumption 3.2. As explained in the previous section, we assume that $\tilde{\gamma}_u^j$ is a post-regularization weighted least squares estimator of γ_u^j (or $\bar{\gamma}_u^j$). To define this estimator, let $\hat{\gamma}_u^j$ be an ℓ_1-penalized (weighted Lasso) estimator

$$\hat{\gamma}_u^j \in \arg\min_\gamma \left(\frac{1}{2}\mathbb{E}_n\left[\hat{f}_u^2\left(D_j - X^j\gamma\right)^2\right] + \frac{\lambda}{n}\left\|\hat{\Psi}_{uj}\gamma\right\|_1\right), \quad (4.4)$$

where λ and $\hat{\Psi}_{uuj}$ are the associated penalty level and the diagonal matrix of penalty loadings specified below in Algorithm 4 and where $\hat{f}_u^2$'s are estimated weights. As in Algorithm 1 in the previous section, we set $\hat{f}_u^2 = \hat{f}_u^2(D, X) = \Lambda'\left(D'\tilde{\theta}_u + X'\tilde{\beta}_u\right)$. Using $\hat{\gamma}_u^j$, we define a post-regularized weighted least squares estimator:

$$\tilde{\gamma}_u^j \in \arg\min_\gamma \frac{1}{2}\mathbb{E}_n\left[\hat{f}_u^2\left(D_j - X^j\gamma\right)^2\right] : \mathrm{supp}(\gamma) \subseteq \mathrm{supp}\left(\hat{\gamma}_u^j\right). \quad (4.5)$$

We derive the rate of convergence and sparsity properties of $\tilde{\gamma}_u^j$ as well as of $\hat{\gamma}_u^j$ in Theorem 4.2 below.

Algorithm 4 (Penalty level and loadings for weighted Lasso). Choose $\gamma \in [1/n, 1/\log n]$ and $c > 1$ (in practice, we set $c = 1.1$ and $\gamma = 0.1/\log n$). Define $\lambda = c\sqrt{n}\Phi^{-1}\left(1 - \gamma/\left(2(p + \tilde{p})N_n\right)\right)$ with $N_n = p\tilde{p}^2n^2$. To select $\hat{\Psi}_{uuj}$, choose a constant $\bar{m} \geq 1$ as an upper bound on the number of loops and proceed as follows: (0) Set $m = 0$ and $\hat{l}_{ujk,0} = \max_{1 \leq i \leq n}\left\|\hat{f}_{ui}X_i^j\right\|_\infty \left\{\mathbb{E}_n\left[\hat{f}_u^2 D_j^2\right]\right\}^{1/2}$. (1) Compute $\hat{\gamma}_u^j$ and $\tilde{\gamma}_u^j$ based on $\hat{\Psi}_{uj} = \mathrm{diag}\left(\left\{\hat{l}_{ujk,m}, k \in [p + \tilde{p} - 1]\right\}\right)$. (2) Set $\hat{l}_{ujk,m+1} := \left\{\mathbb{E}_n\left[\hat{f}_u^4\left(D_j - X^j\tilde{\gamma}_u^j\right)^2\left(X_k^j\right)^2\right]\right\}^{1/2}$. (3) If $m \geq \bar{m}$, report the current value of $\hat{\Psi}_{uuj}$ and stop; otherwise set $m \leftarrow m + 1$ and go to step (1).

Theorem 4.2 (Rates and sparsity for Lasso with estimated weights). *Suppose that Assumptions* 3.1–3.5 *hold for all* $P \in \mathcal{P}_n$. *In addition, suppose that the penalty level* λ *and the matrices of penalty loadings* $\hat{\Psi}_{uj}$ *are chosen according to Algorithm* 4. *Moreover, suppose that the following growth condition holds*: $\delta_n^2 \log a_n = o(1)$. *Then there exists a constant* $\bar{C}$ *such that uniformly over all* $P \in \mathcal{P}_n$ *with probability* $1 - o(1)$,

$$\max_{j \in [\tilde{p}]} \sup_{u \in \mathcal{U}} \left\| \hat{\gamma}_u^j - \bar{\gamma}_u^j \right\| \le \bar{C}\sqrt{\frac{s_n \log a_n}{n}}, \qquad \max_{j \in [\tilde{p}]} \sup_{u \in \mathcal{U}} \left\| \hat{\gamma}_u^j - \bar{\gamma}_u^j \right\|_1 \le \bar{C}\sqrt{\frac{s_n^2 \log a_n}{n}},$$

and the estimator $\hat{\gamma}_u^j$ *is uniformly sparse*, $\max_{j \in [\tilde{p}]} \sup_{u \in \mathcal{U}} \left\| \hat{\gamma}_u^j \right\|_0 \le \bar{C} s_n$. *Also, uniformly over all* $P \in \mathcal{P}_n$, *with probability* $1 - o(1)$,

$$\max_{j \in [\tilde{p}]} \sup_{u \in \mathcal{U}} \left\| \tilde{\gamma}_u^j - \bar{\gamma}_u^j \right\| \le \bar{C}\sqrt{\frac{s_n \log a_n}{n}}, \qquad \max_{j \in [\tilde{p}]} \sup_{u \in \mathcal{U}} \left\| \tilde{\gamma}_u^j - \bar{\gamma}_u^j \right\|_1 \le \bar{C}\sqrt{\frac{s_n^2 \log a_n}{n}}.$$

