BAYESIAN VARIABLE SELECTION IN LINEAR REGRESSION MODELS WITH INSTRUMENTAL VARIABLES

GAUTAM SABNIS, YVES ATCHADÉ, AND PROSPER DOVONON

Abstract. This paper introduces semi-parametric Bayesian methods for high-dimensional linear instrumental variables (IV) models. A selection method for relevant regressors in presence of endogeneity is proposed and its validity established within a quasi-Bayesian framework. We study the properties of the quasi-posterior distribution as the number of regressors increases and provide a set of assumptions under which this posterior distribution puts asymptotically most of its probability mass around the true value of the parameter. An easy-to-implement and efficient Markov Chain Monte Carlo algorithm is proposed to sample from the quasi-posterior distribution. The finite sample performance of the proposed methods are investigated by Monte Carlo experiments. We also carry out an empirical investigation that estimates the return on education using US census data.

1. Introduction

Endogeneity issues arise in a linear regression model when a subset of regressors are correlated with the regression error. Endogenous variables are those influenced by some of the same forces that influence the response variable. For example, economists examining the effects of education on earnings have long been concerned about the endogeneity of education (Angrist and Krueger, 1991). "Ability" is often cited as one factor possibly correlated with earnings (those with higher ability earn more) and education (those with higher ability obtain more education). Endogeneity also arises when measurement errors are present in explanatory variables. It is well-known in

2010 *Mathematics Subject Classification.* 62F15, 62Jxx.

Key words and phrases. High-dimensional Bayesian inference, Endogeneity, Variable selection, Posterior contraction, Markov Chain Monte Carlo, linear regression.

This work is partially supported by the NSF grant DMS 1513040.

G. Sabnis: Boston University, 111 Cummington Mall, Boston, 02215, MA, United States. *E-mail address:* gsabnis@bu.edu.

A. Y. Atchadé: Boston University, 111 Cummington Mall, Boston, 02215, MA, United States. *E-mail address:* atchade@bu.edu.

P. Dovonon: Concordia University, 1455 de Maisonneuve Blvd. West Montreal, Quebec, H3G 1M8, Canada. *E-mail address:* prosper.dovonon@concordia.edu.

regression analysis that endogeneity causes standard estimators such as the ordinary least squares estimator to be inconsistent.

The most common cure to endogeneity issues consists in resorting to instrumental variable (IV) inference (Stock and Trebbi, 2003; Imbens, 2014). Consistent estimation is obtained by relying on the so-called valid instrumental variables; i.e. variables uncorrelated with the regression error but correlated with the endogenous regressors. The absence of correlation between error and IVs yields the IV model:

$$\mathbb{E}[w_{ik}(y_i - x_i'\theta)] = 0, \quad k = 1, \ldots, q, \quad i = 1, \ldots, n$$

where $y_i \in \mathbb{R}$ is the response variable, $x_i \in \mathbb{R}^p$ is the vector of explanatory variables, $w_i \in \mathbb{R}^q$ the vector of instruments, $\theta \in \mathbb{R}^p$ the vector of parameters, and n is the sample size.

High-dimensional regression models (Fan and Li (2001); Candes et al. (2007); Bühlmann and van de Geer (2011); Hastie et al. (2015)) where p can be much larger than the sample size n are not immune to the endogeneity problems outlined above. In fact, as mentioned by Fan and Liao (2014), endogeneity can arise incidentally from the pooling of a large number of regressors. Importantly, Fan and Liao (2014) show that even if they are not ultimately relevant, the presence of endogenous variables may cause penalized least square methods to be inconsistent, as in standard linear regression models.

The objective of this paper is to perform variable selection and estimation of regression parameters in high-dimensional linear regression models in presence of endogenous variables. As a particular case of moment condition model, the IV model is commonly estimated by the generalized method of moments (GMM) introduced by Hansen (1982). However when $q \geq n$, the GMM objective function is typically too noisy to be directly useful. This has led Fan and Liao (2014) to propose the focused GMM (FGMM) which is obtained by minimizing a GMM criterion that sets focus solely on instruments associated to included regressors. Hence an important assumption in FGMM is that the endogeneous variables, as well as the relevant instruments to correct for them are known.

Other recent related work include Belloni et al. (2012); Gautier and Tsybakov (2014); Belloni et al. (2017). Belloni et al. (2012) propose a two-step lasso/post-lasso approach for instrument selection and inference in linear IV models where the number of explanatory variables (p) is fixed but the number of instrumental variables (q) is large. Gautier et al. (2011) consider p large and q possibly large and propose the so-called self-tuning IV estimator and non-asymptotic confidence intervals based on the Dantzig selection of Candes et al. (2007). Belloni et al. (2017) consider p and q large and propose estimators and confidence regions that are honest and asymptotically

correct by relying on a two-step procedure that builds suitably orthogonalized instruments. Unlike FGMM, the methods by Gautier et al. (2011); Belloni et al. (2017) include an automatic moments selection step, and can thus be applied more widely, even when the endogenous variables are not known. However an important drawback with these methods is a potential lack of interpretability of the selected instruments.

This objective of this paper is to develop a Bayesian instrumental variables model using the framework of focused GMM. There are several advantages of taking a Bayesian approach. Firstly, the Bayesian approach makes it straightforward to incorporate into the inference procedure any existing prior information on the relevance of the regressors. This is currently an important issue in many applied research (Greenfield et al., 2013; Studham et al., 2014; Peng et al., 2013). There is also a computational advantage. Indeed, when dealing with discrete parameters or highly multimodal objective functions, sampling actually tends to scale better than optimization. (Chernozhukov and Hong, 2003) made a similar case. Some rigorous results on this phenomenon can be found in (Ma et al., 2018).

By only restricting the moments of the data, IV models obviate the need to assume an underlying data distribution (or complete specification of a likelihood function), and allow inferences about the parameter of interest based only on the partial information supplied by a set of moment conditions. We carry out the same semi-parametric inference in the Bayesian framework, by employing a quasi-Bayesian approach (Chernozhukov and Hong, 2003; Liao et al., 2011; Kato et al., 2013; Atchade, 2017).

The main contributions of this paper are threefold. First, using a working quasi-likelihood combined with a spike-and-slab prior distribution (Mitchell and Beauchamp, 1988; George and McCulloch, 1997), we develop a Bayesian semi-parametric method for variable selection in high-dimensional linear models with endogenous regressors. Second, we study the statistical properties of the quasi-posterior distribution, Π (defined below in (4)), as the dimension p increases. Under some minimal assumptions, we show that Π puts most of its probability mass around the true value of the parameter as $p \to \infty$ (see Theorem 3). Third, we develop an easy-to-implement and efficient Markov Chain Monte Carlo algorithm to sample from Π. To the best of our knowledge, ours is the first paper to present a Bayesian approach tackling endogeneity issues in high-dimensional linear IV models. This work builds on the general approach to high-dimensional Bayesian inference developed in (Atchade and Bhattacharyya (2018)). However due to the specific form of the IV model, results developed in Atchade and Bhattacharyya (2018)) cannot be directly applied. The performance of the methods is highlighted by Monte Carlo simulations. The paper

also includes an empirical application that assesses the return on education using US data by revisiting the work of Angrist and Krueger (1991).

The rest of the paper is organized as follows. The model and the Bayesian method proposed are presented in Section 2. This section also presents our main results establishing the consistency of the selection method proposed. The MCMC sampling algorithm is introduced in Section 3 which also contains our simulation results. Section 4 contains the empirical application and concluding remarks are included in Section 5.

1.1. **Notation.** For integer $a > 0$, we equip $\mathbb{R}^a$, the space of a-dimensional vectors with its usual Euclidean inner product $\langle \cdot, \cdot \rangle$, associated norm $\|\cdot\|_2$, and its Borel sigma-algebra. Unless stated otherwise all vectors are column vectors. We will also use the following norms on $\mathbb{R}^a$: $\|\theta\|_1 \stackrel{\text{def}}{=} \sum_{j=1}^{a} |\theta_j|$, $\|\theta\|_0 \stackrel{\text{def}}{=} \sum_{j=1}^{a} \mathbf{1}_{\{}|\theta_j| > 0\}$ and $\|\theta\|_\infty \stackrel{\text{def}}{=} \max_{1 \le j \le a} |\theta_j|$.

We set $\Delta \stackrel{\text{def}}{=} \{0,1\}^p$. For $\delta \in \Delta$, we set $\delta^c \stackrel{\text{def}}{=} 1 - \delta$, that is $\delta_j^c \stackrel{\text{def}}{=} 1 - \delta_j, 1 \le j \le p$. For $\theta \in \mathbb{R}^p$, the sparsity structure (or support) of θ is the element $\delta \in \Delta$ defined as $\delta_j = \mathbf{1}_{\{|\theta_j|>0\}}, 1 \le j \le p$. Given $\theta \in \mathbb{R}^p$, and $\delta \in \Delta$, the notation $\theta \cdot \delta$ denotes their component-wise product: $(\theta \cdot \delta)_j = \theta_j \delta_j$, $1 \le j \le p$. We then set $\mathbb{R}^p_\delta \stackrel{\text{def}}{=} \{\theta \cdot \delta : \theta \in \mathbb{R}^p\}$. At times we will write θ_δ as a short for $\theta \cdot \delta$. It may help to think of $\delta \in \Delta$ as a selection of regressors, or as a model.

For a given matrix A, we will write A_j to denote its j-th column, and A_δ to denote the sub-matrix of A obtained by selecting the columns j of A for which $\delta_j = 1$.

Throughout the paper e denotes the Euler number and $[p]$ represents the sequence $1, \ldots, p$.

2. Model and Main results

Suppose that we have n independent subjects, and observe on subject i the random vector $(y_i, x_i, w_i) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$. More precisely we make the following data-generating assumption.

H1 **H1.** *$\{(y_i, x_i, w_i, \epsilon_i),\ 1 \le i \le n\}$ are n independent and identically distributed random vectors, where $(y_i, x_i, w_i, \epsilon_i) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}$ with $q \ge p$, and there exists $\theta_\star \in \mathbb{R}^p$ such that*

$$y_i = \langle x_i, \theta_\star \rangle + \epsilon_i, \ \textit{for all} \quad i = 1, \ldots, n. \tag{1}$$

reg:mod

Furthermore we assume that $\epsilon \stackrel{\text{def}}{=} (\epsilon_1, \ldots, \epsilon_n)'$ is conditionally sub-Gaussian in the sense that there exists $\sigma_0 > 0$ such that for all $u \in \mathbb{R}^n$,

$$\mathbb{E}(\epsilon | W) = 0, \quad \textit{and} \quad \mathbb{E}\left(e^{\langle u, \epsilon \rangle} | W\right) \le e^{\frac{\sigma_0^2 \|u\|_2^2}{2}}, \tag{2}$$

sub:gaussian:eq

almost surely, where $W \in \mathbb{R}^{n\times q}$ is the matrix with i-th rows given by w_i'.

Throughout we set $y \stackrel{\text{def}}{=} (y_1, \ldots, y_n)' \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ with i-th row x_i', and we set $\mathbf{z} \stackrel{\text{def}}{=} (y, X, W) \in \mathbb{R}^n \times \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times q}$.

We consider the situation where some of the components of the regressor x_i are endogenous, in the sense that there are correlated with the error ϵ_i, so that $\mathbb{E}(\epsilon_i x_i) \neq 0$. As documented in the introduction, this issue is very common in applications, and it is well-known that inferential procedures that ignore endogeneity are inconsistent in general. A well-established approach to mitigate endogeneity is the use of instrumental variables. This is the approach taken here, and the set of instruments at our disposal is $w_i \in \mathbb{R}^q$. Inference with instrumental variables is classically done via the GMM estimator that is obtained by minimizing the GMM functional

$$\theta \mapsto (y - X\theta)' W D W' (y - X\theta),$$

or penalized versions thereof, where $D \in \mathbb{R}^{q\times q}$ is a symmetric positive definite weight matrix. However in a context where q is potentially larger than n, the GMM functional is typically too noisy. To circumvent this problem we adopted the focused GMM approach of Fan and Liao (2014) that incorporates a moment selection step: only instruments associated to selected regression parameters are included in the model. Note here that the idea of moment selection differs from previous works on moments selection (as in for instance Caner et al. (2018)) which deal with the question of how to retain only valid moment conditions. In our case, all the moments conditions are assumed valid, but we face the challenge of having too many of them, given the available sample size. The purpose of this work is to develop a Bayesian version of focused GMM.

