# Inference on Treatment and Structural Effects Conditional on Observables

**Methods and Theoretical Results.** We consider the following partially linear model,

$$y_{1i} = d_i\alpha_0 + g(z_i) + \zeta_i, \tag{6.45}$$

$$d_i = m(z_i) + v_i, \tag{6.46}$$

$$\begin{pmatrix} \zeta_i \\ v_i \end{pmatrix} \mid z_i \sim N\left(0, \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix}\right) \tag{6.47}$$

where $d_i$ is a policy/treament variable whose impact we would like to infer, and $z_i$ represents confounding factors on which we need to condition. This model is of interest in our international

**Instrumental Variables Model Simulation Results**

| | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|
| Estimator | RMSE | Med. Bias | rp(.05) | $\|\widehat{\Pi}\|_0 = 0$ | RMSE | Med. Bias | rp(.05) | $\|\widehat{\Pi}\|_0 = 0$ |
| No Signal | | | | | | | | |
| 2SLS(All) | 0.318 | 0.305 | 0.862 | | 0.312 | 0.297 | 0.852 | |
| FULL(All) | 2.398 | 0.248 | 0.704 | | 1.236 | 0.318 | 0.066 | |
| IV-Lasso | 0.511 | 0.338 | 0.014 | 455 | 0.477 | 0.296 | 0.012 | 486 |
| FULL-Lasso | 0.509 | 0.338 | 0.010 | 455 | 0.477 | 0.296 | 0.012 | 486 |
| IV-Lasso-CV | 0.329 | 0.301 | 0.652 | 0 | 0.478 | 0.299 | 0.064 | 348 |
| FULL-Lasso-CV | 0.359 | 0.305 | 0.384 | 0 | 0.474 | 0.299 | 0.054 | 348 |
| Sup-Score | | | 0.004 | | | | 0.010 | |
| $F^* = 10$ | | | | | | | | |
| 2SLS(All) | 0.058 | 0.058 | 0.806 | | 0.026 | 0.025 | 0.808 | |
| FULL(All) | 0.545 | 0.050 | 0.690 | | 0.816 | 0.006 | 0.052 | |
| IV-Lasso | 0.055 | 0.020 | 0.042 | 147 | 0.027 | 0.009 | 0.056 | 160 |
| FULL-Lasso | 0.054 | 0.020 | 0.032 | 147 | 0.027 | 0.009 | 0.044 | 160 |
| IV-Lasso-CV | 0.052 | 0.024 | 0.072 | 10 | 0.027 | 0.009 | 0.054 | 202 |
| FULL-Lasso-CV | 0.051 | 0.022 | 0.068 | 10 | 0.027 | 0.009 | 0.044 | 202 |
| Sup-Score | | | 0.006 | | | | 0.004 | |
| $F^* = 40$ | | | | | | | | |
| 2SLS(All) | 0.081 | 0.072 | 0.626 | | 0.036 | 0.032 | 0.636 | |
| FULL(All) | 0.951 | 0.050 | 0.690 | | 0.038 | 0.000 | 0.036 | |
| IV-Lasso | 0.051 | 0.012 | 0.048 | 1 | 0.022 | 0.003 | 0.048 | 0 |
| FULL-Lasso | 0.051 | 0.011 | 0.046 | 1 | 0.022 | 0.002 | 0.038 | 0 |
| IV-Lasso-CV | 0.048 | 0.016 | 0.058 | 0 | 0.022 | 0.004 | 0.052 | 0 |
| FULL-Lasso-CV | 0.049 | 0.014 | 0.050 | 0 | 0.022 | 0.003 | 0.042 | 0 |
| Sup-Score | | | 0.004 | | | | 0.006 | |
| $F^* = 160$ | | | | | | | | |
| 2SLS(All) | 0.075 | 0.062 | 0.306 | | 0.034 | 0.029 | 0.334 | |
| FULL(All) | 1.106 | 0.023 | 0.622 | | 0.026 | 0.002 | 0.044 | |
| IV-Lasso | 0.049 | 0.005 | 0.064 | 0 | 0.022 | 0.002 | 0.044 | 0 |
| FULL-Lasso | 0.049 | 0.002 | 0.056 | 0 | 0.022 | 0.001 | 0.040 | 0 |
| IV-Lasso-CV | 0.048 | 0.006 | 0.054 | 0 | 0.022 | 0.002 | 0.040 | 0 |
| FULL-Lasso-CV | 0.049 | 0.003 | 0.048 | 0 | 0.022 | 0.000 | 0.038 | 0 |
| Sup-Score | | | 0.004 | | | | 0.010 | |

TABLE 3. Results are based on 500 simulation replications. $F^*$ measures the strength of the instruments as outlined in the text. We report root-mean-square-error (RMSE), median bias (Med. Bias), rejection frequency for 5% level tests (rp(.05)), and the number of times the Lasso-based procedures select no instruments ($\|\widehat{\Pi}\|_0 = 0$). Further details are provided in the text.

growth example discussed in the next section as well as in many empirical studies (Heckman, LaLonde, and Smith 1999, Imbens 2004). The confounding factors affect the policy variable via $m(z_i)$. We assume that $m(z_i)$ and $g(z_i)$ each admit an approximately sparse form and use linear combinations of technical control terms $x_i = P(z_i)$ to approximate them.