5. Monte Carlo simulations.

Here, we investigate the finite sample properties of the proposed estimators and the associated confidence regions. We report only the performance of the estimator based on the double selection procedure due to space constraints and note that it is very similar to the performance of the estimator based on score functions with near-orthogonality property. We will compare the proposed procedure with the traditional estimator that refits the model selected by the corresponding ℓ_1-penalized M-estimator (naive post-selection estimator).

We consider a logistic regression model with functional response data where the response $Y_u = 1\{y \le u\}$ for $u \in \mathcal{U}$ a compact set. We specify two different designs: (1) a location model, $y = x'\beta_0 + \xi$, where ξ is distributed as a logistic random variable, the first component of x is the intercept and the other $p - 1$ components are distributed as $N(0, \Sigma)$ with $E_{k,j} = |0.5|^{|k-j|}$; (2) a location-shift model, $y = \{(x'\beta_0 + \xi)/x'\vartheta_0\}^3$, where ξ is distributed as alogistic random variable, $x_j = |w_j|$ where w is a p-vector distributed as $N(0, \Sigma)$ with $\Sigma_{k,j} = |0.5|^{|k-j|}$, and ϑ_0 has nonnegative components. Such specification implies that for each $u \in \mathcal{U}$:

Design1: $\theta_u = u(1,0,\ldots,0)' - \beta_0$ and Design 2: $\theta_u = u^{1/3}\vartheta_0 - \beta_0$. In our simulations, we will consider $n = 500$ and $p = 2000$. For the location model (Design 1), we will consider two different choices for β_0: (i)$\beta_{0j}^{(i)} = 2/j^2$ for $j = 1,\ldots,p$, and (ii) $\beta_{0j}^{(ii)} = (1/2)/\{j - 3.5\}^2$ for $j > 1$ with the intercept coefficient $\beta_{0j}^{(ii)} = -10$. [These choices ensure $\max_{j>1} |\beta_{0j}| = 2$ and that y is around zero in Design 2(ii).] We set $\vartheta_0 = \frac{1}{8}(1, 1, 1, 1, 0, 0, \ldots, 0, 0, 1, 1, 1, 1)'$. For Design 1, we have $\mathcal{U} = [1, 2.5]$ and for Design 2 we have $\mathcal{U} = [-0.5, 0.5]$. The results are based on 500 replications (the bootstrap procedure is performed 5000 times for each replication).

We report the (empirical) rejection frequencies for confidence regions with 95% nominal coverage, so that 0.05 is the target rejection frequency. We report the rejection frequencies for the proposed estimator and the post-naive selection estimator. Table 1 presents the performance of the methods when applied to construct a confidence interval for a single parameter ($\tilde{p} = 1$ and $\mathcal{U}$ is a singleton). Since the setting is not symmetric, we investigate the performance for different components. Specifically, we consider $\{u\} \times \{j\}$ for $j = 1,\ldots, 5$.

First, consider the location model (Design 1). The difference between the performance of the naive estimator for Design 1(i) and 1(ii) highlights its fragile performance which is highly dependent on the unknown parameters. We can see from Table 1 that in Design 1(i) the Naive method achieve (pointwise) rejection frequencies up to 0.162 when the nominal level is 0.05. In Design 1(ii), it can be as high as 0.886. We also note that it is important to look at the performance of each component and avoid averaging across components (large j components are essentially not in the model, indeed for $j > 50$ we obtain rejection frequencies very close to 0.05 regardless of the model selection procedure). In contrast, the proposed estimator exhibits a much more robust behavior. For Design 1(i), the rejection frequencies are between 0.040 and 0.062 while for Design 1(ii) the rejection frequencies of the proposed estimator are between 0.040 and 0.056.