For broader applicability we extended the focused GMM framework of Fan and Liao (2014), to allow for any set of instruments that satisfies H1. However as in Fan and Liao (2014), we shall make the crucial assumption that each explanatory variable X_j has a known small set of instruments. If X_j is exogenous, then it may be taken as its own instrument, although other choices are allowed. More generally, given $\delta \in \Delta$, we let $T(\delta) \in \{0,1\}^q$ be such that $(T(\delta))_k = 1$ if instrument k is needed for some variable in X_δ, and $(T(\delta))_k = 0$ otherwise. We will assume that for a given model δ there is always as many instruments as explanatory variable:

$$\|T(\delta)\|_0 \geq \|\delta\|_0, \quad \delta \in \Delta. \tag{3}$$ cond:Td

Without further notice we will assume below that (3) holds. We stress again that the mapping $\delta \mapsto T(\delta)$ is assumed known. Hence our set up applies mainly to problems where there is a small number of known endogenous variables (among a large number

p of variables), each of which with a known set of instruments. This assumption is consistent with classical instrumental variable analysis which typically presupposes that the data analyst has a good understanding of the data-generating process, and has relevant instruments to deal with endogenous variables.

Since $T_\delta \in \{0,1\}^q$, we write $W_{T(\delta)}$ to denote the submatrix of W obtained by keeping only the columns of W for which the corresponding components of T_δ are 1. Let $\mathcal{Z} \stackrel{\text{def}}{=} \mathbb{R}^n \times \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times q}$. For $\delta \in \Delta$, $\mathbf{z} = (y, X, W) \in \mathcal{Z}$, and $\sigma > 0$, we consider the following negative focused GMM functional

$$\ell(\delta, \theta; z) \stackrel{\text{def}}{=} -\frac{1}{2n\sigma^2} \left(y - X(\theta \cdot \delta)\right)' W_{T(\delta)} W'_{T(\delta)} \left(y - X(\theta \cdot \delta)\right),$$

that we view as a working log-likelihood (a quasi-log-likelihood). To carry on our Bayesian analysis we need a prior distribution on (δ, θ). Our prior information on the problem is that $\theta_\star$ is sparse. We use this to formulate the following prior distribution.

H2 **H 2.** *We assume that the prior distribution for δ on Δ is*

$$\pi(\delta) = \alpha^{\|\delta\|_0}(1-\alpha)^{p-\|\delta\|_0}, \quad \delta \in \Delta$$

where $\alpha \in (0,1)$ is such that $\alpha/(1-\alpha) = \frac{1}{p^{u+1}}$, for some constant $u > 0$. Furthermore, given δ the components of θ are independent and

$$\theta_j|\delta \sim \begin{cases} \boldsymbol{N}\left(0, \frac{1}{\rho}\right), & \text{if} \quad \delta_j = 1 \\ \boldsymbol{N}(0, \gamma), & \text{if} \quad \delta_j = 0 \end{cases}$$

for constants $\rho > 0, \gamma > 0$.

Remark 1. The prior distribution in H2 is fairly common in high-dimensional Bayesian statistics (see for instance Castillo et al. (2015); Atchade and Bhattacharyya (2018) and the references therein). It can be easily modified to account for prior information on the importance of the regressors. This is done by turning ρ into a vector $(\rho_1, \ldots, \rho_p)$, and setting ρ_j appropriately to reflect the a prior information available on the relevance of X_j. □

Let $C_\delta \in \mathbb{R}^{p\times p}$ be the diagonal matrix such that $C_{\delta,jj} = \frac{1}{\rho}$ if $\delta_j = 1$, and $C_{\delta,jj} = \gamma$ if $\delta_{jj} = 0$. Our proposed quasi-posterior distribution for inference on (δ, θ) is given by

$$\Pi(\delta, \mathrm{d}\theta|z) \propto \pi(\delta) e^{-\frac{1}{2n\sigma^2}(y-X(\theta\cdot\delta))' W_{T(\delta)} W'_{T(\delta)} (y-X(\theta\cdot\delta))} \frac{e^{-\frac{1}{2}\theta' C_\delta^{-1}\theta}}{\sqrt{\det(2\pi C_\delta)}} \mathrm{d}\theta, \tag{4}$$

post:dist

for hyper-parameters $(u, \sigma^2, \rho, \gamma)$, supplied by the user. We provide some guidelines below on choosing these constants. Inference on the support of $\theta_\star$ is done from the marginal posterior distribution $\Pi(\delta|\mathbf{z})$, and inference on the magnitude of $\theta_{\star,j}$ given

that it is significant is carried from the conditional distribution of θ_j given $\mathbf{z}$, and $\delta_j = 1$. We show below that when H1-H2 holds, this Bayesian approach to inferring $\theta_\star$ can be consistent even when p, q are much large than the sample size n.

If A is a matrix, we write $\bar{\Sigma}(A)$ (resp. $\underline{\Sigma}(A)$) to denote the largest (resp. smallest) singular value of A. Given finite constants $c_0, \bar{\kappa}$, we introduce the event

$$\mathcal{E}_0 = \mathcal{E}_0(c_0, \bar{\kappa}) \stackrel{\text{def}}{=} \Big\{ (y, X, W) \in \mathcal{Z} : \max_{1 \le k \le q} \|W_k\|_2 \le c_0 \sqrt{n}, \;\; \bar{\Sigma}\left(W'_{T(\delta_\star)} X_{\delta_\star} \right) \le n\sqrt{\bar{\kappa}}, \\ \text{and} \quad \max_{1 \le k \le q} |\langle W_k, \epsilon \rangle| \le 2c_0 \sigma_0 \sqrt{n \log(q)}, \;\; \Big\}, \quad (5)$$ def:e0

where σ_0 is as in H1. We will see below in Section 2.1 that when the data $\mathbf{z}$ is generated as described in H1, then with high probability $\mathbf{z} \in \mathcal{E}_0$ for appropriate choice of c_0 and $\bar{\kappa}$. The next result shows that in this case, the quasi-posterior distribution $\Pi(\cdot|\mathbf{z})$ puts most of its probability mass on sparse elements of Δ.

lem:sparsity **Proposition 2.** *Assume H1-H2, and let $\mathcal{E}_0$ be as defined in (5) for some constants c_0 and $\bar{\kappa}$. Suppose that $n\bar{\kappa} \le \sigma^2 \rho p$, and set*

$$\bar{k} \stackrel{\text{def}}{=} \frac{1}{u} \left[s_\star + m_0 + \frac{2c_0^2 \left(\frac{\sigma_0}{\sigma}\right)^2 \|T(\delta_\star)\|_0 \log(q)}{\log(p)} + \frac{\rho \|\theta_\star\|_2^2}{2 \log(p)} \right].$$

Then

$$\boldsymbol{1}_{\mathcal{E}_0}(\mathbf{z}) \Pi\left(\|\delta\|_0 > s_\star + \bar{k} | \mathbf{z} \right) \le \frac{2}{p^{m_0}}.$$

Proof. See Section 5.1. □

We set

$$\bar{s} \stackrel{\text{def}}{=} s_\star + \frac{1}{u} \left[s_\star + m_0 + \frac{2c_0^2 \left(\frac{\sigma_0}{\sigma}\right)^2 \|T(\delta_\star)\|_0 \log(q)}{\log(p)} + \frac{\rho \|\theta_\star\|_2^2}{2 \log(p)} \right]. \quad (6)$$ def:bars

Hence by Proposition 2, $\mathbf{1}_{\mathcal{E}_0}(\mathbf{z}) \Pi\left(\|\delta\|_0 > \bar{s} | \mathbf{z} \right) \le \frac{2}{p^{m_0}}$. Note that if the number of instruments per variable is small, $\|\theta_\star\|_\infty = O(1)$ as p grows, and ρ is taken small, then $\bar{s} = O(s_\star)$ as $p \to \infty$.

We show next that $\Pi(\cdot|z)$ put most of its probability mass around $(\delta_\star, \theta_\star)$, where $\delta_\star$ denotes the sparsity structure (or support) of $\theta_\star$, that is $\delta_{\star j} = \mathbf{1}(|\theta_{\star j}| > 0)$. To that end, we note that when $\mathbf{z} \in \mathcal{E}_0$, the maximum number of instruments used in any given model cannot exceed

$$\bar{t} \stackrel{\text{def}}{=} \max_{\delta \in \Delta_{\bar{s}}} \|T_\delta\|_0. \quad (7)$$ def:bart

Given finite constant $\underline{\kappa} > 0$, we set

$$\mathcal{E} = \mathcal{E}(c_0, \bar{\kappa}, \underline{\kappa}) \stackrel{\text{def}}{=} \mathcal{E}_0 \cap \Big\{ (y, X, W) \in \mathcal{Z} : \max_{\delta_2:\, \|\delta_2\|_0 \le \bar{t}} \max_{\delta_1:\, \|\delta_1\|_0 \le \bar{s}+s_\star} \bar{\Sigma}\left(W_{\delta_2}' X_{\delta_1}\right) \le n\sqrt{\bar{\kappa}}, \\ \text{and} \quad \min_{\delta_2:\, 0<\|\delta_2\|_0 \le \bar{t}} \min_{\delta_1:\, 0<\|\delta_1\|_0 \le \bar{s}+s_\star} \underline{\Sigma}\left(W_{\delta_2}' X_{\delta_1}\right) \ge n\sqrt{\underline{\kappa}} \Big\}. \quad (8) \quad \boxed{\text{def:e}}$$

thm:1 **Theorem 3.** *Assume H1-H2, , and let $\mathcal{E}$ be as defined in (8) for some constants $c_0, \bar{\kappa}$, and $\underline{\kappa}$. Choose $\rho > 0$ such that $n\bar{\kappa} \le \sigma^2 \rho p$, and*

$$\rho \|\theta_\star\|_2 \le \frac{c_0}{16} \left(\frac{\sigma_0}{\sigma^2}\right) \sqrt{\bar{\kappa} \bar{t} n \log(q)}. \quad (9) \quad \boxed{\text{cond:rho}}$$

Set

$$r \stackrel{\text{def}}{=} \frac{8 c_0 \sigma_0 \bar{\kappa}^{1/2}}{\underline{\kappa}} \sqrt{\frac{\bar{t} \log(q)}{n}}, \quad (10) \quad \boxed{\text{def:eps}}$$

and for absolute constants $m > 1$, $M > 2$, set

$$\mathsf{B} \stackrel{\text{def}}{=} \bigcup_{\delta \in \Delta_{\bar{s}}} \{\delta\} \times \{\theta \in \mathbb{R}^p : \ \|(\theta \cdot \delta) - \theta_\star\|_2 \le M r, \ \|\theta - (\theta \cdot \delta)\|_2 \le m\sqrt{\gamma p}\}.$$

If $c_0^2 M^2 \min(1, (\sigma_0/\sigma)^2) \ge 4(2+u)$, then

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z}) \left(1 - \Pi(\mathsf{B}|\mathbf{z})\right)\right] \le \frac{2}{p^{m_0}} + \frac{8}{q^{\bar{t}}} + 2e^{-\frac{(m-1)^2 p}{2}}. \quad (11) \quad \boxed{\text{cv:rate}}$$

Proof. See Section 5.2.1. □

Theorem 3 shows that when ρ is well chosen and $\mathbf{z} \in \mathcal{E}$, if $(\delta, \theta) \in \Pi(\cdot \mid \mathbf{z})$ then the sparse vector $(\theta \cdot \delta)$ satisfies $\|(\theta \cdot \delta) - \theta_\star\| \le M\mathsf{r}$, and $\|\theta - (\theta \cdot \delta)\|_2 \le m\sqrt{\gamma p}$ with high probability. The contraction rate is

$$\mathsf{r} = \frac{8 c_0 M \sigma_0 \bar{\kappa}^{1/2}}{\underline{\kappa}} \sqrt{\frac{\bar{t} \log(q)}{n}}.$$

Hence for $n \ge C_0 \bar{t} \log(q)$ (for some constant C_0), valid inference on $\theta_\star$ is possible using $\Pi(\cdot|\mathbf{z})$, provided that $\sqrt{\bar{\kappa}}/\underline{\kappa}$ is well-behaved. The parameter $\underline{\kappa}$ measures the finite sample strength (or relevance) of the instruments. That is, how much correlation there is between the instruments and the explanatory variables. In general $\underline{\kappa}$ depends on the dimension p and the sample size n. We will show in Section 2.1 that for modestly large sample size, the finite sample relevance of the instruments is driven mainly by the population relevance of the instruments. This result implies that, provided that the sample size is modestly large, the quasi-posterior distribution delivers the best possible inference, given the available instruments.