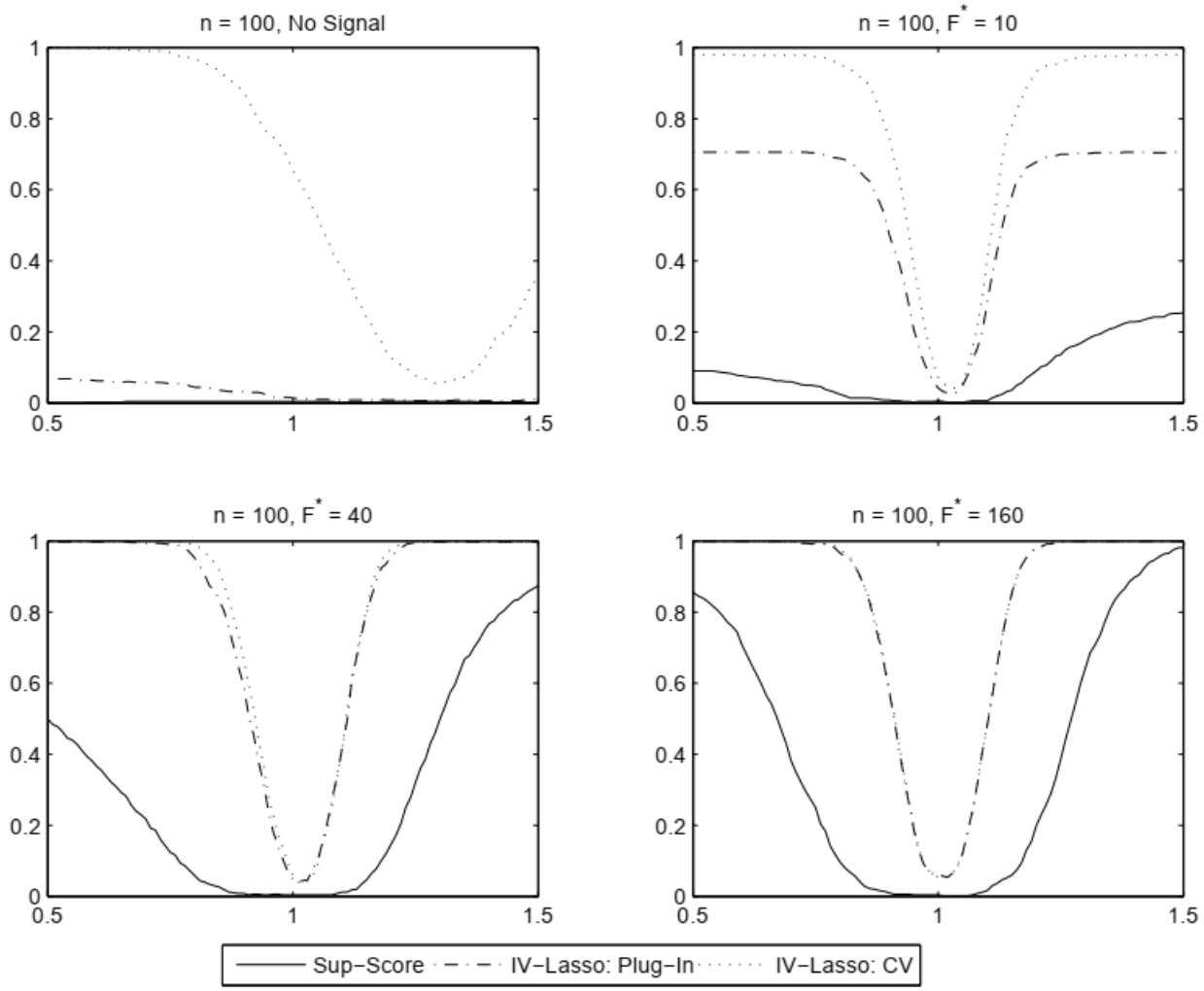

FIGURE 2. Power curves for Sup-Score test, IV-Lasso with Iterated penalty, and IV-Lasso with penalty selected by 10-Fold Cross-Validation from IV simulation with 100 observations.