Table 2 presents the performance for simultaneous confidence bands of the form $\left\{\left[\widetilde{\theta}_{uj} - cv\,\tilde{\sigma}_{uj}, \widetilde{\theta}_{uj} + cv\,\tilde{\sigma}_{uj}\right] \text{ for } u \in \mathcal{U} \times [\tilde{p}]\right\}$ where $\widetilde{\theta}_{uj}$ is a point estimate, $\tilde{\sigma}_{uj}$ is an estimate of the pointwise standard deviation and cv is a critical value that accounts for the uniform estimation. For the point estimate, we consider the proposed estimator and the post-naive selection estimator which have estimates of standard deviation. We consider two critical values: from the multiplier bootstrap (MB) procedure and the Bonferroni (BF) correction (which we expect to be conservative). For each of the four different designs [1(i), 1(ii), 2(i) and 2(ii) described above], we consider four different choices of $\mathcal{U} \times [\tilde{p}]$. Table 2 displays rejection frequencies for confidence regions with 95% nominal coverage (and again 0.05 would be the ideal performance). The simulation results confirms the differences between the performance of the methods and overall the proposed procedure is closer to the nominal value of 0.05. The proposed estimator performed within a factor of two to the nominal value in 10 out of the 16 designs considered (and 13 out 16 within a factor of three). The post-naive selection estimator performed within a factor of two only in 3 out of the 16 designs when using the multiplier bootstrap as critical value (7 out of 16 within a factor of three) and similarly with the Bonferroni correction as the critical value.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

APPENDIX A:: PROOFS FOR SECTION 2

In this appendix, we use C to denote a strictly positive constant that is independent of n and $P \in \mathcal{P}_n$. The value of C may change at each appearance. Also, the notation $an \lesssim b_n$ means that $an \leq Cb_n$ for all n and some C. The notation $a_n \gtrsim b_n$ means that $b_n \lesssim an$. Moreover, the notation $a_n = o(1)$ means that there exists a sequence $(b_n)_{n \geq 1}$ of positive numbers such that (i) $|a_n| \leq b_n$ for all n, (ii) b_n is independent of $P \in \mathcal{P}_n$ for all n and (iii) $b_n \to 0$ as $n \to \infty$. Finally, the notation $an = O_p(b_n)$ means that for all $\epsilon > 0$, there exists C such that $P_p(a_n > Cb_n) \leq 1 - \epsilon$ for all n. Using this notation allows us to avoid repeating "uniformly over $P \in \mathcal{P}_n$" many times in the proofs of Theorem 2.1 and Corollaries 2.1 and 2.2. Throughout this appendix, we assume that $n \geq n_0$.

Proof of Theorem 2.1.

We split the proof into five steps.

Step 1. (Preliminary rate result). We claim that with probability $1 - o(1)$, $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| \check{\theta}_{uj} - \theta_{uj} \right| \lesssim B_{1n}\tau_n$. To show that, note that by definition of $\check{\theta}_{uj}$, we have for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$,

$$\left| \mathbb{E}_n \left[\psi_{uj} \left(W, \check{\theta}_{uj}, \hat{\eta}_{uj} \right) \right] \right| \leq \inf_{\theta \in \Theta_{uj}} \left| \mathbb{E}_n \left[\psi_{uj} \left(W, \theta, \hat{\eta}_{uj} \right) \right] \right| + \epsilon_n,$$

which implies via the triangle inequality that uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, with probability $1 - o(1)$,

$$\left| \mathrm{E}_P \left[\psi_{uj} \left(W, \theta, \eta_{uj} \right) \right] \Big|_{\theta = \check{\theta}_{uj}} \right| \leq \epsilon_n + 2I_1 + 2I_2 \lesssim B_{1n}\tau_n, \quad \text{(A.1)}$$

where

$$I_1 := \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}} \left| \mathbb{E}_n \left[\psi_{uj} \left(W, \theta, \hat{\eta}_{uj} \right) \right] - \mathbb{E}_n \left[\psi_{uj} \left(W, \theta, \eta_{uj} \right) \right] \right| \lesssim B_{1n}\tau_n,$$
$$I_2 := \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}} \left| \mathbb{E}_n \left[\psi_{uj} \left(W, \theta, \eta_{uj} \right) \right] - \mathrm{E}_P \left[\psi_{uj} \left(W, \theta, \theta_{uj} \right) \right] \right| \lesssim \tau_n,$$

and the bounds on I_1 and I_2 are derived in Step 2 [note also that $\epsilon_n = o(\tau_n)$ by construction of the estimator and Assumption 2.2(vi)]. Since by Assumption 2.1(iv), $2^{-1} \left| J_{uj} \left(\check{\theta}_{uj} - \theta_{uj} \right) \right| \wedge c_0$ does not exceed the left-hand side of (A.1), $\inf_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| J_{uj} \right| \gtrsim 1$, and by Assumption 2.2(vi), $B_{1n}\tau_n = o(1)$, we conclude that

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| \check{\theta}_{uj} - \theta_{uj} \right| \lesssim \left(\inf_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| J_{uj} \right| \right)^{-1} B_{1n}\tau_n \lesssim B_{1n}\tau_n, \quad \text{(A.2)}$$

with probability $1 - o(1)$ yielding the claim of this step.