The two constraints on ρ in Theorem 3 (namely $n\bar{\kappa} \le \sigma^2 \rho p$, and (9)) suggests choosing ρ such that

$$\frac{n}{p} \lesssim \rho \lesssim \sqrt{\bar{t} n \log(p)},$$

which implies that our results applies particularly in the context where p is larger than n. Importantly (9) suggests that the prior variance (that is $1/\rho$) should not be taken too small. In the applications we recommend the choice

$$\rho = \frac{C_0 \log(pq)}{\sqrt{n}},$$

for some tuning constant C_0. In all the computation we set $C_0 = 1$. Theorem 3 imposes very little constraint on γ and we recommend choosing γ as

$$\gamma = \frac{\gamma_0}{n},$$

for some tuning parameter $\gamma_0 \in (0.1, 1]$. We noted in practice that choosing γ_0 overly small negatively affects the mixing of the Markov Chain Monte Carlo algorithm.

Theorem 3 suggests choosing $\sigma = \sigma_0$. Indeed, if σ is taken smaller than σ_0, we loose in terms of model sparsity: as apparent from (6), choosing $\sigma < \sigma_0$ makes $\bar{s}$ larger, which in turn makes $\bar{t}$ and $\underline{\kappa}$ larger, and a slow convergence rate r ensues. On the other hand, having $\sigma > \sigma_0$ implies that we need a bigger constant M to satisfy the condition $c_0^2 M^2 \min(1, (\sigma_0/\sigma)^2) \geq 4(2+u)$, and a slower convergence also follows. We note however that in most cases it is better to over estimate σ_0 than to underestimate. This was also observed in our numerical experiments (not reported here).

The parameter u determines the sparsity of the prior. The quasi-posterior Π is fairly robust to the choice of u in the range $[1, 2]$. In all the numerical experiments below we set $u = 1$.

sec:event:e

2.1. **On the parameter $\underline{\kappa}$ and the event $\mathcal{E}$.** The population relevance of the instruments is measured by the matrix

$$\mathcal{C} \stackrel{\text{def}}{=} \mathbb{E}(w_1' x_1) \in \mathbb{R}^{q \times p}.$$

Given a matrix A, we write $[A]_{\delta,\delta'}$ to denote the submatrix of A corresponding to the rows of A for which $\delta_i = 1$, and columns of A for which $\delta_j' = 1$. We make the following assumption on $\mathcal{C}$.

H12 **H 3.** *There exist $0 < \underline{\lambda} \leq \bar{\lambda}$, and an integer $s_1, s_2 \geq 1$ such that for all $\delta^{(1)} \in \{0,1\}^q$, $\delta^{(2)} \in \{0,1\}^p$ with $\|\delta^{(1)}\|_0 \leq s_1$, $\|\delta^{(2)}\|_0 \leq s_2$, all the singular values of the submatrix $[\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}$ lie between $\underline{\lambda}$ and $\bar{\lambda}$.*

Suppose that H3 holds with $s_1 = \bar{t}$ and $s_2 = \bar{s} + s_\star$. Then the parameter $\underline{\lambda}$ measures the population relevance of the instruments for the considered model. We show in the next result if the sample size n satisfies $n \geq \bar{t}(\bar{s}+s_\star)(\bar{t}\log(q) + (\bar{s}+s_\star)\log(p))/\underline{\lambda}^2$, then with high probability we can take $\bar{\kappa} = (9/4)\bar{\lambda}^2$, and $\underline{\kappa} = \underline{\lambda}^2/4$ in the definition of $\mathcal{E}$. That is the finite sample relevance of the instruments is similar to the population relevance.

lem:e21 **Lemma 4.** *Assume H3, and suppose that* $\|w_1\|_\infty \|x_1\|_\infty \leq K$ *almost surely for some constant* K*. If the sample size* n *satisfies*

$$\sqrt{n} \geq \frac{6}{\underline{\lambda}} \sqrt{s_1 s_2 \left(s_1 \log(q) + s_2 \log(p)\right)}, \tag{12}$$ ss:1

then with probability at least $1 - 2(s_1+s_2)/[q^{s_1}p^{s_2}]$ *the following holds: for all* $\delta^{(1)} \in \{0,1\}^q$, $\delta^{(2)} \in \{0,1\}^p$ *with* $\|\delta^{(1)}\|_0 \leq s_1$, $\|\delta^{(2)}\|_0 \leq s_2$, *all the singular values of the submatrix* $(W_{\delta^{(1)}})' X_{\delta^{(2)}}$ *lie between* $n\underline{\lambda}/2$ *and* $3n\bar{\lambda}/2$*.*

Proof. See Section 5.3. □

If H3 and (12) hold with $s_1 = \bar{t}$ and $s_2 = \bar{s} + s_\star$, then taking $c_0 = K$, $\bar{\kappa} = (9/4)\bar{\lambda}^2$, and $\underline{\kappa} = \underline{\lambda}^2/4$, by Lemma 4 we have

$$\mathbb{P}(\mathbf{z} \notin \mathcal{E}) \leq \frac{2(\bar{s}+\bar{t})}{q^{\bar{t}} p^{\bar{s}}} + \mathbb{P}\left(\max_{1 \leq k \leq q} |\langle W_k, \epsilon \rangle| > 2c_0 \sigma_0 \sqrt{n \log(q)} \right).$$

By the sub-Gaussian assumption in H1, it holds

$$\mathbb{P}\left(\max_{1 \leq k \leq q} |\langle W_k, \epsilon \rangle| > t | W \right) \leq 2qe^{-\frac{t^2}{2\sigma_0^2 \max_{1 \leq k \leq q} \|W_k\|_2}} \leq 2qe^{-\frac{t^2}{2c_0^2 \sigma_0^2 n}}.$$

At a result, taking $t = 2c_0\sigma_0\sqrt{n\log(q)}$, it follows that

$$\mathbb{P}(\mathbf{z} \notin \mathcal{E}) \leq \frac{2(\bar{s}+\bar{t})}{q^{\bar{t}} p^{\bar{s}}} + \frac{2}{q}.$$

This shows that with high probability we have $\mathbf{z} \in \mathcal{E}$. However the assumption that the variables x and w have uniformly bounded components is restrictive. That assumption is not needed to obtain that $\max_{1 \leq k \leq q} \|W_j\|_2 \leq c_0\sqrt{n}$. For instance, if the random variables W_{ij} are mean-zero independent and identically distributed, and $W_{1j}/\sqrt{\mathsf{Var}(W_{1j})}$ is sub-Gaussian with sub-Gaussian parameter τ_0 say, then it follows from Lemma 1 of Ravikumar et al. (2011) that

$$\max_{1 \leq k \leq q} \|W_j\|_2 \leq \sqrt{3(1+4\tau_0^2)} \max_{1 \leq j \leq q} \sqrt{\mathsf{Var}(W_{1j})}\sqrt{n},$$

with probability at least $1 - 4e^{-n/16}$, provided that $n \geq 16\log(p)$. However the boundedness assumption is needed to apply the matrix Bernstein inequality (Tropp (2015)) used to control the singular values of $W'_{T(\delta)}X_\delta$. We show in the next result that a similar behavior of the singular values of $W'_{T(\delta)}X_\delta$ continue to hold even without the boundedness assumption, albeit in a slightly weaker form.

lem:e22 **Lemma 5.** *Assume H3, and suppose that for some constants* $\alpha > 0$ *and* $m > 0$*,*

$$\max_{1 \leq j \leq p, 1 \leq k \leq q} \mathbb{E}\left(e^{\alpha \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (x_{ij}w_{ik} - \mathbb{E}(x_{ij}w_{ik})) \right|} \right) \leq m < \infty,$$

and

$$\max_{1\le j\le p, 1\le k\le q} \mathbb{E}[(x_{1j}w_{1k})^4] \le m < \infty.$$

Let $\ell_n = \left(\frac{5}{\alpha}\frac{\log(pq)}{\sqrt{n}}\right)^{1/2}$. *If* $\log(pq) \ge \max\left(\frac{\alpha^2 m}{2}, \alpha m^{1/4}\right)$ *and* $\ell_n \le n\underline{\lambda}/(2\sqrt{s_1 s_2})$, *then, with probability at least* $1-\ell_n$, *the following holds: for all* $\delta^{(1)} \in \{0,1\}^q$, $\delta^{(2)} \in \{0,1\}^p$ *with* $\|\delta^{(1)}\|_0 \le s_1$, $\|\delta^{(2)}\|_0 \le s_2$, *all the singular values of all the submatrix* $(W_{\delta^{(1)}})' X_{\delta^{(2)}}$ *lie between* $n\underline{\lambda}/2$ *and* $3n\bar{\lambda}/2$.

Proof. See Section 5.4. □

sec:num

3. Markov Chain Monte Carlo computation and numerical experiments

In this section we develop a practical Markov Chain Monte Carlo algorithm to sample from the posterior distribution Π, and explore the behavior of Π on two simulated data examples. We refer the reader to Tierney (1994); Robert and Casella (2004) for introduction to basic MCMC algorithms.

mcmc˙sampler

3.1. A MCMC sampler for Π .

We begin with a description of the MCMC sampler. We sample from Π using a Metropolized-Gibbs sampler which iteratively alternates between an update of θ given δ, and an update of δ given θ. Given δ, we partition θ into $\theta = ([\theta]_\delta, [\theta]_{\delta^c})$, where $[\theta]_\delta$ groups the components of θ for which $\delta_j = 1$, and $[\theta]_{\delta^c}$ groups the remaining components. These two groups are conditionally independent given δ. Furthermore, given δ, the components of $[\theta]_{\delta^c}$ are independent and identically distributed with distribution $\mathbf{N}(0,\gamma)$. And given δ, $[\theta]_\delta \sim \mathbf{N}(m_\delta, V_\delta)$, where

$$V_\delta = \left(\frac{1}{n\sigma^2} X'_\delta W_{T(\delta)} W'_{T(\delta)} X_\delta + \rho I_{\|\delta\|_0}\right)^{-1}, \text{ and } m_\delta = \frac{1}{n\sigma^2} V_\delta X'_\delta W_{T(\delta)} W'_{T(\delta)} y. \quad (13)$$

mV

It should be noted that the matrix to be inverted above is of dimension $\|\delta\|_0$. Hence if δ is sparse, the update of θ can be done very efficiently.

To update δ, we update each component at the time using an Independent Metropolis-Hastings algorithm where the proposal is a Bernoulli distribution with probability $1/2$. To develop the details, note that,

$$\Pi(\delta|\theta, z) \propto \pi(\delta) \frac{e^{-\frac{1}{2}\theta' C_\delta^{-1}\theta}}{\sqrt{\det(2\pi C_\delta)}}\, e^{-\frac{1}{2n\sigma^2} \sum_{\ell:\,(T(\delta))_\ell = 1} \langle y - X\theta_\delta, W_\ell\rangle^2}. \quad (14)$$

Given an index $j \in [p]$, we form the new state $\delta_{(j)}$ by drawing $(\delta_{(j)})_j \sim \mathsf{Ber}(0.5)$, and $(\delta_{(j)})_k = \delta_k$ for all $k \ne j$. The Metropolis-Hastings acceptance probability is given by

$$\min\left\{1, \frac{\Pi(\delta_{(j)}|\theta, \mathbf{z})}{\Pi(\delta|\theta, \mathbf{z})}\right\},$$

where

$$\frac{\Pi(\delta_{(j)}|\theta,\mathbf{z})}{\Pi(\delta|\theta,\mathbf{z})} = \left(\frac{\alpha}{1-\alpha}\frac{f_1(\theta_j)}{f_0(\theta_j)}\right)^{(\delta_{(j)})_j-\delta_j} \exp\left(-\frac{1}{2n\sigma^2}\sum_{\ell:\,(T(\delta_{(j)}))_\ell=1}\langle y - X(\theta\cdot\delta_{(j)}), W_\ell\rangle^2 + \frac{1}{2n\sigma^2}\sum_{\ell:\,(T(\delta))_\ell=1}\langle y - X(\theta\cdot\delta), W_\ell\rangle^2\right), \quad (15)$$ eq:Aj

where f_0 (resp. f_1) is the density of $\mathbf{N}(0,\gamma)$ (resp. $\mathbf{N}(0,\rho^{-1})$). We note that the computations in (15) can be done efficiently by pre-computing the matrix $W'X$, and $W'y$. We summarize the algorithm as follows.

algo:basic **Algorithm 1.** Draw $(\delta^{(0)},\theta^{(0)}) \in \Delta\times\mathbb{R}^p$ from some initial distribution. For $k = 0,\ldots,$ repeat the following. Given $(\delta^{(k)},\theta^{(k)}) = (\delta,\theta) \in \Delta\times\mathbb{R}^p$:

(STEP 1): For all j such that $\delta_j = 0$, draw $\theta_j^{(k+1)} \sim \mathbf{N}(0,\gamma)$. And if $\|\delta\|_0 > 0$, draw $[\theta^{(k+1)}]_\delta \sim \mathbf{N}(m_\delta, V_\delta)$, with m_δ, V_δ as in (13).