There are at least three obvious strategies for inference:

(i) Estimate $\alpha_0$ by applying a Feasible Lasso method to model (6.45) without penalizing $\alpha_0$,
(ii) Estimate $\alpha_0$ by applying a Post-Lasso method to model (6.45) without penalizing $\alpha_0$,
(iii) Estimate $\alpha_0$ by applying an Indirect Post-Lasso where $\alpha_0$ is estimated by running standard least squares regression of $y$ on $d$ and control terms selected in a preliminary Feasible Lasso regression of $d_i$ on $x_i$ in (6.46).

Note that it is most natural not to penalize $\alpha_0$ since the goal is to quantify the impact of $d_i$. (The previous rate results derived in Theorems 1 and 2 for the regression function extend to the case where the coefficients on a fixed number of variables are not penalized.) In what follows, we shall refer to options (i), (ii), and (iii) respectively as Lasso, Post-Lasso, and Indirect Post-Lasso.

Regarding inference, "intuition" suggests that if $g$ can be estimated at faster than the $n^{1/4}$ rate then any of (i)-(iii) could be $\sqrt{n}$-consistent and asymptotically normal. It turns out that this "intuition" is often correct for options (ii) and (iii) but is wrong for option (i). Indeed, it is possible to show that under rather strong regularity conditions that

$$(\sigma_\zeta^2[\mathbb{E}_n v_i^2]^{-1})^{-1/2}\sqrt{n}(\bar{\alpha} - \alpha_0) = N(0,1) + o_P(1), \tag{6.48}$$

where $\sigma_\zeta^2[\mathbb{E}_n v_i^2]^{-1}$ is the semi-parametric efficiency bound for estimating $\alpha_0$, for $\bar{\alpha}$ denoting the estimators (ii) or (iii) above. Unfortunately, the distributional result (6.48) is not very robust to modest violations of regularity conditions and may provide a poor approximation to the finite-sample distributions of the estimators for $\alpha_0$. The reason is that Lasso applied to (6.45) may miss important terms relating $d_i$ to $z_i$ through $m(z_i)$ and thus suffer from substantial omitted variables bias. On the other hand, Lasso applied only to (6.46), even if successful in selecting adequate controls for the relationship between $d_i$ and $z_i$, may miss important terms in $g(z_i)$ and thus be highly inefficient. We illustrate this lack of robustness through a simulation experiment reported below.

Instead of using Lasso, Post-Lasso, or Indirect Post-Lasso, we advocate a "double-Post-Lasso" method. To define this estimator, we write the reduced form corresponding to (6.45)-(6.46):

$$y_{1i} = \alpha_0 m(z_i) + g(z_i) + \alpha_0 v_i + \zeta_i, \tag{6.49}$$

$$d_i = m(z_i) + v_i. \tag{6.50}$$

Now we have two equations and hence can apply Lasso methods to each equation to select control terms. That is, we run Lasso regression of $y_{1i}$ on $x_i = P(z_i)$ and Lasso regression of $d_i$ on $x_i = P(z_i)$. Then we can run least squares of $y_{1i}$ on $d_i$ and the union of the controls selected in each equation to estimate and perform inference on $\alpha_0$. By using this procedure we increase the chances for successfully recovering terms that approximate the key control term $m(z_i)$, which results in improved robustness properties. Indeed, the resulting procedure is considerably more robust in computational experiments and requires much weaker regularity conditions than the obvious strategies outlined above.

Now we formally define the double-Post-Lasso estimator. Let $\widehat{I}_1 = \text{support}(\widehat{\beta}_1)$ denote the control terms selected by a feasible Lasso estimator $\widehat{\beta}_1$ computed using data $(y_i, x_i) = (d_i, x_i), i = 1, ..., n$. Let $\widehat{I}_2 = \text{support}(\widehat{\beta}_2)$ denote the control terms selected by a feasible Lasso estimator $\widehat{\beta}_2$ computed using data $(y_i, x_i) = (y_{1i}, x_i), i = 1, ..., n$. The double-Post-Lasso estimator $\check{\alpha}$ of $\alpha_0$ is defined as the least squares estimator obtained by regressing $y_{1i}$ on $d_i$ and the selected control terms $x_{ij}$ with $j \in \widehat{I} \supseteq \widehat{I}_1 \cup \widehat{I}_2$:

$$(\check{\alpha}, \check{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\text{argmin}} \{\mathbb{E}_n[(y_{1i} - d_i\alpha - x_i'\beta)^2] \ : \ \beta_j = 0, \forall j \notin \widehat{I}\}.$$

The set $\widehat{I}$ can contain other variables with names $\widehat{I}_3$ that the analyst may think are important for ensuring robustness. Thus, $\widehat{I} = \widehat{I}_1 \cup \widehat{I}_2 \cup \widehat{I}_3$; let $\widehat{s} = |\widehat{I}|$ and $\widehat{s}_j = |\widehat{I}_j|$ for $j = 1, 2, 3$.