Step 2. (Bounds on I_1 and I_2). We claim that with probability $1 - o(1)$, $I_1 \lesssim B_{1n}\tau_n$ and $I_2 < \tau_n$. To show these relations, observe that with probability $1 - o(1)$, we have $I_1 \leq 2I_{1a} + I_{1b}$ and $I_2 \leq I_{1a}$, where

$$I_{1a} := \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}, \eta \in \mathcal{T}_{uj}} \left| \mathbb{E}_n \left[\psi_{uj}(W, \theta, \eta) \right] - \mathrm{E}_P \left[\psi_{uj}(W, \theta, \eta) \right] \right|,$$
$$I_{1b} := \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}, \eta \in \mathcal{T}_{uj}} \left| \mathrm{E}_P \left[\psi_{uj}(W, \theta, \eta) \right] - \mathrm{E}_P \left[\psi_{uj} \left(W, \theta, \eta_{uj} \right) \right] \right|.$$

To bound I_{1b}, we employ Taylor's expansion:

$$I_{1b} \leq \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}, \eta \in \mathcal{T}_{uj}, r \in [0,1)} \partial_r \mathrm{E}_P\left[\psi_{uj}\left(W, \theta, \eta_{uj} + r\left(\eta - \eta_{uj}\right)\right)\right]$$
$$\leq B_{1n} \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \eta \in \mathcal{T}_{uj}} \left\|\eta - \eta_{uj}\right\|_e \leq B_{1n}\tau_n,$$

by Assumptions 2.1(v) and 2.2(ii).

To bound I_{1a}, we apply the maximal inequality of Lemma P.2 to the class $\mathcal{F}_1$ defined in Assumption 2.2 to conclude that with probability $1 - o(1)$,

$$I_{1a} \lesssim n^{-1/2}\left(\sqrt{v_n \log a_n} + n^{-1/2+1/q} v_n K_n \log a_n\right). \quad \text{(A.3)}$$

Here, we used: $\log \sup_Q N\left(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}\right) \leq v_n \log\left(a_n/\epsilon\right)$ for all $0 < \epsilon \leq 1$ with $\|F_1\|_{P,q} \leq K_n$ by Assumption 2.2(iv); $\sup_{f \in \mathcal{F}_1} \|f\|_{P,2}^2 \leq C_0$ by Assumption 2.2(v); $a_n \geq n \vee K_n$ and $v_n \geq 1$ by the choice of a_n and v_n. In turn, the right-hand side of (A.3) is bounded from above by $O(\tau_n)$ by Assumption 2.2(vi) since $\left(v_n \log a_n/n\right)^{1/2} \lesssim \tau_n$ and

$$n^{-1/2} n^{-1/2+1/q} v_n K_n \log a_n \lesssim n^{-1/2}\delta_n \lesssim n^{-1/2} \lesssim \tau_n.$$

Combining presented bounds gives the claim of this step.

Step 3. (Linearization). Here, we prove the claim of the theorem. Fix $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. By definition of $\check{\theta}_{uj}$, we have

$$\sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \check{\theta}_{uj}, \hat{\eta}_{uj}\right)\right]\right| \leq \inf_{\theta \in \Theta_{uj}} \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \theta, \hat{\eta}_{uj}\right)\right]\right| + \epsilon_n\sqrt{n}. \quad \text{(A.4)}$$

Also, for any $\theta \in \Theta_{uj}$ and $\eta \in \mathcal{T}_{uj}$, we have

$$\sqrt{n}\mathbb{E}_n\left[\psi_{uj}(W, \theta, \eta)\right] = \sqrt{n}\mathbb{E}_n\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right] - \mathbb{G}_n\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right) \quad \text{(A.5)}$$
$$- \sqrt{n}\left(\mathrm{E}_P\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right] - \mathrm{E}_P\left[\psi_{uj}(W, \theta, \eta)\right]\right) + \mathbb{G}_n\psi_{uj}(W, \theta, \eta).$$

Moreover, by Taylor's expansion of the function $r \mapsto \mathrm{E}_P[\psi_{uj}(W, \theta_{uj} + r(\theta - \theta_{uj}), \eta_{uj} + r(\eta - \eta_{uj}))]$,

$$\mathrm{E}_P\left[\psi_{uj}(W,\theta,\eta)\right] - \mathrm{E}_P\left[\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right] = J_{uj}\left(\theta-\theta_{uj}\right) + \mathrm{D}_{u,j,0}\left[\eta-\eta_{uj}\right] \quad \text{(A.6)}$$
$$+ 2^{-1}\partial_r^2 \mathrm{E}_P\left[W,\theta_{uj}+r\left(\theta-\theta_{uj}\right),\eta_{uj}+r\left(\eta-\eta_{uj}\right)\right]\Big|_{r=\bar{r}}$$