(STEP 2): Given $\theta^{(k+1)} = \theta$, for each $j \in [p]$, update $\delta_j^{(k+1)}$ using the Independent Metropolis algorithm with proposal $\mathsf{Ber}(0.5)$ and acceptance probability given by (15). □

3.2. **Numerical Experiments.** In this section we investigate the performance of our proposed approach via numerical simulations, using the same set up as in Fan and Liao (2014); Belloni et al. (2017). We simulate from a linear model

$$Y = X^T\theta_0 + \epsilon$$

For each component of X, we write $X_j = X_j^e$ if X_j is endogeneous, and $X_j = X_j^x$ if X_j is exogeneous. X_j^e, X_j^x and ϵ are generated according to two different setups which we outline below.

Setup 1:

$$X_j^e = (F_j + H_j + 1)(3\epsilon + 1), \quad X_j^x = F_j + H_j + u_j$$

where $\{\epsilon, u_1,\ldots,u_p\}$ are independent $N(0,1)$. Here $F = (F_1,\ldots,F_p)^T$ and $H = (H_1,\ldots,H_p)^T$ are the transformations of a three-dimensional instrumental variable $V = (V_1,V_2,V_3)^T \sim N(0,\mathrm{I}_3)$ and $W = (F,H)$. There are m endogeneous variables $(X_1,X_2,X_3,X_6,\ldots,X_{2+m})^T$ with $m = \{10,50\}$.

The Fourier basis are applied as the working instruments,

$$F = \sqrt{2}\{\sin(j\pi V_1) + \sin(j\pi V_2) + \sin(j\pi V_3) : j \le p\}$$
$$H = \sqrt{2}\{\cos(j\pi V_1) + \cos(j\pi V_2) + \cos(j\pi V_3) : j \le p\}$$

Setup 2:

$$X_j^e = \tilde{X}_j + \sum_{t=1}^{L} Z_{L(j-1)+t}, \quad \epsilon = \zeta + \tilde{X}'\gamma_0$$

where $\gamma_0 = (.1, .2, .3, \ldots, 1, 0, \ldots)' \in \mathbb{R}^p$, $Z \in \mathbb{R}^{n\times(Lp)}$ with i.i.d. $N(0,1)$ entries, and the rows of $\tilde{X}$ are i.i.d. draws from the p-dimensional $N(0,\Sigma)$, where $\Sigma_{ij} = 0.3^{|i-j|}$, and $\zeta \sim N(0, 1/4^2)$. In this set up, each variable X_j has L instruments $Z_{L(j-1)+1}, \ldots, Z_{L(j-1)+L}$. The first 10 variables are endogenous. Setup 2 is more challenging than Setup 1, because of the correlation between the variables $\tilde{X}_j$ and the small components of γ_0 which creates weaker instruments.

The two setups are taken from Fan and Liao (2014) and Belloni et al. (2017) respectively. For both setups, we choose the true parameter vector $\theta_0 \in \mathbb{R}^p$ with number of non-zero components, $s_\star = 5$, that takes the value

$$\theta_\star = \text{SNR} \times (5, -4, 7, -2, 1.5, 0, \ldots, 0)'$$

where SNR > 0 is a signal-to-noise parameter. Varying the SNR parameter allows us to explore the performance of our approach for varying levels of signal strength. We performed simulations for SNR $= \{0.25, 1\}$, sample size $n = 100$, and number of covariates $p \in \{100, 200\}$. SNR $= 1$ corresponds to high SNR (hSNR) while SNR $= 0.25$ corresponds to weak SNR (wSNR).

In our experiments, we used 100 replications to aggregate the results. Four performance measures are used to compare the methods. The first measure is the number of true positive (TP), that is the number of correctly identified nonzero coefficients. The second measure is the number of incorrectly identified coefficients, the false positive (FP).

$$\text{TP} = \sum_{\delta} \left(\sum_{i:\, \delta_{\star,i}=1} \delta_i \right) \Pi(\delta|\mathbf{z}), \quad \text{and} \quad \text{FP} = \sum_{\delta} \left(\sum_{i:\, \delta_{\star,i}=0} \delta_i \right) \Pi(\delta|\mathbf{z}).$$

The last two measures are mean squared errors, MSE_S and MSE_N, defined as

$$\text{MSE}_S = \int \frac{1}{s_\star} \sum_{i:\, \delta_{\star,i}=1} (\theta_i - \theta_{\star,i})^2 \, \Pi(\mathrm{d}\theta|\mathbf{z}), \quad \text{and} \quad \text{MSE}_N = \int \frac{1}{p - s_\star} \sum_{i:\, \delta_{\star,i}=0} \theta_i^2 \, \Pi(\mathrm{d}\theta|\mathbf{z}).$$

The expectations in these definitions are approximated by averaging over the MCMC run. Standard errors on these measures are obtained from the 100 MCMC replications. In each run of the MCMC sampler, θ is initialized using the lasso solution and δ is initialized as the support of the lasso solution. Our proposed method has four

tuning parameters. In all our empirical work, we set

$$u = 1, \quad \rho = \frac{\log(p * q)}{\sqrt{n}}, \quad \gamma = \frac{1}{n}, \quad \sigma^2 = 1.$$

FGMM results are obtained using the code on the authors' website by setting the FGMM parameter $\lambda_{\text{fgmm}} = 0.3$. The summary of our results is presented in Tables 1 - 2 which compare our method, quasi-Bayesian moment restrictions model (BMRM), with FGMM and penalized least squares (PLS).

In the high signal-to-noise regime (SNR = 1), PLS performs well in selecting the true coefficients but, at the same time, includes a significantly large number of false positives. FGMM reduces the number of unimportant coefficients while keeping the important coefficients in the model. In contrast, BMRM not only selects all the important coefficients but also succeeds in weeding out all the unimportant coefficients. Our proposed method stands out in this regard. Comparisons along the mean square errors is not as clear cut, which is expected since posterior means tends to be high uncertainly due to the use of a prior distribution. The lower panels of the tables display results for the weak signal-to-noise regime (SNR = 0.25) case. Again, BMRM outperforms FGMM in selecting the important regressors and removing the unimportant regressors.

TABLE 1. Setup 1: Endogeneity in both important and unimportant regressors, $n = 100$, $m = 10$, $s_0 = 5$. Top and bottom panels correspond to hSNR and wSNR regimes respectively.

	BMRM				**FGMM**				**PLS**			
p	TP	FP	MSE_S	MSE_N	TP	FP	MSE_S	MSE_N	TP	FP	MSE_S	MSE_N
100	5 · 0 (0·0)	0 · 0 (0·0)	0 · 10 (0·001)	0 · 0002 (0·0001)	5 · 00 (0·0)	3 · 14 (1·14)	0 · 002 (0·002)	0 · 0 (0·0)	5 · 0 (0·0)	59 · 08 (14·98)	0 · 02 (0·03)	0 · 003 (0·004)
200	5 · 0 (0·0)	0 (0)	0 · 097 (0·0005)	0 · 0001 (0·0000)	4 · 99 (0·10)	3 · 29 (1·42)	0 · 007 (0·05)	0 (0)	5 (0)	98 · 48 (28·62)	0 · 15 (0·22)	0 · 01 (0·02)
100	4 · 24 (0·548)	0 · 01 (0·03)	0 · 095 (0·001)	0 · 001 (0·001)	4 · 36 (0·67)	3 · 18 (1·20)	0 · 03 (0·04)	0 · 000 (0·000)	4 · 99 (0·1)	21 · 41 (10·77)	0 · 01 (0·001)	0 · 000 (0·000)
200	4 · 40 (0·56)	0 · 05 (0·17)	0 · 020 (0·001)	0 · 0005 (0·0005)	4 · 36 (0·66)	3 · 29 (1·13)	0 · 03 (0·04)	0 (0)	4 · 96 (0·20)	30 · 02 (16·91)	0 · 01 (0·01)	0 · 000 (0·000)

table:setup1

4. Application to Angrist & Krueger's (1991) model ??? Endogeneity in Angrist & Krueger Data

Angrist and Krueger (1991) use the large samples available in the 1980 U.S. Census to estimate return to schooling. The endogeneity of *level of education* in the *wage*

TABLE 2. Setup 2: Endogeneity in all regressors, $n = 100$, $L = 2$, $s_0 = 5$. Top and bottom panels correspond to hSNR and wSNR regimes respectively.

	BMRM				**FGMM**				**PLS**			
p	TP	FP	MSE_S	MSE_N	TP	FP	MSE_S	MSE_N	TP	FP	MSE_S	MSE_N
100	$5 \cdot 0$ (0·0)	$0 \cdot 2$ (0·4)	$0 \cdot 03$ (0·10)	$0 \cdot 002$ (0·001)	$4 \cdot 79$ (0·50)	$2 \cdot 93$ (1·98)	$0 \cdot 34$ (0·45)	$0 \cdot 002$ (0·005)	5 (0)	$8 \cdot 28$ (3·58)	$0 \cdot 07$ (0·05)	$0 \cdot 008$ (0·002)
200	$5 \cdot 0$ (0·0)	$0 \cdot 5$ (0·68)	$0 \cdot 22$ (0·05)	$0 \cdot 001$ (0·001)	$4 \cdot 69$ (0·58)	$3 \cdot 06$ (2·14)	$0 \cdot 39$ (0·45)	$0 \cdot 001$ (0·003)	5 (0)	$10 \cdot 70$ (5·68)	$0 \cdot 10$ (0·08)	$0 \cdot 005$ (0·002)
100	$3 \cdot 25$ (0·6)	$0 \cdot 20$ (0·40)	$0 \cdot 11$ (0·03)	$0 \cdot 006$ (0·002)	$2 \cdot 98$ (1·09)	$2 \cdot 54$ (1·91)	$0 \cdot 54$ (0·88)	$0 \cdot 004$ (0·008)	$4 \cdot 35$ (0·48)	$4 \cdot 73$ (1·31)	$0 \cdot 08$ (0·03)	$0 \cdot 007$ (0·002)
200	$3 \cdot 23$ (0·64)	$0 \cdot 42$ (0·81)	$0 \cdot 22$ (0·06)	$0 \cdot 003$ (0·002)	$3 \cdot 05$ (1·02)	$2 \cdot 78$ (2·13)	$0 \cdot 42$ (0·45)	$0 \cdot 002$ (0·004)	$4 \cdot 25$ (0·52)	$5 \cdot 61$ (2·37)	$0 \cdot 09$ (0·04)	$0 \cdot 004$ (0·001)

table:setup2

equation has led them to consider in their seminal work *quarter of birth* as instrumental variable for *level of education*. They argue that individuals born in early quarters are more likely to drop out of school earlier than those born in late quarters and they back up this correlation with data. It also makes sense that, controlling for schooling, *quarter of birth* is independent of *wage*, hence this instrument is exogenous. Using standard IV methods, they estimate return to schooling to be 0.0928 for a sample of 329,509 males born in 1930-1939 using the 1980 U.S. Census.

The use of Quarter of birth as IV has been criticised by Bound et al. (1995) who mention that the correlation between *level of education* and *quarter of birth* is actually weak and as a result, the IV estimator is likely inconsistent and IV inference misleading. Subsequently, Cruz and Moreira (2005) find that the instrument is informative enough to allow for meaningful inference and Hoogerheide and van Dijk (2006) carried out a Bayesian inference on Angrist and Krueger's (1991) model and data. They obtain a median of posterior distribution of return-to-schooling of 0.106 with [0.083; 0.129] as 95% credible interval; results close to those of Angrist and Krueger (1991).

The linear IV model of Angrist and Krueger (1991), formally introduced below, controls for a large number of covariates including a total of 501 explanatory variables and this qualifies as a large model (even though the number of observations exceeds the number of explanatory variables and instruments). Our goal in this section is to carry out inference on this model using the high-dimensional Bayesian model selection method developed in this paper which we consider to be more reliable in this context.

The linear IV model of Angrist and Krueger (1991) is given by:

$$y_i = \langle x_i, \theta \rangle + \epsilon_i, \quad \mathbb{E}(\epsilon_i | w_i) = 0$$

where y_i is the *log(wage)* of individual i and x_i denotes a set of 510 variables: education, 9 year-of-birth (YOB) dummies, 50 state-of-birth (SOB) dummies, and 450 state-of-birth $\times$ year-of-birth (YOB$\times$SOB) interactions. For individual i, we write

$$x_i = [\text{Education}_i, \text{YOB}_i, \text{SOB}_i, (\text{YOB} \times \text{SOB})_i] \in \mathbb{R}^{510\times 1}$$

As instruments, w_i, we use 3 quarter-of-birth dummies (QOB) for the endogeneous variable education, and allow the exogeneous variables to be instruments for themselves. For individual i, we write

$$w_i = [\text{QOB}_i, \text{YOB}_i, \text{SOB}_i, (\text{YOB} \times \text{SOB})_i] \in \mathbb{R}^{512\times 1}$$

Note that there is an irregular dependence between the variables x_i and their corresponding instruments w_i. For example, if the endogenous variable *level of education* is active, then all 3 instruments, corresponding to QOB, are included in the model.