**Condition ASTE**. *(i) The data $(y_{1i}, d_i, z_i), i = 1, ..., n$, obeys model (6.45)-(6.47) for each $n$, and $x_i = P(z_i)$ is a dictionary of transformations of $z_i$. (ii) The parameter values $\sigma_v^2$ and $\sigma_\zeta^2$ are bounded from above by $\bar{\sigma}$ and away from zero, uniformly in $n$, and $|\alpha_0|$ is bounded uniformly in $n$. (iii) Regressor values $x_i, i = 1, ..., n$, obey the normalization condition $\mathbb{E}_n[x_{ij}^2] = 1$ for all $j \in \{1, ..., p\}$ and sparse eigenvalue condition SE. (iv) There exists $s \geqslant 1$ and $\beta_{m0}$ and $\beta_{g0}$ such that*

$$m(z_i) = x_i'\beta_{m0} + r_{mi}, \quad \|\beta_{m0}\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{mi}^2]\}^{1/2} \leqslant K\bar{\sigma}\sqrt{s/n}, \tag{6.51}$$

$$g(z_i) = x_i'\beta_{g0} + r_{gi}, \quad \|\beta_{g0}\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{gi}^2]\}^{1/2} \leqslant K\bar{\sigma}\sqrt{s/n}, \tag{6.52}$$

*where $K$ is an absolute constant, independent of $n$, but all other parameter values can depend $n$. (v) $s^2 \log^2(p \vee n) = o(n)$ and $\widehat{s}_3 \lesssim 1 \vee \widehat{s}_1 \vee \widehat{s}_2$.*

**Theorem 5** (Inference on Treatment Effects)**.** *Suppose condition ASTE holds. The double-Post-Lasso estimator $\check{\alpha}$ obeys,*

$$(\sigma_\zeta^2[\mathbb{E}_n v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha} - \alpha_0) = N(0,1) + o_P(1).$$

*Moreover, the result continues to apply if $\sigma_\zeta^2$ is replaced by $\widehat{\sigma}_\zeta^2 = \mathbb{E}_n[(y_{1i} - d_i\check{\alpha} - x_i'\check{\beta})^2](n/(n - \widehat{s} - 1))$ and $\mathbb{E}_n[v_i^2]$ by $\mathbb{E}_n[\widehat{v}_i^2] = \min_{\beta \in \mathbb{R}^p}\{\mathbb{E}_n[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}\}$.*

**Comment 6.1.** Theorem 5, derived by the second-named author, shows that the double-Post-Lasso estimator asymptotically achieves the semi-parametric efficiency bound under a set of technical conditions and the following key growth condition: $s^2 \log^2(p \vee n) = o(n)$. This rate condition requires the conditional expectations to be sufficiently smooth so that a relatively small number of series terms can be used to approximate them well. As in the case of the IV estimator, this condition can be replaced with the weaker condition that $s \log(p \vee n) = o(n)$ by employing a sample splitting method of Fan, Guo, and Hao (2011). This is done in a companion paper, which also deals with a more general setup, covering non-Gaussian, heteroscedastic disturbances (Belloni, Chernozhukov, and Hansen 2011). □

**Comment 6.2.** The post double selection estimator is formulated in response to the inferential non-robustness properties of the post single selection procedures. The non-robustness of the latter is in line with the uniformity/robustness critique developed by Potscher (2009). The post double selection procedure developed here is in part motivated as a constructive response to this uniformity critique. The need for such constructive response was stressed by Hansen (2005). The goal here is to produce an inferential method which gives useful confidence intervals that are as robust as possible. Indeed, this robustness is captured by the fact that Theorem 5 permits the data-generating process (dgp) to change with $n$, as explicitly stated in the Notation section. Thus conclusions of the theorem are valid for a wide variety of sequences of dgps. However, while this construction partly addresses the uniformity critique, it does not achieve "full" uniformity, that is, it does not achieve validity over *all* potential sequences of dgps. However, we should not interpret this as a deficiency, if the potential sequences causing invalidity are thought of as implausible or unlikely (see Giné and Nickl (2010)). Finally, it would be desirable to have a *useful* procedure that is valid under *all* sequences of dgps, but such a procedure does not exist. □