for some $\bar{r} \in (0,1)$. Substituting this equality into (A.5), taking $\theta = \check{\theta}_{uj}$ and $\eta = \hat{\eta}_{uj}$, and using (A.4) gives

$$\sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right] + J_{uj}\left(\check{\theta}_{uj}-\theta_{uj}\right) + \mathrm{D}_{u,j,0}\left[\hat{\eta}_{uj}-\eta_{uj}\right]\right| \le \epsilon_n\sqrt{n} \quad \text{(A.7)}$$
$$+ \inf_{\theta\in\Theta_{uj}} \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W,\theta,\hat{\eta}_{uj}\right)\right]\right| + \left|II_1(u,j)\right| + \left|II_2(u,j)\right|,$$

where

$$II_1(u,j) := \sqrt{n}\sup_{r\in[0,1)}\left|\partial_r^2 \mathrm{E}_P\left[\psi_{uj}\left(W,\theta_{uj}+r\left(\theta-\theta_{uj}\right),\eta_{uj}+r\left\{\eta-\eta_{uj}\right\}\right)\right]\right|,$$
$$II_2(u,j) := \mathbb{G}_n\left(\psi_{uj}(W,\theta,\eta) - \psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right)$$

with $\theta = \check{\theta}_{uj}$ and $\eta = \hat{\eta}_{uj}$. It will be shown in Step 4 that

$$\sup_{u\in U, j\in[\tilde{p}]}\left(\left|II_1(u,j)\right| + \left|II_2(u,j)\right|\right) = O_P(\delta_n). \quad \text{(A.8)}$$

In addition, it will be shown in Step 5 that

$$\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}\inf_{\theta\in\Theta_{uj}} \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W,\theta,\hat{\eta}_{uj}\right)\right]\right| = O_P(\delta_n). \quad \text{(A.9)}$$

Moreover, $\epsilon_n\sqrt{n} = o(\delta_n)$ by construction of the estimator. Therefore, the expression in (A.7) is $O_P(\delta_n)$. Also, $\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}\left|\mathrm{D}_{u,j,0}\left[\hat{\eta}_{uj}-\eta_{uj}\right]\right| = O_P\left(\delta_n n^{-1/2}\right)$ by the near-orthogonality condition since $\hat{\eta}_{uj} \in \mathcal{T}_{uj}$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ with probability 1 — $o(1)$ by Assumption 2.2(i). Therefore, Assumption 2.1(iv) gives

$$\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}\left|J_{uj}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_{uj}\left(W,\theta_{uj},\eta_{uj}\right)\right] + \sqrt{n}\left(\check{\theta}_{uj}-\theta_{uj}\right)\right| = O_P(\delta_n).$$

The asserted claim now follows by dividing both parts of the display above by σ_{uj} (under the supremum on the left-hand side) and noting that σ_{uj} is bounded below from zero uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ by Assumptions 2.2(iii) and 2.2(v).

Step 4. [Bounds on $II_1(u, j)$ and $II_2(u, j)$]. Here, we prove (A.8). First, with probability $1 - o(1)$,

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |II_1(u, j)| \leq \sqrt{n} B_{2n} \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \left| \check{\theta}_{uj} - \theta_{uj} \right|^2 \vee \left\| \hat{\eta}_{uj} - \eta_{uj} \right\|_e^2 \lesssim \sqrt{n} B_{1n}^2 B_{2n} \tau_n^2 \lesssim \delta_n,$$

where the first inequality follows from Assumptions 2.1(v) and 2.2(i), the second from Step 1 and Assumptions 2.2(ii) and 2.2(vi) and the third from Assumption 2.2(vi).

Second, we have with probability $1 - o(1)$ that $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |II_2(u, j)| \lesssim \sup_{f \in \mathcal{F}_2} |\mathbb{G}_n(f)|$, where

$$\mathcal{F}_2 = \left\{ \psi_{uj}(\cdot, \theta, \eta) - \psi_{uj}\left(\cdot, \theta_{uj}, \eta_{uj}\right) : u \in \mathcal{U}, j \in [\tilde{p}], \eta \in \mathcal{T}_{uj}, \left| \theta - \theta_{uj} \right| \leq C B_{1n} \tau_n \right\}$$

for sufficiently large constant C. To bound $\sup_{f \in \mathcal{F}_2} |\mathbb{G}_n(f)|$, we applyLemmaP.2. Observe that

$$\begin{aligned} &\sup_{f \in \mathcal{F}_2} \|f\|_{P,2}^2 \\ &\leq \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \left| \theta - C B_{1n} \tau_n, \eta \in \mathcal{T}_{uj} \right.} E_P\left[\left(\psi_{uj}(W, \theta, \eta) - \psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right) \right)^2 \right] \\ &\leq \sup_{u \in \mathcal{U}, j \in [\tilde{p}], \left| \theta - C B_{1n} \tau_n, \eta \in \mathcal{T}_{uj} \right.} C_0 \left(\left| \theta - \theta_{uj} \right| \vee \left\| \eta - \eta_{uj} \right\|_e \right)^{\omega} \lesssim \left(B_{1n} \tau_n \right)^{\omega}, \end{aligned}$$