We apply our inference method following the steps in Algorithm 1 to jointly select the most relevant variables/instruments out the 510/512 considered and obtain an estimate of the posterior distribution of the parameter of interest: the return-to-schooling (θ_1). In our implementation, we set the tuning parameters **... to... values???**.

The posterior distribution of θ_1 is plotted in Figure 1. The estimated mean of this distribution is 0.1096 with a 95% credible interval for θ_1 given by $[0.096, 0.129]$. A total of 9 covariates are selected, namely: selected are **Add the list of selected variables...** along with the set of instruments given by **Add the list of related instruments....**.

This result means that, everything being equal, an extra year of education increases expected wage by about 0.1096%. This value is close to the posterior median reported by Hoogerheide and van Dijk (2006). Our 95%-credible interval has the save upper bound as theirs but interestingly is narrower. This advantage highlights the efficiency gain expected from model selection that is built in our procedure.

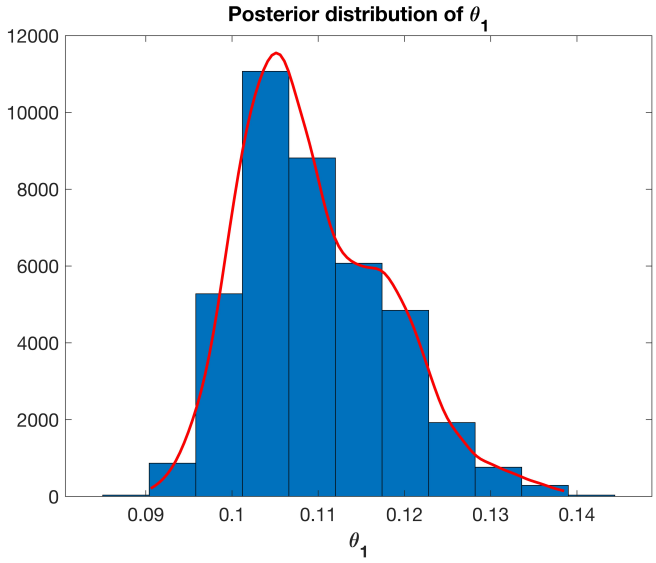

FIGURE 1. Posterior distribution of θ_1 which summarizes the causal impact of education on earning. The posterior mean is .1096.

theta˙post

5. PROOFS

sec:proofs

Our methods of proof follows closely Atchade and Bhattacharyya (2018). For $\delta \in \Delta \stackrel{\text{def}}{=} \{0,1\}^p$, we will write $\mu_\delta(\mathrm{d}\theta)$ to denote the product measure on $\mathbb{R}^p$ given by

$$\mu_\delta(\mathrm{d}\theta) \stackrel{\text{def}}{=} \prod_{j=1}^{p} \mu_{\delta_j}(\mathrm{d}\theta_j),$$

where $\mu_0(\mathrm{d}x)$ is the Dirac mass at 0, and $\mu_1(\mathrm{d}x)$ is the Lebesgue measure on $\mathbb{R}$. Given $(\delta,\theta) \in \Delta \times \mathbb{R}^p$, we set

$$q_{\delta,\theta}(\mathbf{z}) \stackrel{\text{def}}{=} \exp\left[-\frac{1}{2n\sigma^2}\,(y - X\theta)'\, W_{T(\delta)} W'_{T(\delta)}\,(y - X\theta)\right]. \tag{16}$$

def:q

First we derive a lower bound on the normalizing constant. Although the quasi-likelihood function used here is slightly more general than in (Atchade and Bhattacharyya (2018)), the proof of the lower-bound proceeds similarly.

lem:control:nc

Lemma 6. *Assume H1-H2. Let $\mathcal{C}(\mathbf{z})$ denote the normalizing constant of $\Pi(\cdot|z)$. For $z \in \mathcal{E}_0$,*

$$\mathcal{C}(\mathbf{z}) \geq \pi(\delta_\star) q_{\delta_\star,\theta_\star}(\mathbf{z}) e^{-\frac{\rho}{2}\|\theta_\star\|_2^2} \left(\frac{\rho}{\frac{n\bar{\kappa}}{\sigma^2} + \rho}\right)^{\frac{s_\star}{2}}. \tag{17}$$

eq:control:nc:check

Proof. By definition we have

$$\mathcal{C}(\mathbf{z}) = \sum_{\delta\in\Delta} \pi(\delta) \int_{\mathbb{R}^p} q_{\delta,\theta}(\mathbf{z}) \frac{e^{-\frac{1}{2}\theta' B_\delta^{-1}\theta}}{\sqrt{\det(2\pi B_\delta)}} \mathrm{d}\theta$$
$$\geq \pi(\delta_\star) q_{\delta_\star,\theta_\star}(\mathbf{z}) \left(\frac{\rho}{2\pi}\right)^{\frac{s_\star}{2}} \int_{\mathbb{R}^p} \frac{q_{\delta_\star,\theta}(\mathbf{z})}{q_{\delta_\star,\theta_\star}(\mathbf{z})} e^{-\frac{\rho}{2}\|\theta\|_2^2} \mu_{\delta_\star}(\theta).$$

With $G(\mathbf{z}) = \nabla \log q_{\delta_\star,\theta_\star}(\mathbf{z})$, we have

$$\log q_{\delta_\star,\theta}(\mathbf{z}) - \log q_{\delta_\star,\theta_\star}(\mathbf{z}) = \langle G(\mathbf{z}), \theta - \theta_\star\rangle - \frac{1}{2n\sigma^2}(\theta-\theta_\star)' X' W_{T(\delta_\star)} W'_{T(\delta_\star)} X(\theta-\theta_\star).$$

We recall that $M_\delta = W'_{T(\delta)} X_\delta$, so that for $z \in \mathcal{E}_0$,

$$\log q_{\delta_\star,\theta}(\mathbf{z}) - \log q_{\delta_\star,\theta_\star}(\mathbf{z}) \geq \langle G(\mathbf{z}), \theta - \theta_\star\rangle - \frac{n\bar{\kappa}}{2\sigma^2}\|\theta-\theta_\star\|_2^2.$$

Hence,

$$\mathcal{C}(\mathbf{z}) \geq \pi(\delta_\star) q_{\delta_\star,\theta_\star}(\mathbf{z}) \left(\frac{\rho}{2\pi}\right)^{\frac{s_\star}{2}} e^{-\frac{\rho}{2}\|\theta_\star\|_2^2}$$
$$\int_{\mathbb{R}^p} e^{\langle G(\mathbf{z}), \theta-\theta_\star\rangle - \frac{\rho}{2}\left(\|\theta\|_2^2 - \|\theta_\star\|_2^2\right) - \frac{n\bar{\kappa}}{2\sigma^2}\|\theta-\theta_\star\|_2^2} \mu_{\delta_\star}(\mathrm{d}\theta).$$

We have $-\frac{\rho}{2}\left(\|\theta\|_2^2 - \|\theta_\star\|_2^2\right) = -\frac{\rho}{2}\|\theta-\theta_\star\|_2^2 - \rho\langle\theta_\star, \theta-\theta_\star\rangle$. Therefore,

$$\int_{\mathbb{R}^p} e^{\langle G(\mathbf{z}), \theta-\theta_\star\rangle - \frac{\rho}{2}\left(\|\theta\|_2^2 - \|\theta_\star\|_2^2\right) - \frac{n\bar{\kappa}}{2\sigma^2}\|\theta-\theta_\star\|_2^2} \mu_{\delta_\star}(\mathrm{d}\theta)$$
$$= \int_{\mathbb{R}^p} e^{\langle G(\mathbf{z}) - \rho\theta_\star, u-\theta_\star\rangle - \frac{\frac{n\bar{\kappa}}{\sigma^2}+\rho}{2}\|u-\theta_\star\|_2^2} \mu_{\delta_\star}(\mathrm{d}u) \geq \left(\frac{2\pi}{\frac{n\bar{\kappa}}{\sigma^2}+\rho}\right)^{\frac{s_\star}{2}},$$

and (17) follows easily. □

sec:proof:lem:sparsity

5.1. **Proof of Proposition 2.** By the control on the normalizing constant provided by Lemma 6, for any $k \geq 0$, we have

$$\mathbf{1}_{\mathcal{E}_0}(z)\Pi\left(\|\delta\|_0 > k | z\right) \leq \left(1 + \frac{n\bar{\kappa}}{\sigma^2\rho}\right)^{\frac{s_\star}{2}}$$
$$\sum_{\delta\in\Delta:\ \|\delta\|_0 > k} \frac{\pi(\delta)}{\pi(\delta_\star)} \left(\frac{\rho}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \mathbf{1}_{\mathcal{E}_0}(z) \int_{\mathbb{R}^p} \frac{e^{\ell(\delta,u;z) - \frac{\rho}{2}\|u\|_2^2}}{e^{\ell(\delta_\star,\theta_\star;z) - \frac{\rho}{2}\|\theta_\star\|_2^2}} \mu_\delta(\mathrm{d}u).$$

We note that $1 + n\bar{\kappa}/(\sigma^2\rho) \leq 2p \leq p^2$ by assumption. And for $z \in \mathcal{E}_0$,

$$\frac{e^{\ell(\delta,\theta_\star;z)}}{e^{\ell(\delta_\star,\theta_\star;z)}} \leq e^{-\ell(\delta_\star,\theta_\star;z)} = \exp\left(\frac{1}{2n\sigma^2}\epsilon'\left[W_{T(\delta_\star)} W'_{T(\delta_\star)}\right]\epsilon\right) \leq e^{2c_0^2\left(\frac{\sigma_0}{\sigma}\right)^2 \|T(\delta_\star)\|_0 \log(q)}.$$

Hence

$$\mathbf{1}_{\mathcal{E}_0}(Z)\Pi\left(\|\delta\|_0 > k|Z\right) \leq p^{s_\star} e^{2c_0^2\left(\frac{\sigma_0}{\sigma}\right)^2\|T(\delta_\star)\|_0\log(q)} e^{\frac{\rho}{2}\|\theta_\star\|_2^2} \sum_{\delta\in\Delta:\ \|\delta\|_0>k} \frac{\pi(\delta)}{\pi(\delta_\star)}\left(\frac{\rho}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \times \int_{\mathbb{R}^p} e^{-\frac{\rho}{2}\|u\|_2^2}\mu_\delta(\mathrm{d}u)$$

$$= p^{s_\star} e^{2c_0^2\left(\frac{\sigma_0}{\sigma}\right)^2\|T(\delta_\star)\|_0\log(q)} e^{\frac{\rho}{2}\|\theta_\star\|_2^2} \sum_{\delta\in\Delta:\ \|\delta\|_0>k} \frac{\pi(\delta)}{\pi(\delta_\star)}. \quad (18)$$

control:prob:3

By H2, we have

$$\sum_{\delta:\ \|\delta\|_0\geq s_\star+k} \frac{\pi(\delta)}{\pi(\delta_\star)} \quad = \quad \sum_{j=s_\star+k}^{p} \binom{p}{j}\left(\frac{\alpha}{1-\alpha}\right)^{j-s_\star} \quad \leq \quad \binom{p}{s_\star}\sum_{j=s_\star+k}^{p}\left(\frac{1}{p^u}\right)^{j-s_\star},$$

using the fact that $\frac{\mathsf{q}}{1-\mathsf{q}} = \frac{1}{p^{u+1}}$, and $\binom{p}{j} \leq p^{j-s_\star}\binom{p}{s_\star}$. Hence for $p^u \geq 2$,

$$\sum_{\delta:\ \|\delta\|_0\geq s_\star+k} \frac{\pi(\delta)}{\pi(\delta_\star)} \leq \frac{2\binom{p}{s_\star}}{p^{ku}} \leq \frac{2}{p^{ku-s_\star}}.$$

We conclude that

$$\mathbf{1}_{\mathcal{E}_0}(Z)\Pi\left(\|\delta\|_0 > s_\star + k|Z\right) \leq e^{2c_0^2\left(\frac{\sigma_0}{\sigma}\right)^2\|T(\delta_\star)\|_0\log(q)} e^{\frac{\rho}{2}\|\theta_\star\|_2^2}\frac{2}{p^{ku-s_\star}} \leq \frac{2}{p^{m_0}},$$

by choosing k as in the statement of the theorem. This completes the proof. □

5.2. **Proof of Theorem 3.** Our proofs rely on the existence of some testing procedures that we derive following the same arguments as in Atchade and Bhattacharyya (2018). Let $\mathcal{Z} = \mathbb{R}^n \times \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times q}$ equipped with its Lebesgue measure. Let $f_\star$ be a density on $\mathcal{Z}$.