6.2. **Monte Carlo Example: Partially Linear Models.** In this section, we compare the estimation strategies proposed above in the following model:

$$y_i = d_i'\alpha_0 + \tilde{x}_i'\beta_0 + \zeta_i, \quad \zeta_i \sim N(0, \sigma_\zeta^2) \tag{6.53}$$

where the covariates $\tilde{x} \sim N(0, \Sigma)$, $\Sigma_{kj} = (0.5)^{|j-k|}$, and

$$d_i = \tilde{x}_i'\eta_0 + v_i, \quad v_i \sim N(0, \sigma_v^2) \tag{6.54}$$

with $\sigma_\zeta = \sigma_v = 1$, and $\sigma_{\zeta v} = 0$. The dimension $p$ of the covariates $x$ is 200, and the sample size $n$ is 100. We set $\alpha_0 = 1$ and

$$\begin{aligned} \beta_0 &= \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5},\ 0,\ 0,\ 0,\ 0,\ 0,\ 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, \ldots, 0\right)', \\ \eta_0 &= \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \frac{1}{9}, \frac{1}{10},\ 0, \ldots\ldots\ldots\ldots\ldots, 0\right)'. \end{aligned}$$

We set $\lambda$ according to the $X$-dependent rule with $1 - \gamma = .95$. For each repetition we draw new $x$'s, $\zeta$'s and $v$'s.

We summarize the inference performance of these methods in Table 4 which illustrates mean bias, standard deviation, and rejection probabilities of 95% confidence intervals. As we had expected, Lasso and Post-Lasso exhibit a large mean bias which dominates the estimation error and results in poor performance of conventional inference methods. On the other hand, the Indirect Post-Lasso has a small bias relative to estimation error but is substantially more variable than double-Post-Lasso and produces a conservative test, a test with size much smaller than the nominal level. Notably, the double-Post-Lasso provides coverage that is close to the promised 5% level and has the smallest mean bias and standard deviation.

**Partial Linear Model Simulation Results**

| Estimator | Mean Bias | Std. Dev. | rp(0.05) |
|---|---|---|---|
| Lasso | 0.644 | 0.093 | 1.000 |
| Post-Lasso | 0.415 | 0.209 | 0.877 |
| Indirect Post-Lasso | 0.0908 | 0.194 | 0.004 |
| Double selection | -0.0041 | 0.111 | 0.054 |
| Double selection Oracle | 0.0001 | 0.110 | 0.051 |
| Oracle | -0.0003 | 0.100 | 0.044 |

TABLE 4. Results are based on 1000 simulation replications of the partially linear model (6.53) where $p = 200$ and $n = 100$. We report mean bias (Mean Bias), standard deviation (Std. Dev.), and rejection frequency for 5% level tests (rp(.05)) for the four estimators described in Section 7.1.

## 7. EMPIRICAL EXAMPLES.

In this section, we illustrate the performance of sparse methods in two empirical examples. In the first, we revisit the classic Angrist and Krueger (1991)'s instrumental variables estimation of the returns to schooling. In this example, there are many instruments which can potentially

be used in forming the IV estimator and there are concerns about the potential biases and inferential problems introduced from using many instruments. Our results show that sparse methods can be effectively used to alleviate these concerns. The second example concerns the use of $\ell_1$-penalized methods to select control variables for growth regressions in which there are many possible country level controls relative to the number of countries. Using Square-root Lasso to select control variables, we find that there is evidence in favor of the hypothesis of convergence.

7.1. **Angrist and Krueger Example with 1530 instruments.** We consider the Angrist and Krueger (1991) model

$$\begin{aligned} y_{1i} &= \theta_1 y_{2i} + w_i'\gamma + \zeta_i, \quad \mathrm{E}[\zeta_i|w_i, z_i] = 0, \\ y_{2i} &= z_i'\beta + w_i'\delta + v_i, \quad \mathrm{E}[v_i|w_i, z_i] = 0, \end{aligned}$$

where $y_{1i}$ is the log(wage) of individual $i$, $y_{2i}$ denotes education, $w_i$ denotes a vector of control variables, and $z_i$ denotes a vector of instrumental variables that affect education but do not directly affect the wage. The data were drawn from the 1980 U.S. Census and consist of 329,509 men born between 1930 and 1939. In this example, $w_i$ is a set of 510 variables: a constant, 9 year-of-birth dummies, 50 state-of-birth dummies, and 450 state-of-birth $\times$ year-of-birth interactions. As instruments, we use three quarter-of-birth dummies and interactions of these quarter-of-birth dummies with the set of state-of-birth and year-of-birth controls in $w_i$ giving a total of 1530 potential instruments. Angrist and Krueger (1991) discusses the endogeneity of schooling in the wage equation and provides an argument for the validity of $z_i$ as instruments based on compulsory schooling laws and the shape of the life-cycle earnings profile. We refer the interested reader to Angrist and Krueger (1991) for further details. The coefficient of interest is $\theta_1$, which summarizes the causal impact of education on earnings.