where we used Assumption 2.1(v) and Assumption 2.2(ii). Also, observe that $(B_{1n}\, \tau_n)^{\omega/2} \geq n^{-\omega/4}$ by Assumption 2.2(vi) since $B_{1n} \geq 1$. Therefore, an application of Lemma P.2 with an envelope $F_2 = 2F_1$ and $\sigma = (CB_{1n}\, \tau_n)^{\omega/2}$ for sufficiently large constant C gives with probability $1 - o(1)$,

$$\sup_{f \in \mathcal{F}_2} |\mathbb{G}_n(f)| \lesssim \left(B_{1n} \tau_n \right)^{\omega/2} \sqrt{v_n \log a_n} + n^{-1/2 + 1/q} v_n K_n \log a_n, \quad \text{(A.10)}$$

since $\sup_{f \in \mathcal{F}_2} |f| \leq 2 \sup_{f \in \mathcal{F}_1} |f| \leq 2F_1$ and $\|F_1\|_{P,q} \leq K_n$ by Assumption 2.2(iv) and

$$\log \sup_Q N\left(\epsilon \|F_2\|_{Q,2}, \mathcal{F}_2, \|\cdot\|_{Q,2} \right) \lesssim v_n \log\left(a_n / \epsilon \right) \quad \text{for all } 0 < \epsilon \leq 1$$

by Lemma O.1 because $\mathcal{F}_2 \subset \mathcal{F}_1 - \mathcal{F}_1$ for $\mathcal{F}_1$ defined in Assumption 2.2(iv). The claim of this step now follows from an application of Assumption 2.2(vi) to bound the right-hand side of (A.10).

Step 5. Here, we prove (A.9). For all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, let $\bar{\theta}_{uj} = \theta_{uj} - J_{uj}^{-1}\mathbb{E}_n\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right]$. Then $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left|\bar{\theta}_{uj} - \theta_{uj}\right| = O_P\left(\mathcal{S}_n/\sqrt{n}\right)$ since $\mathcal{S}_n = \mathrm{E}_P\left[\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left|\sqrt{n}\mathbb{E}_n\left[\psi_{uj}\left(W_{uj}, \theta_{uj}, \eta_{uj}\right)\right]\right|\right]$ and J_{uj} is bounded in absolute value below from zero uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ by Assumption 2.1(iv). Therefore, $\bar{\theta}_{uj} \in \Theta_{uj}$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ with probability $1 - o(1)$ by Assumption 2.1(i). Hence, with the same probability, for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$,

$$\inf_{\theta \in \Theta_{uj}} \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \theta, \hat{\eta}_{uj}\right)\right]\right| \leq \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \bar{\theta}_{uj}, \hat{\eta}_{uj}\right)\right]\right|,$$

and so it suffices to show that

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \bar{\theta}_{uj}, \hat{\eta}_{uj}\right)\right]\right| = O_P(\delta_n). \quad \text{(A.11)}$$

To prove (A.11), for given $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, substitute $\theta = \bar{\theta}_{uj}$ and $\eta = \hat{\eta}_{uj}$ into (A.5) and use Taylor's expansion in (A.6). This gives

$$\sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \bar{\theta}_{uj}, \hat{\eta}_{uj}\right)\right]\right| \leq \left|\widetilde{II}_1(u, j)\right| + \left|\widetilde{II}_2(u, j)\right| + \sqrt{n}\left|\mathbb{E}_n\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right] + J_{uj}\left(\bar{\theta}_{uj} - \theta_{uj}\right)\right.$$
$$\left. + \mathrm{D}_{u,j,0}\left[\hat{\eta}_{uj} - \eta_{uj}\right]\right|,$$

where $\widetilde{II}_1(u, j)$ and $\widetilde{II}_2(u, j)$ are defined as $II_1(u, j)$ and $II_2(u, j)$ in Step 3 but with $\breve{\theta}_{uj}$ replaced by $\bar{\theta}_{uj}$. Then, given that $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left|\bar{\theta}_{uj} - \theta_{uj}\right| \lesssim \mathcal{S}_n \log n/\sqrt{n}$ with probability $1 - o(1)$, the argument in Step 4 shows that

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left(\left|\widetilde{II}_1(u, j)\right| + \left|\widetilde{II}_2(u, j)\right|\right) = O_P\left(\delta_n\right).$$

In addition, $\mathbb{E}_n\left[\psi_{uj}\left(W, \theta_{uj}, \eta_{uj}\right)\right] + J_{uj}\left(\bar{\theta}_{uj} - \theta_{uj}\right) = 0$ by the definition of $\bar{\theta}_{uj}$ and $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]}\left|\mathrm{D}_{u,j,0}\left[\hat{\eta}_{uj} - \eta_{uj}\right]\right| = O_P\left(\delta_n n^{-1/2}\right)$ by the near-orthogonality condition. Combining these bounds gives (A.11), so that the claim of this step follows, and completes the proof of the theorem.