test **Lemma 7.** *Assume H1, and let $f_\star$ denote the distribution of $\mathbf{z} = (y, X, W)$. For any constant $M > 2$ such that $M^2c_0^2 \geq 2(\sigma/\sigma_0)^2$, there exists a measurable function $\phi : \mathcal{Z} \to [0, 1]$ such that*

$$\int_{\mathcal{Z}} \phi(\mathbf{z}) f_\star(\mathbf{z})\mathrm{d}z \leq \frac{4}{\bar{q}^{\bar{t}}}.$$

Furthermore, for all $\delta \in \Delta_{\bar{s}}$ and all $\theta \in \mathbb{R}^p_\delta$ such that $\|\theta - \theta_\star\|_2 > jMr$ for some $j \geq 1$, we have

$$\int_{\mathcal{E}_1} (1-\phi(\mathbf{z}))\frac{q_{\delta,\theta}(\mathbf{z})}{q_{\delta,\theta_\star}(\mathbf{z})} f_\star(\mathbf{z})\mathrm{d}z \leq e^{-\frac{n\kappa}{32\sigma^2}(jMr)^2}.$$

Proof. Fix $M > 2$. Fix $\delta \in \Delta_{\bar{s}}$, and with $q_{\delta,\theta}$ as in (16), we define

$$\bar{q}_{\delta,u}(\mathbf{z}) \stackrel{\text{def}}{=} \frac{q_{\delta,u}(\mathbf{z})}{q_{\delta,\theta_\star}(\mathbf{z})} f_\star(\mathbf{z}) \mathbf{1}_{\mathcal{E}}(\mathbf{z}), \quad u \in \mathbb{R}^p_\delta, \ \ \mathbf{z} \in \mathcal{Z},$$

and

$$\begin{aligned} \mathcal{L}(\delta, u; \mathbf{z}) &\stackrel{\text{def}}{=} \log q_{\delta,u}(\mathbf{z}) - \log q_{\delta,\theta_\star}(\mathbf{z}) - \langle \nabla \log q_{\delta,\theta_\star}(\mathbf{z}), u - \theta_\star \rangle, \\ &= -\frac{1}{2n\sigma^2} (u - \theta_\star)' X' W_{T(\delta)} W'_{T(\delta)} X (u - \theta_\star). \end{aligned}$$

Hence

$$\bar{q}_{\delta,u}(\mathbf{z}) = e^{\left\langle \nabla \log q_{\delta,\theta_\star}(\mathbf{z}), u - \theta_\star \right\rangle + \mathcal{L}(\delta,u;\mathbf{z})} f_\star(\mathbf{z}) \mathbf{1}_{\mathcal{E}}(\mathbf{z}).$$

For $\mathbf{z} \in \mathcal{E}$,

$$\begin{aligned} |\langle \nabla \log q_{\delta,\theta_\star}(\mathbf{z}), u - \theta_\star \rangle| &= \frac{1}{n\sigma^2} \left| \left\langle X' W_{T(\delta)} W'_{T(\delta)} \epsilon, u - \theta_\star \right\rangle \right| \\ &\leq \frac{1}{n\sigma^2} \|W'_{T(\delta)} X (u - \theta_\star)\|_1 \max_{1 \leq j \leq q} \|W'_j \epsilon\|_2 \\ &\leq 2c_0 \left(\frac{\sigma_0}{\sigma} \right) \sqrt{\frac{\|T(\delta)\|_0 \bar{\kappa} n \log(q)}{\sigma^2}} \|u - \theta_\star\|_2. \end{aligned}$$

Then, using the properties of the set $\mathcal{E}$, we conclude that for all $\delta \in \Delta_{\bar{s}}$, $u \in \mathbb{R}^p_\delta$, and $\mathbf{z} \in \mathcal{E}$,

$$\begin{aligned} \bar{q}_{\delta,u}(\mathbf{z}) &\leq e^{2c_0 \left(\frac{\sigma_0}{\sigma}\right) \sqrt{\frac{\|T(\delta)\|_0 \bar{\kappa} n \log(q)}{\sigma^2}} \|u - \theta_\star\|_2 - \frac{n\kappa}{2\sigma^2} \|u - \theta_\star\|_2^2} f_\star(\mathbf{z}) \mathbf{1}_{\mathcal{E}_1}(\mathbf{z}) \\ &\leq e^{\frac{n\kappa}{4\sigma^2} \|u - \theta_\star\|_2^2} f_\star(\mathbf{z}) \mathbf{1}_{\mathcal{E}_1}(\mathbf{z}), \quad (19) \end{aligned}$$

bound:ratio:test

whenever $\|u - \theta_\star\|_2 > \mathsf{r}$, where r is as defined in (10). Therefore $\int_{\mathcal{E}} \bar{q}_{\delta,u}(\mathbf{z}) \mathrm{d}\mathbf{z} < \infty$, for all $\delta \in \Delta_{\bar{s}}$, and $u \in \mathbb{R}^p_\delta$. Now, fix $\eta \geq 2\mathsf{r}$, $\delta \in \Delta_{\bar{s}}$, and $\theta \in \mathbb{R}^p_\delta$, such that $\|\theta - \theta_\star\|_2 > \eta$. Let

$$\mathcal{P}_{\delta,\theta} \stackrel{\text{def}}{=} \left\{ \bar{q}_{\delta,u} : \ u \in \mathbb{R}^p_\delta, \ \|u - \theta\|_2 \leq \frac{\eta}{2} \right\}.$$

According to Lemma 13 of Atchade and Bhattacharyya (2018), applied with $p = f_\star$, and $\mathcal{Q} = \mathcal{P}_{\delta,\theta}$, there exists a test function $\phi_{\delta,\theta}$ such that

$$\sup_{q \in \mathcal{P}_{\delta,\theta}} \left[\int \phi_{\delta,\theta} f_\star + \int (1 - \phi_{\delta,\theta}) q \right] \leq \sup_{q \in \mathsf{conv}(\mathcal{P}_{\delta,\theta})} \int_{\mathcal{Z}} \sqrt{f_\star q}. \quad (20)$$

lem:test:eq1

Any element $q \in \mathsf{conv}(\mathcal{P}_{\delta,\theta})$ can be written as $q = \sum_j \alpha_j \bar{q}_{\delta,u_j}(z)$, where $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$, and $\|u_j - \theta\|_2 \leq \eta/2$ (hence $\|u_j - \theta_\star\|_2 > \eta/2$). It then follows from (19) that

$$\int_{\mathcal{Z}} \sqrt{f_\star q} \leq e^{\frac{n\kappa\eta^2}{32\sigma^2}}.$$

Hence (20) becomes

$$\sup_{q\in\mathcal{P}_{\delta,\theta}}\left[\int\phi_{\delta,\theta}f_\star+\int(1-\phi_{\delta,\theta})q\right]\le e^{\frac{n\kappa\eta^2}{32\sigma^2}}. \tag{21}$$

lem:test:eq2

Now write $\cup_\delta\{\theta\in\mathbb{R}^p_\delta:\ \|\theta-\theta_\star\|_2>M\mathsf{r}\}$ as $\cup_\delta\cup_{j\ge1}\mathcal{A}_\epsilon(\delta,j)$, where the unions in δ are taken over all δ such that $\|\delta\|_0\le\bar{s}$, and

$$\mathcal{A}_\epsilon(\delta,j)\stackrel{\text{def}}{=}\left\{\theta\in\mathbb{R}^p_\delta:\ jM\mathsf{r}<\|\theta-\theta_\star\|_2\le(j+1)M\mathsf{r}\right\}.$$

For $\mathcal{A}_\epsilon(\delta,j)\neq\emptyset$, let $\mathcal{S}(\delta,j)$ be a maximally $(jM\mathsf{r}/2)$-separated points in $\mathcal{A}_\epsilon(\delta,j)$. It is easily checked that the cardinality of $\mathcal{S}(\delta,j)$ is upper bounded by $9^{\|\delta\|_0}\le9^{\bar{s}}$. For $\theta_{\delta,jk}\in\mathcal{S}(\delta,j)$, let $\phi_{\delta,\theta_{\delta,jk}}$ denote the test function obtained above with $\theta=\theta_{\delta,jk}$ and $\eta=jM\mathsf{r}$. From (21) $\phi_{\delta,\theta_{\delta,jk}}$ satisfies

$$\max_{u\in\mathbb{R}^p_\delta,\ \|u-\theta_{\delta,jk}\|_2\le\frac{jM\mathsf{r}}{2}}\left[\mathbb{E}_\star(\phi_{\delta,\theta_{\delta,jk}}(Z))+\int_{\mathcal{E}_1}(1-\phi_{\delta,\theta_{\delta,jk}}(\mathbf{z}))\bar{q}_{\delta,u}(\mathbf{z})\mathrm{d}\mathbf{z}\right]\le e^{-\frac{n\kappa}{32\sigma^2}(jM\mathsf{r})^2}. \tag{22}$$

lem:test:eq3

Then we set

$$\phi=\max_{\delta:\ \|\delta\|_0\le\bar{s}}\ \sup_{j\ge1}\ \max_{\theta_{\delta,jk}\in\mathcal{S}(\delta,j)}\phi_{\delta,\theta_{\delta,jk}}.$$

It then follows that

$$\begin{aligned}\mathbb{E}\left(\phi(Z)\right)&\le\sum_{k=0}^{\bar{s}}\sum_{\delta:\ \|\delta\|_0=k}\sum_{j\ge1}\sum_{\theta_{\delta,jk}\in\mathcal{S}(\delta,j)}\mathbb{E}_\star\left(\phi_{\delta,\theta_{\delta,jk}}(Z)\right)\\&\le\sum_{k=0}^{\bar{s}}\binom{p}{k}9^k\sum_{j\ge1}e^{-\frac{n\kappa}{32\sigma^2}(jM\mathsf{r})^2}\le\frac{2(9p)^{\bar{s}}e^{-\frac{n\kappa}{32\sigma^2}(M\mathsf{r})^2}}{1-e^{-\frac{n\kappa}{32\sigma^2}(M\mathsf{r})^2}}\le4e^{-\frac{n\kappa}{32\sigma^2}(M\mathsf{r})^2}\le\frac{4}{q^{\bar{t}}},\end{aligned}$$

for all $p\ge9$ if $M>2$ is taken such that $M^2c_0^2\ge2(\sigma/\sigma_0)^2$.

If for some δ, such that $\|\delta\|_0\le\bar{s}$ and some $\theta\in\mathbb{R}^p_\delta$ we have $\|\theta-\theta_\star\|_2>jM\bar{\mathsf{r}}$, then θ resides within $(iM\bar{\mathsf{r}})/2$ of some point $\theta_{\delta,ik}\in\mathcal{S}(\delta,i)$ for some $i\ge j$. Hence, by (22),

$$\int_{\mathcal{E}}(1-\phi(\mathbf{z}))\bar{q}_{\delta,\theta}(\mathbf{z})\mathrm{d}\mathbf{z}\le\int_{\mathcal{E}}(1-\phi_{\delta,\theta_{\delta,jk}}(\mathbf{z}))\bar{q}_{\delta,\theta}(\mathbf{z})\mathrm{d}\mathbf{z}\le e^{-\frac{n\kappa}{32\sigma^2}(jM\mathsf{r})^2}.$$

This ends the proof. □

sec:proof:thm:1

5.2.1. *Proof of Theorem 3.* We have $\Delta\times\mathbb{R}^p=((\Delta\setminus\Delta_{\bar{s}})\times\mathbb{R}^p)\cup\bar{\mathcal{F}}_1\cup\bar{\mathcal{F}}_2\cup\mathsf{B}$, where

$$\bar{\mathcal{F}}_1\stackrel{\text{def}}{=}\bigcup_{\delta\in\Delta_{\bar{s}}}\{\delta\}\times\mathcal{F}_1^{(\delta)},\qquad\bar{\mathcal{F}}_2\stackrel{\text{def}}{=}\bigcup_{\delta\in\Delta_{\bar{s}}}\{\delta\}\times\mathcal{F}_2^{(\delta)},$$

where $\mathcal{F}_1^{(\delta)}\stackrel{\text{def}}{=}\{\theta\in\mathbb{R}^p:\ \|\theta\cdot\delta-\theta_\star\|_2>M\epsilon\}$, and
$\mathcal{F}_2^{(\delta)}\stackrel{\text{def}}{=}\left\{\theta\in\mathbb{R}^p:\ \|\theta\cdot\delta-\theta_\star\|_2\le M\epsilon,\ \text{ and }\|\theta-\theta\cdot\delta\|_2>m\sqrt{\gamma p}\right\}$. Therefore we have

$$1-\Pi(\bar{\mathsf{B}}|\mathbf{z})=\Pi(\|\delta\|_0>\bar{s}|\mathbf{z})+\Pi(\bar{\mathcal{F}}_1|\mathbf{z})+\Pi(\bar{\mathcal{F}}_2|\mathbf{z}).$$

We know from Lemma 2 that

$$\mathbf{1}_{\mathcal{E}}(\mathbf{z})\Pi(\|\delta\|_0 > \bar{s}|\mathbf{z}) \leq \frac{2}{p^{m_0}}.$$

Setting $\mathcal{F}_{22}^{(\delta)} = \{\theta \in \mathbb{R}^p : \ \|\theta - \theta_\delta\|_2 > m\sqrt{\gamma p}\}$, it is straightforward to see that

$$\Pi(\bar{\mathcal{F}}_2|\mathbf{z}) \leq \max_{\delta \in \Delta_{\bar{s}}} \mathbb{P}\left(V \in \mathcal{F}_{22}^{(\delta)}\right),$$

where $V \sim \mathbf{N}_p(0, \gamma I_p)$. By standard Guassian deviation bound, $\mathbb{P}(V \in \mathcal{F}_{22}^{(\delta)}) \leq 2e^{-\frac{(m-1)^2 p}{2}}$ for all $\delta \in \Delta_{\bar{s}}$. It follows that for all $\mathbf{z} \in \mathcal{Z}$, $\Pi(\bar{\mathcal{F}}_2|\mathbf{z}) \leq 2e^{-\frac{(m-1)^2 p}{2}}$.