There are two basic options for estimating $\theta_1$ that have been used in the literature: one uses just the three basic quarter-of-birth dummies and the other uses 180 instruments corresponding to the three quarter-of-birth dummies and their interactions with the 9 main effects for year-of-birth and 50 main effects for state-of-birth. It is commonly-held that using the set of 180 instruments results in 2SLS estimates of $\theta_1$ that have a substantial bias, while using just the three quarter-of-birth dummies results in an estimator with smaller bias but a large variance; see, e.g., Hansen, Hausman, and Newey (2008). Another approach uses the 180 instruments and the Fuller estimator (Fuller 1977) (FULL) with an adjustment for the use of many instruments. Of course, using sparse methods for the first-stage estimation offers another option that could be used in place of any of the aforementioned approaches.

Table 5 presents estimates of the returns to schooling coefficient using 2SLS and FULL[13] and different sets of instruments. Given knowledge of the construction of the instruments, the first three rows of the table correspond to the natural groupings of the instruments into the

[13]We set the user-defined choice parameter in the Fuller estimator equal to one which results in a higher-order unbiased estimator.

Estimates of the Returns to Schooling in the Angrist-Krueger Data

| Number of Instruments | 2SLS Estimate | 2SLS Std. Error | Fuller Estimate | Fuller Std. Error |
|---|---|---|---|---|
| 3 | 0.1079 | 0.0196 | 0.1087 | 0.0200 |
| 180 | 0.0928 | 0.0097 | 0.1063 | 0.0143 |
| 1530 | 0.0712 | 0.0049 | 0.1019 | 0.0422 |
| | | Lasso - Iterated | | |
| 1 | 0.0862 | 0.0254 | | |
| | | Lasso - 10-Fold Cross-Validation | | |
| 12 | 0.0982 | 0.0137 | 0.0997 | 0.0139 |

Sup-Score/Inverse Lasso 95% Confidence Interval

| Number of Instruments | Center of CI | Quasi Std. Error | Confidence Interval |
|---|---|---|---|
| 3 | .100 | 0.0255 | (0.05,0.15) |
| 180 | .110 | 0.0459 | (0.02,0.20) |
| 1530 | .095 | 0.0689 | (-0.04,0.23) |

TABLE 5. This table reports estimates of the returns-to-schooling parameter in the Angrist and Krueger 1991 data for different sets of instruments. The columns 2SLS and 2SLS Std. Error give the 2SLS point estimate and associated estimated standard error, and the columns Fuller Estimate and Fuller Std. Error give the Fuller point estimate and associated estimated standard error. We report Post-Lasso results based on instruments selected using the plug-in penalty described in Section 3.1 (Lasso - Iterated) and based on instruments selected using a penalty level chosen by 10-Fold Cross-Validation (Lasso - 10-Fold Cross-Validation). For the Lasso-based results, Number of Instruments is the number of instruments selected by Lasso.

three main quarter of birth effects, the three quarter-of-birth dummies and their interactions with the 9 main effects for year-of-birth and 50 main effects for state-of-birth, and the full set of 1530 potential instruments. The remaining two rows give results based on using Lasso to select instruments with penalty level given by the simple plug-in rule in Section 3 or by 10-fold cross-validation. Using the plug-in rule, Lasso selects only the dummy for being born in the fourth quarter; and with the cross-validated penalty level, Lasso selects 12 instruments which include the dummy for being born in the third quarter, the dummy for being born in the fourth quarter, and 10 interaction terms. The reported estimates are obtained using Post-Lasso.

The results in Table 5 are interesting and quite favorable to the idea of using Lasso to do variable selection for instrumental variables. It is first worth noting that with 180 or 1530 instruments, there are modest differences between the 2SLS and FULL point estimates that theory as well as evidence in Hansen, Hausman, and Newey (2008) suggests is likely due to bias induced by overfitting the 2SLS first-stage which may be large relative to precision. In the remaining cases, the 2SLS and FULL estimates are all very close to each other suggesting that this bias is likely not much of a concern. This similarity between the two estimates

is reassuring for the Lasso-based estimates as it suggests that Lasso is working as it should in avoiding overfitting of the first-stage and thus keeping bias of the second-stage estimator relatively small.