APPENDIX B:: REMAINING PROOFS FOR SECTION 2

See the Supplementary Material.

APPENDIX C:: PROOFS FOR SECTIONS 3 AND 4

See the Supplementary Material.

REFERENCES

[1]. Andrews DWK (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica 62 43–72.

[2]. 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 2369–2429.

[3]. Belloni A and Chernozhukov V (2011). ℓ_1-Penalized quantile regression for high dimensional sparse models. Ann. Statist 39 82–130.

[4]. Belloni A and Chernozhukov V (2013). Least squares after model selection in high-dimensional sparse models. Bernoulli 19 521–547. Available at arXiv:1001.0188.

[5]. Belloni A, Chernozhukov V, Chetverikov D and Wei Y (2018). Supplementto "Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework." DOI:10.1214/17-AOS1671SUPP.

[6]. Belloni A, Chernozhukov V, Fernández-Val I and Hansen C (2013). Program evaluation with high-dimensional data. Available at arXiv:1311.2645.

[7]. Belloni A, Chernozhukov V and Hansen C (2010). Lasso methods for Gaussian instrumental variables models. Available at arXiv:1012.1297.

[8]. Belloni A, Chernozhukov V and Hansen C (2013). Inference for high-dimensional sparse econometric models. In Advances in Economics and Econometrics. 10th World Congress of Econometric Society, August 2010, Vol. III. 245–295. Available at arXiv:1201.0220.

[9]. Belloni A, Chernozhukov V and Hansen C (2014). Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud 81 608–650.

[10]. Belloni A, Chernozhukov V and Kato K (2013). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. Available at arXiv:1312.7186.

[11]. Belloni A, Chernozhukov V and Kato K (2015). Uniform post selection inference for LAD regression models and other Z-estimators. Biometrika 102 77–94.

[12]. Belloni A, Chernozhukov V and Wang L (2011). Square-root-lasso: Pivotal recovery of sparse signals via conic programming. Biometrika 98 791–806.

[13]. Belloni A, Chernozhukov V and Wang L (2014). Pivotal estimation via square-root Lasso in nonparametric regression. Ann. Statist 42 757–788.

[14]. Chamberlain G (1992). Efficiency bounds for semiparametric regression. Econometrica 60 567–596.

[15]. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W and Robins J (2018). Double/debiased machine learning for treatment and structural parameters. Econom. J 21 C1–C68.

[16]. Chernozhukov V, Chetverikov D and Kato K (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Statist 41 2786–2819.

[17]. Chernozhukov V, Chetverikov D and Kato K (2014). Anti-concentration and honest, adaptive confidence bands. Ann. Statist 42 1787–1818.

[18]. Chernozhukov V, Chetverikov D and Kato K (2017). Central limit theorems and bootstrap in high dimensions. Ann. Probab 4 2309–2352.

[19]. Chernozhukov V, Chetverikov D and Kato K (2014). Gaussian approximation of suprema of empirical processes. Ann. Statist 42 1564–1597.

[20]. Chernozhukov V, Chetverikov D and Kato K (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. Probab. Theory Related Fields 162 47–70.

[21]. Chernozhukov V, Chetverikov D and Kato K (2015). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. Available at arXiv:1502.00352.

[22]. Chernozhukov V, Fernández-Val I and Melly B (2013). Inference on counterfactual distributions. Econometrica 81 2205–2268.

[23]. Chernozhukov V, Hansen C and Spindler M (2015). Post-selection and post regularization inference in linear models with very many controls and instruments. Am. Econ. Rev. Pap. Proc 105 486–490.

[24]. Deng H and ZHANG C-H (2017). Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. Available at arXiv:1705.09528.

[25]. Dudley R (1999). Uniform Central Limit Theorems Cambridge Studies in Advanced Mathematics 63. Cambridge Univ. Press, Cambridge.

[26]. Hothorn T, Kneib T and Bühlmann P (2014). Conditional transformation models.J. Roy. Statist. Soc. Ser. B 76 3–27.

[27]. Javanmard A and Montanari A (2014). Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res 15 2869–2909.

[28]. Javanmard A and Montanari A (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. IEEE Trans. Inform. Theory 60 6522–6554.

[29]. Kosorok M (2008). Introduction to Empirical Processes and Semiparametric Inference. Springer, Berlin.