Let ϕ denote the test function asserted by Lemma 7. We can then write

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})\Pi(\bar{\mathcal{F}}_1|\mathbf{z})\right] \leq \mathbb{E}\left[\phi(\mathbf{z})\right] + \mathbb{E}\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})\left(1 - \phi(\mathbf{z})\right)\Pi(\bar{\mathcal{F}}_1|\mathbf{z})\right].$$

Lemma 7 gives

$$\mathbb{E}\left[\phi(\mathbf{z})\right] \leq \frac{4}{q^{\bar{t}}}.$$

By Lemma 6, we have

$$\Pi(\bar{\mathcal{F}}_1|\mathbf{z})\mathbf{1}_{\mathcal{E}}(Z) \leq \left(1 + \frac{n\bar{\kappa}}{\sigma^2\rho}\right)^{\frac{s_\star}{2}} \times \mathbf{1}_{\mathcal{E}}(\mathbf{z}) \sum_{\delta \in \Delta_{\bar{s}}} \frac{\pi(\delta)}{\pi(\delta_\star)} \left(\frac{\rho}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \int_{\mathcal{F}_1} \frac{q_{\delta,\theta}(\mathbf{z})}{q_{\delta_\star,\theta_\star}(\mathbf{z})} e^{-\frac{\rho}{2}\left(\|\theta\|_2^2 - \|\theta_\star\|_2^2\right)} \mu_\delta(\mathrm{d}\theta),$$

where $\mathcal{F}_1 \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p : \ \|\theta - \theta_\star\|_2 \leq M\epsilon\}$. We have

$$\frac{q_{\delta,\theta_\star}(\mathbf{z})}{q_{\delta_\star,\theta_\star}(\mathbf{z})} = \exp\left(\frac{1}{2n\sigma^2}\epsilon\left[W_{T(\delta_\star)}W'_{T(\delta_\star)} - W_{T(\delta)}W'_{T(\delta)}\right]\epsilon\right) \leq \exp\left(\frac{1}{2n\sigma^2}\epsilon\left[W_{T(\delta_\star)}W'_{T(\delta_\star)}\right]\epsilon\right),$$

and for $\mathbf{z} \in \mathcal{E}$, $\epsilon\left[W_{T(\delta_\star)}W'_{T(\delta_\star)}\right]\epsilon \leq 4c_0^2\sigma_0^2\bar{t}n\log(q)$. It follows from the above and Fubini's theorem that

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})(1 - \phi(\mathbf{z}))\Pi(\bar{\mathcal{F}}_1|\mathbf{z})\right] \leq e^{2c_0^2\bar{t}\left(\frac{\sigma_0}{\sigma}\right)^2\log(q)}p^{s_\star} \times \sum_{\delta \in \Delta_{\bar{s}}} \frac{\omega_\delta}{\omega_{\delta_\star}} \left(\frac{\rho}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \int_{\mathcal{F}_1} \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})\left(1 - \phi(\mathbf{z})\right)\frac{q_{\delta,\theta}(\mathbf{z})}{q_{\delta,\theta_\star}(\mathbf{z})}\right] e^{-\frac{\rho}{2}\left(\|\theta\|_2^2 - \|\theta_\star\|_2^2\right)} \mu_\delta(\mathrm{d}\theta), \quad (23)$$

eq:proof:thm:contra

We write $\mathcal{F}_1 = \cup_{j\geq 1}\mathcal{F}_{1,j}$, where $\mathcal{F}_{1,j} \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p : \ jM\epsilon < \|\theta-\theta_\star\|_2 \leq (j+1)M\epsilon\}$. Using this and Lemma 7, we have

$$\int_{\mathcal{F}_{1,j}} \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})\,(1-\phi(\mathbf{z}))\,\frac{q_{\delta,\theta}(\mathbf{z})}{q_{\delta,\theta_\star}(\mathbf{z})}\right] e^{-\frac{\rho}{2}\left(\|\theta\|_2^2-\|\theta_\star\|_2^2\right)}\mu_\delta(\mathrm{d}\theta)$$
$$\leq e^{-\frac{n\underline{\kappa}}{32\sigma^2}(jM\mathsf{r})^2}\int_{\mathcal{F}_{1,j}} e^{-\frac{\rho}{2}\left(\|\theta\|_2^2-\|\theta_\star\|_2^2\right)}\mu_\delta(\mathrm{d}\theta),$$

and

$$\int_{\mathcal{F}_{1,j}} e^{-\frac{\rho}{2}(\|\theta\|_2^2-\|\theta_\star\|_2^2)}\mu_\delta(\mathrm{d}\theta) = \int_{\mathcal{F}_{1,j}} e^{-\frac{\rho}{2}\left(\|\theta-\theta_\star\|_2^2+2\langle\theta_\star,\theta-\theta_\star\rangle\right)}\mu_\delta(\mathrm{d}\theta)$$
$$\leq e^{2\rho\|\theta_\star\|_2(jM\mathsf{r})}\int_{\mathbb{R}^p} e^{-\frac{\rho}{2}\|\theta-\theta_\star\|_2^2}\mu_\delta(\mathrm{d}\theta) \leq e^{2\rho\|\theta_\star\|_2(jM\mathsf{r})}\left(\frac{2\pi}{\rho}\right)^{\frac{\|\delta\|_2}{2}}.$$

Therefore (23) becomes

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})(1-\phi(\mathbf{z}))\Pi(\bar{\mathcal{F}}_1|\mathbf{z})\right]$$
$$\leq p^{s_\star}e^{2c_0^2\bar{t}\left(\frac{\sigma_0}{\sigma}\right)^2\log(q)}\sum_{\delta\in\Delta_{\bar{s}}}\frac{\pi(\delta)}{\pi(\delta_\star)}\sum_{j\geq 1}e^{-\frac{n\underline{\kappa}}{32\sigma^2}(jM\mathsf{r})^2+2\rho\|\theta_\star\|_2(jM\mathsf{r})}$$
$$\leq p^{s_\star}e^{2\bar{t}\left(\frac{\sigma_0}{\sigma}\right)^2\log(q)}\sum_{\delta\in\Delta_{\bar{s}}}\frac{\pi(\delta)}{\pi(\delta_\star)}\frac{e^{-\frac{n\underline{\kappa}}{64\sigma^2}(M\mathsf{r})^2}}{1-e^{-\frac{n\underline{\kappa}}{64\sigma^2}(M\mathsf{r})^2}}$$
$$\leq 2p^{s_\star}e^{2\bar{t}\left(\frac{\sigma_0}{\sigma}\right)^2\log(q)}e^{-\frac{n\underline{\kappa}}{64\sigma^2}(M\mathsf{r})^2}\sum_{\delta\in\Delta_{\bar{s}}}\frac{\pi(\delta)}{\pi(\delta_\star)}, \quad (24)$$ eq:proof:thm:contra

where we use (9) to conclude that for all $j \geq 1$,

$$-\frac{n\underline{\kappa}}{64}(jM\mathsf{r})^2 + 2\rho\|\theta_\star\|_2(jM\mathsf{r}) \leq 0.$$

We note that for $\alpha \leq 1/2$, and since $\binom{p}{s} \leq p^s$,

$$\sum_{\delta\in\Delta_{\bar{s}}}\frac{\pi(\delta)}{\pi(\delta_\star)} = \left(\frac{1-\alpha}{\alpha}\right)^{s_\star}\sum_{\delta\in\Delta_{\bar{s}}}\left(\frac{\alpha}{1-\alpha}\right)^{\|\delta\|_0} \leq \left(\frac{1-\alpha}{\alpha}\right)^{s_\star}\sum_{s=0}^{\bar{s}}\binom{p}{s}(2\alpha)^s$$
$$\leq p^{s_\star(1+u)}\sum_{s=0}^{\bar{s}}(2p\alpha)^s \leq 2p^{s_\star(1+u)},$$

provided that $p^u \geq 4$. It follows readily from (24) that

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{E}}(\mathbf{z})(1-\phi(\mathbf{z}))\Pi(\bar{\mathcal{F}}_1|\mathbf{z})\right] \leq \frac{4}{q^{\bar{t}}}. \quad (25)$$ eq:proof:thm:contra

The result follows by putting the pieces together.

□

sec:proof:lem:e21

5.3. **Proof of Lemma 4.** Let $\mathcal{C}_n \stackrel{\text{def}}{=} W'X$ which can also be written as $\mathcal{C}_n = \sum_{i=1}^n w_i x_i' = n\mathcal{C} + \sum_{i=1}^n (w_i x_i' - \mathcal{C})$. Likewise, given $\delta^{(1)}$, $\delta^{(2)}$, with $\|\delta^{(1)}\|_0 \leq s_1$, $\|\delta^{(2)}\|_0 \leq s_2$, the submatrix $(W_{\delta^{(1)}})'X_{\delta^{(2)}} = [\mathcal{C}_n]_{\delta^{(1)},\delta^{(2)}}$ can be written as

$$(W_{\delta^{(1)}})'X_{\delta^{(2)}} = [n\mathcal{C}]_{\delta^{(1)},\delta^{(2)}} + \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}).$$

Note that for all $\delta^{(1)}$, $\delta^{(2)}$, with $\|\delta^{(1)}\|_0 \leq s_1$, $\|\delta^{(2)}\|_0 \leq s_2$

$$\left\| [w_i x_i']_{\delta^{(1)},\delta^{(2)}} \right\|_2 \leq \sqrt{s_1 s_2}\|x_1\|_\infty \|w_1\|_\infty \leq K\sqrt{s_1 s_2}. \tag{26}$$

spec:norm

Standard results on perturbation of singular values (see e.g. Golub and Van Loan (2013) Corollary 2.4.4), together with H3 imply that if $\Sigma_{\delta^{(1)},\delta^{(2)}}$ denotes a singular value of $(W_{\delta^{(1)}})'X_{\delta^{(2)}}$, we have

$$\begin{aligned} n\underline{\lambda} - \left\| \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}) \right\|_2 &\leq \Sigma_{\delta^{(1)},\delta^{(2)}} \\ &\leq n\bar{\lambda} + \left\| \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}) \right\|_2 . \end{aligned} \tag{27}$$

eq:res:iso

Using (26), the matrix Bernstein inequality (see e.g. Theorem 6.1.1 of Tropp (2015)), and union bound, for any $a > 0$, we have

$$\begin{aligned} &\mathbb{P}\left[\max_{\delta^{(1)},\delta^{(2)}} \left\| \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}) \right\|_2 > a \right] \\ &\qquad \leq 2\binom{q}{s_1}\binom{p}{s_2}(s_1+s_2) e^{-\frac{a^2}{2ns_1s_2K^2 + \frac{2}{3}\sqrt{s_1 s_2}Ka}}. \end{aligned}$$

Taking $a^2 = 8s_1s_2(s_1\log(q) + s_2\log(p))K^2 n$, we work out that the right hand side of the last display is upper bounded by $2(s_1+s_2)/[q^{s_1}p^{s_2}]$, provided that $\sqrt{n} \geq \sqrt{s_1\log(q) + s_2\log(p)}$. Hence (27) reads: with probability at least $2(s_1+s_2)/[q^{s_1}p^{s_2}]$ it holds: for all $\delta^{(1)}$, $\delta^{(2)}$, with $\|\delta^{(1)}\|_0 \leq s_1$, $\|\delta^{(2)}\|_0 \leq s_2$

$$n\underline{\lambda} - K\sqrt{8s_1s_2(s_1\log(q) + s_2\log(p))n} \leq \Sigma_{\delta^{(1)},\delta^{(2)}} \leq n\bar{\lambda} + K\sqrt{8s_1s_2(s_1\log(q) + s_2\log(p))n},$$

which easily imply the stated result.