For comparing standard errors, it is useful to remember that one can regard Lasso as a way to select variables in a situation in which there is no *a priori* information about which of the set of variables is important; i.e. Lasso does not use the knowledge that the three quarter of birth dummies are the "main" instruments and so is selecting among 1530 *a priori* "equal" instruments. Given this, it is again reassuring that Lasso with the more conservative plug-in penalty selects the dummy for birth in the fourth quarter which is the variable that most cleanly satisfies Angrist and Krueger (1991)'s argument for the validity of the instrument set. With this instrument, we estimate the returns-to-schooling to be .0862 with an estimated standard error of .0254. The best comparison is FULL with 1530 instruments which also does not use any *a priori* information about the relevance of the instruments and estimates the returns-to-schooling as .1019 with a much larger standard error of .0422. One can be less conservative than the plug-in penalty by using cross-validation to choose the penalty level. In this case, 12 instruments are chosen producing a Fuller point estimate (standard error) of .0997 (.0139) or 2SLS point estimate (standard error) of .0982 (.0137). These standard errors are smaller than even the standard errors obtained using information about the likely ordering of the instruments given by using 3 or 180 instruments where FULL has standard errors of .0200 and .0143 respectively. That is, Lasso finds just 12 instruments that contain nearly all information in the first stage and, by keeping the number of instruments small, produces a 2SLS estimate that likely has relatively small bias. We believe that these empirical results are reliable. In particular, we note that the first stage $F$ statistic on the selected 12 instruments is approximately 20; our computational experiments in the previous section employ designs with $F = 10$ and $F = 40$ to show that this method works well for both estimation and inference purposes.

As a final check, we report the 95% confidence interval obtained from the Sup-Score test of Section 5.2 based on the three natural groupings of 3, 180, and 1530 instruments. This test is robust to weak or non-identification and is simple to implement. For the three different sets of instruments, we obtain intervals that are much wider but roughly in line with the intervals discussed above. We note that our preferred method from the simulation section only makes use of the Sup-Score test when no instruments are selected, does a good job at controlling size in the simulation, and is more powerful than the Sup-Score test when the instruments contain signal about the endogenous variable. Using this procedure would lead us to use the much more precise IV-Lasso results.

Overall, these results demonstrate that Lasso instrument selection is feasible and produces sensible and what appear to be relatively high-quality estimates in this application. The results from the Lasso-based IV estimators are similar to those obtained from other leading approaches to estimation and inference with many-instruments and do not require *ex ante*

information about which are the most relevant instruments. Thus, the Lasso-based IV procedures should provide a valuable complement to existing approaches to estimation and inference in the presence of many instruments.

7.2. **Growth Example.** In this section, we consider variable selection in an international economic growth example. We use the Barro and Lee (1994) data consisting of a panel of 138 countries for the period of 1960 to 1985. We consider the national growth rates in GDP per capita as the dependent variable. In our analysis, we consider a model with $p = 62$ covariates which allows for a total of $n = 90$ complete observations. Our goal here is to provide estimates which shed light on the convergence hypothesis discussed below by selecting controls from among these covariates.[14]

One of the central issues in the empirical growth literature is the estimation of the effect of an initial (lagged) level of GDP per capita on the growth rates of GDP per capita. In particular, a key prediction from the classical Solow-Swan-Ramsey growth model is the hypothesis of convergence which states that poorer countries should typically grow faster than richer countries and therefore should tend to catch up with the richer countries over time. This hypothesis implies that the effect of a country's initial level of GDP on its growth rate should be negative. As pointed out in Barro and Sala-i-Martin (1995), this hypothesis is rejected using a simple bivariate regression of growth rates on the initial level of GDP. (In our case, regression yields a statistically insignificant coefficient of .00132.) In order to reconcile the data and the theory, the literature has focused on estimating the effect *conditional* on characteristics of countries. Covariates that describe such characteristics can include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others; see (Barro and Sala-i-Martin 1995). The theory then predicts that the effect of the initial level of GDP on the growth rate should be negative among otherwise similar countries.

Given that the number of covariates we can condition on is comparable to the sample size, covariate selection becomes an important issue in this analysis; see Levine and Renelt (1992), Sala-i-Martin (1997), Sala-i-Martin, Doppelhofer, and Miller (2004). In particular, previous findings came under severe criticisms for relying upon *ad hoc* procedures for covariate selection; see, e.g., Levine and Renelt (1992). Since the number of covariates is high, there is no simple way to resolve the model selection problem using only standard tools. Indeed the number of possible lower-dimensional model is very large, though see Levine and Renelt (1992), Sala-i-Martin (1997) and Sala-i-Martin, Doppelhofer, and Miller (2004) for attempts to search over millions of these models. Here we use $\ell_1$-penalized methods to attempt to resolve this important issue.

We first present results for covariate selection using the different methods discussed in Section 6: (a) a simple Post-Square-root-Lasso method which uses controls selected from applying the

[14] We can compare our results to those obtained in other standard models in the growth literature such as (Barro and Sala-i-Martin 1995, Koenker and Machado 1999).

**Model Selection Results for the International Growth Regressions**
**Real GDP per capita (log) is included in all models**

| Selection Method | Additional Variables Selected |
|---|---|
| Square-root Lasso | Black Market Premium (log) |
| Double selection | Terms of trade shock |
| | Infant Mortality Rate (0-1 age) |
| | Female gross enrollment for secondary education |
| | Percentage of "no schooling" in the female population |
| | Percentage of "higher school attained" in the male population |
| | Average schooling years in the female population over the age of 25 |

TABLE 6. The controls selected by different methods.