[30]. Leeb H and Pötscher B (2008). Can one estimate the unconditional distribution of post-model-selection estimators? Econometric Theory 24 338–376.

[31]. Leeb H and Pötscher B (2008). Recent developments in model selection and related areas. Econometric Theory 24 319–322.

[32]. Leeb H and Pötscher BM (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. J. Econometrics 142 201–211.

[33]. Linton O (1996). Edgeworth approximation for MINPIN estimators in semiparametric regression models. Econometric Theory 12 30–60. Cowles Foundation Discussion Papers 1086 (1994).

[34]. Mammen E (1993). Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist 21 255–285.

[35]. Newey W (1990). Semiparametric efficiency bounds. J. Appl. Econometrics 5 99–135.

[36]. Newey W (1994). The asymptotic variance of semiparametric estimators. Econometrica 62 1349–1382.

[37]. Neyman J (1959). Optimal asymptotic tests of composite statistical hypotheses In Probability and Statistics: The Harald Cramér Volume (Grenander U, ed.) 213–234. Almqvist & Wiksell, Stockholm.

[38]. Neyman J (1979). $c(\alpha)$ tests and their use. Sankhyā 41 1–21.

[39]. Ning Y and Liu H (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Available at arXiv:1412.8765.

[40]. Pötscher B and Leeb H (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. J. Multivariate Anal 100 2065–2082.

[41]. Robins J and Rotnitzky A (1995). Semiparametric efficiency in multivariate regression models with missing data. J. Amer. Statist. Assoc 90 122–129.

[42]. Stein C (1956). Efficient nonparametric testing and estimation In Proc. 3rd Berkeley Symp. Math. Statist. and Probab. 1 187–195. Univ. California Press, Berkeley, CA.

[43]. Van de Geer S, Bühlmann P, Ritov Y and Dezeure R (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist 42 1166–1202.

[44]. Van der Vaart A (1998). Asymptotic Statistics. Cambridge Univ. Press, Cambridge.

[45]. Van der Vaart A and Wellner J (1996). Weak Convergence and Empirical Processes.

[46]. Zhang C-H and Zhang S (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. J. Roy. Statist. Soc. Ser. B 76 217–242.

[47]. Zhao T, Kolar M and Liu H (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. Available at arXiv:1412.8724.

Table 1

We report the pointwise rejection frequencies of each method for (pointwise) confidence intervals for each $j \in \{1,\ldots, 5\}$. For Design 1, we used $\mathscr{U} = \{1\}$ and for Design 2, we used $\mathscr{U} = \{0.5\}$. The results are based on 500 replications

$p = 2000, n = 500$		Rejection frequencies for $j \in \{1,\ldots, 5\}$				
Design	**Method**	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
1(i)	Proposed	0.042	0.040	0.062	0.050	0.044
	Naive	0.100	0.098	0.108	0.108	0.162
1(ii)	Proposed	0.044	0.040	0.054	0.056	0.056
	Naive	0.038	0.030	0.070	0.886	0.698
2(i)	Proposed	0.046	0.054	0.044	0.052	0.054
	Naive	0.046	0.050	0.038	0.070	0.054
2(ii)	Proposed	0.092	0.074	0.034	0.088	0.082
	Naive	0.034	0.972	0.182	0.312	0.916

Table 2

We report the rejection frequencies of each method for the (uniform) confidence bands for $\mathcal{U} \times \tilde{p}$. The proposed estimator computes the critical value based on the multiplier bootstrap procedure. For the naive post-selection estimator, we report the results for two choices of critical values, one choice based on the multiplier bootstrap (MB) and another based on Bonferroni (BF) correction. The results are based on 500 replications

$p = 2000, n = 500$		Uniform over $\mathcal{U} \times \tilde{p}$			
Design	**Method**	**[1,2.5] × {1}**	**{1} × [10]**	**[1,2.5] × [10]**	**{1} × [1000]**
1(i)	Proposed	0.054	0.036	0.048	0.040
	Naive (MB)	0.126	0.136	0.172	0.032
	Naive (BF)	0.014	0.124	0.026	0.032
1(ii)	Propose	0.270	0.036	0.032	0.142
	Naive (MB)	0.014	0.802	0.934	0.404
	Naive (BF)	0.000	0.802	0.718	0.376
Design	**Method**	**[−0.5, 0.5] × {1}**	**{0.5} × [10]**	**[−0.5,0.5] × [10]**	**{0.5} × [1000]**
2(i)	Proposed	0.364	0.038	0.052	0.062
	Naive (MB)	0.116	0.040	0.022	0.048
	Naive (BF)	0.018	0.038	0.000	0.046
2(ii)	Proposed	0.140	0.090	0.408	0.084
	Naive (MB)	0.002	0.946	0.996	0.362
	Naive (BF)	0.000	0.946	0.944	0.298