□

sec:proof:lem:e22

5.4. **Proof of Lemma 5.** The proof is similar to the proof of Lemma 4, but we control the random variable

$$\max_{\delta^{(1)}:\, \|\delta^{(1)}\|_0 \le s_1} \quad \max_{\delta^{(2)}:\, \|\delta^{(2)}\|_0 \le s_2} \quad \left\| \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}) \right\|_2,$$

differently. Note that

$$\max_{\delta^{(1)}:\, \|\delta^{(1)}\|_0 \le s_1} \quad \max_{\delta^{(2)}:\, \|\delta^{(2)}\|_0 \le s_2} \quad \frac{1}{n} \left\| \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}) \right\|_2 \le \sqrt{s_1 s_2} \left\| \frac{1}{n} \sum_{i=1}^n (w_i x_i' - \mathcal{C}) \right\|_\infty. \quad (28)$$

1

Applying Lemma 8, we have:

$$\mathbb{E}\left(\left\| \frac{1}{n} \sum_{i=1}^n (w_i x_i' - \mathcal{C}) \right\|_\infty \right) \le \ell_n^2.$$

As a result,

$$\mathbb{P}\left(\sqrt{s_1 s_2} \left\| \frac{1}{n} \sum_{i=1}^n (w_i x_i' - \mathcal{C}) \right\|_\infty \ge \sqrt{s_1 s_2}\ell_n \right) \le \ell_n.$$

Using (28), we can claim that, with probability at least as large as $1 - \ell_n$,

$$\sup_{\delta^{(1)},\delta^{(2)}:\|\delta^{(j)}\|_0 \le s_j} \left\| \sum_{i=1}^n ([w_i x_i']_{\delta^{(1)},\delta^{(2)}} - [\mathcal{C}]_{\delta^{(1)},\delta^{(2)}}) \right\|_2 \le n\sqrt{s_1 s_2}\ell_n.$$

The result follows thanks to (27) and the fact that $\sqrt{s_1 s_2}\ell_n \le \frac{n\lambda}{2}$.

lemma˙1

Lemma 8. *Let $\{x_i : i = 1, \ldots\}$ be a sequence of independent $\mathbb{R}^q$-valued random vectors such that $\mathbb{E}(x_i) = 0$ and, there exists $\alpha > 0$ such that*

$$\max_{1 \le j \le q} \mathbb{E}\left(e^{\alpha \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} \right|} \right) \le M < \infty \quad \text{and} \quad \max_{1 \le i \le n, 1 \le j \le q} \mathbb{E}(x_{ij}^4) \le M < \infty,$$

for some constant $M > 0$, where x_{ij} is the jth component of x_i.

Then, if $\log q \ge \alpha^2 M/2 \vee \alpha M^{1/4}$,

$$\mathbb{E}\left(\left\| \frac{1}{n} \sum_{i=1}^n x_i \right\|_\infty \right) \le c_1 \frac{\log q}{\sqrt{n}},$$

with $c_1 = 5/\alpha$.

Proof. Let $Y_n = \left\|\frac{1}{n}\sum_{i=1}^n x_i\right\|_\infty$ and $0 \le t \le \alpha/2$. By the Jensen's inequality, we have

$$\begin{aligned} e^{t\mathbb{E}(Y_n)} &\le \mathbb{E}e^{tY_n} = \mathbb{E}e^{t\max\limits_{1\le j\le q}\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|} = \mathbb{E}\left(\max_{1\le j\le q} e^{t\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|}\right) \\ &\le \sum_{j=1}^q \mathbb{E}e^{t\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|} \le q\max_{1\le j\le q}\mathbb{E}e^{t\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|}. \end{aligned} \tag{29}$$

eq:bound1

By a second-order Taylor expansion around 0, we have

$$e^{\frac{t}{n}\left|\sum_{i=1}^n x_{ij}\right|} = 1 + \left|\sum_{i=1}^n x_{ij}\right|\frac{t}{n} + \frac{1}{2}\left|\sum_{i=1}^n x_{ij}\right|^2 e^{\dot{t}\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|}\cdot\frac{t^2}{n^2},$$

with $\dot{t} \in (0,t)$. Thus,

$$\begin{aligned} \mathbb{E}e^{\frac{t}{n}\left|\sum_{i=1}^n x_{ij}\right|} &= 1 + \mathbb{E}\left|\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n x_{ij}\right|\tfrac{t}{\sqrt{n}} + \tfrac{1}{2}\mathbb{E}\left(\left|\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n x_{ij}\right|^2 e^{\dot{t}\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|}\right)\cdot\frac{t^2}{n} \\ &\equiv 1 + A\tfrac{t}{\sqrt{n}} + \tfrac{1}{2}B\tfrac{t^2}{n}. \end{aligned}$$

We now proceed to bound A and B. Note that:

$$A \le \left(\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n x_{ij}^2\right)\right)^{1/2} \le \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}(x_{ij}^2)\right)^{1/2} \le \left(\max_{i,j}\mathbb{E}(x_{ij})^2\right)^{1/2} \le M^{1/4}.$$

$$\begin{aligned} B &\le \mathbb{E}\left(\left|\tfrac{1}{\sqrt{n}}\sum_{i=1}^n x_{ij}\right|^2 e^{t\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\right|}\right) \le \left(\mathbb{E}\left(\tfrac{1}{\sqrt{n}}\sum_{i=1}^n x_{ij}\right)^4 \mathbb{E}e^{2\frac{t}{\sqrt{n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n x_{ij}\right|}\right)^{1/2} \\ &\le \sqrt{M}\left(\mathbb{E}\left(\tfrac{1}{\sqrt{n}}\sum_{i=1}^n x_{ij}\right)^4\right)^{1/2} \equiv \sqrt{M}C^{1/2}, \end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the third one holds by assumption so long as $2t/\sqrt{n} \le \alpha$. It is not hard to see that

$$\begin{aligned} C &= \tfrac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n x_{ij}\right)^4 = \tfrac{1}{n^2}\left[\sum_{i=1}^n \mathbb{E}(x_{ij}^4) + 6\left(\textstyle\sum_{i\ne i'}\mathbb{E}(x_{ij}^2)\mathbb{E}(x_{i'j}^2)\right)\right] \\ &= \tfrac{1}{n^2}\left(\sum_{i=1}^n \mathbb{E}(x_{ij}^4) + 6\left(\sum_{i=1}^n \mathbb{E}(x_{ij}^2)\right)^2 - 6\sum_{i=1}^n[\mathbb{E}(x_{ij}^2)]^2\right) \\ &\le (6 + 7/n)\tfrac{1}{n}\sum_{i=1}^n \mathbb{E}(x_{ij}^4) \le 13M. \end{aligned}$$

(The second to last inequality is obtain by a repeated application of the Jensen's inequality.)

As a result, $B \leq \sqrt{13M}\sqrt{M} \leq 4M$. Thus we can claim that, for all j,

$$\mathbb{E}e^{t\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right|} \leq 1 + M^{1/4}\frac{t}{\sqrt{n}} + 2M\frac{t^2}{n}.$$

Then, from (29), we have

$$t\mathbb{E}(Y_n) \leq \log q + \log\left(1 + \frac{M^{1/4}}{\sqrt{n}}t + 2M\frac{t^2}{n}\right) \leq \log q + \frac{M^{1/4}}{\sqrt{n}}t + 2M\frac{t^2}{n}.$$

Hence,

$$\mathbb{E}(Y_n) \leq \frac{\log q}{t} + \frac{M^{1/4}}{\sqrt{n}} + \frac{2M}{n}t.$$

Using the lower bound on $\log q$, we have

$$\mathbb{E}(Y_n) \leq \log q\left(\frac{1}{t} + \frac{4}{n\alpha^2}t\right) + \frac{\log q}{\alpha\sqrt{n}}.$$

The right-hand-side of this inequality is minimized at $t^* = \alpha\sqrt{n}/2$ and we have

$$\mathbb{E}Y_n \leq \frac{4\log q}{\alpha\sqrt{n}} + \frac{\log q}{\alpha\sqrt{n}} = \frac{5\log q}{\alpha\sqrt{n}}.$$

Which completes the proof.

□

References

angrist1991does — Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* **106** 979–1014.

AB:18 — Atchade, Y. and Bhattacharyya, A. (2018). Regularization and Computation with high-dimensional spike-and-slab posterior distributions. *ArXiv e-prints* arXiv:1803.10282.

atchade:15b — Atchade, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *Ann. Statist.* **45** 2248–2273.

belloni2012sparse — 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.

belloni2017simultaneous — Belloni, A., Chernozhukov, V., Hansen, C. and Newey, W. (2017). Simultaneous confidence intervals for high-dimensional linear models with many endogenous variables. *arXiv preprint arXiv:1712.08102* .

Bound:etal:1995 — Bound, J., Jeager, D. A. and Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90** 443–450.

buhlGeer11 BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics, Springer, Heidelberg. Methods, theory and applications.

candes2007dantzig CANDES, E., TAO, T. ET AL. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics* **35** 2313–2351.

caner:etal:17 CANER, M., HAN, X. and LEE, Y. (2018). Adaptive elastic net gmm estimation with many invalid moment conditions: Simultaneous model and moment selection. *Journal of Business & Economic Statistics* **36** 24–46.

castillo:etal:14 CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018.

chernozhukov2003mcmc CHERNOZHUKOV, V. and HONG, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics* **115** 293–346.

Cruz:Moreira:2005 CRUZ, L. M. and MOREIRA, M. J. (2005). On the validity of econometric techniques with weak instruments: Inference on returns to education using compulsory school attendance laws. *The Journal of Human Resources* **40** 393–410.

fan2001variable FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360.

fan2014endogeneity FAN, J. and LIAO, Y. (2014). Endogeneity in high dimensions. *Annals of statistics* **42** 872.

gautier2014high GAUTIER, E. and TSYBAKOV, A. (2014). High-dimensional instrumental variables regression and confidence sets. Tech. rep., HAL.

gautier2011high GAUTIER, E., TSYBAKOV, A. and ROSE, C. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454* .

george1997approaches GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica* 339–373.

golub:vl GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix Computations*. 4th ed. Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD.

r9 GREENFIELD, A., HAFEMEISTER, C. and BONNEAU.R (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* **29** 1060 – 067.

hansen1982large HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.

hastie:etal:15 HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Hoogerheide:van:Dijk:2006 HOOGERHEIDE, L. and VAN DIJK, H. K. (2006). A reconsideration of the angrist-krueger analysis on returns to education. Tech. Rep. No. EI 2006-15, Econometric

Institute.

imbens2014instrumental IMBENS, G. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science* **29** 323–358.

kato2013quasi KATO, K. ET AL. (2013). Quasi-bayesian analysis of nonparametric instrumental variables models. *The Annals of Statistics* **41** 2359–2390.

liao2011posterior LIAO, Y., JIANG, W. ET AL. (2011). Posterior consistency of nonparametric conditional moment restricted models. *The Annals of Statistics* **39** 3003–3031.

ma:etal:18 MA, Y.-A., CHEN, Y., JIN, C., FLAMMARION, N. and JORDAN, M. I. (2018). Sampling Can Be Faster Than Optimization. *arXiv e-prints* arXiv:1811.08413.

mitchell1988bayesian MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1032.

r11 PENG, B., ZHU, D., ANDER, B. P., ZHANG, X., XUE, F., SHARP, F. and YANG, X. (2013). An integrative framework for bayesian variable selection with informative priors for identifying genes and pathways. *PLoS ONE* **8(7)**.

ravikumaretal11 RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980.

robertetcasella04 ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo statistical methods*. 2nd ed. Springer Texts in Statistics, Springer-Verlag, New York.

stock2003retrospectives STOCK, J. H. and TREBBI, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives* **17** 177–194.

r10 STUDHAM, M., TJARNBERG, A., NORDLING, T., NELANDER, S. and SONNHAMMER, E. (2014). Functional association networks as priors for gene regulatory network inference. *Bioinformatics* **30** i130 – i138.

tierney94 TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. With discussion and a rejoinder by the author.

tropp:15 TROPP, J. A. (2015). An Introduction to Matrix Concentration Inequalities. *arXiv e-prints* arXiv:1501.01571.