Square-root-Lasso to select controls in the regression of growth rates on log-GDP and other controls, and (b) the Post-double-selection method, which uses the controls selected by Square-root-Lasso in the regression of log-GDP on other controls and in the regression of growth rates on other controls. These were all based on Square-root Lasso to avoid the estimation of $\sigma$. We present the model selection results in Table 6.

Square-root Lasso applied to the regression of growth rates on log-GDP and other controls selected only one control, the log of the black market premium which characterizes trade openness. The double selection method selected infant mortality rate, terms of trade shock, and several education variables (female gross enrollment for secondary education, percentage of "no schooling" in the female population, percentage of "higher school attained" in male population, and average schooling years in female population over the age of 25) to forecast log-GDP but no additional controls were selected to forecast growth. We refer the reader to Barro and Lee (1994) and Barro and Sala-i-Martin (1995) for a complete definition and discussion of each of these variables.

We then proceeded to construct confidence intervals for the coefficient on initial GDP based on each set of selected variables. We also report estimates of the effect of initial GDP in a model which uses the set of controls obtained from the double-selection procedure and additionally includes the log of the black market premium. We expressly allow for such amelioration strategy in our formal construction of the estimator. Table 7 shows these results. We find that in all these models the linear regression coefficients on the initial level of GDP are negative. In addition, zero is excluded from the 90% confidence interval in each case. These findings support the hypothesis of (conditional) convergence derived from the classical Solow-Swan-Ramsey growth model. The findings also agree with and thus support the previous findings reported in Barro and Sala-i-Martin (1995) which relied on ad-hoc reasoning for covariate selection.

**Confidence Intervals after Model Selection**
**for the International Growth Regressions**

| | Real GDP per capita (log) | |
|---|---|---|
| Method | Coefficient | 90% Confidence Interval |
| Post Square-root Lasso | −0.0112 | [−0.0219, −0.0007] |
| Post Double selection | −0.0221 | [−0.0437, −0.0005] |
| Post Double selection (+ Black Market Premium) | −0.0302 | [−0.0509, −0.0096] |

TABLE 7. The table above displays the coefficient and a 90% confidence interval associated with each method. The selected models are displayed in Table 6.

## 8. Conclusion

There are many situations in economics where a researcher has access to data with a large number of covariates. In this article, we have presented results for performing analysis of such data by selecting relevant regressors and estimating their coefficients using $\ell_1$-penalization methods. We gave special attention to the instrumental variables model and the partially linear model, both of which are widely used to estimate structural economic effects. Through simulation and empirical examples, we have demonstrated that $\ell_1$ penalization methods may be usefully employed in these models and can complement tools commonly employed by applied researchers.

Of course, there are many avenues for additional research. The use of $\ell_1$-penalization is only one method of performing estimation with high-dimensional data. It will be interesting to consider and understand the behavior of other methods (e.g. Huang, Horowitz, and Ma (2008), Fan and Li (2001), Zhang (2010), Fan and Liao (2011)) for estimating structural economic objects. In addition, extending HDS models and methods to other types of economic models beyond those considered in this article will be interesting. An important problem in economics is the analysis of high-dimensional data in which there are many weak signals within the set of variables considered in which case the sparsity assumption may provide a poor approximation. The sup-score test presented in this article offers one approach to dealing with this problem, but further additional research dealing with this issue seems warranted. It would also be interesting to consider efficient use of high-dimensional data in cases in which scores are not independent across observations which is a much-considered case in economics. Overall, we believe the results in this article provide useful tools for applied economists but that there are still substantial and interesting topics in the use of high-dimensional economic data that warrant further investigation.

## Appendix A. Iterated Estimation of the Noise Level $\sigma$

In the case of Lasso, the penalty levels (3.9) and (3.10) require the practitioner to fill in a value for $\sigma$. Theoretically, any upper bound on $\sigma$ can be used and the standard approach in the literature is

to use the conservative estimate $\bar{\sigma} = \sqrt{\mathrm{Var}_n[y_i]} := \sqrt{\mathbb{E}_n\left[(y_i - \bar{y})^2\right]}$, where $\bar{y} = \mathbb{E}_n[y_i]$. Unfortunately, in various examples we found that this approach leads to overpenalization. Here we briefly discuss iterative procedures to estimate $\sigma$ similar to the ones described in Belloni and Chernozhukov (2011b). Let $I_0$ be a set of regressors that is included in the model. Note that $I_0$ is always non-empty since it will always include the intercept. Let $\bar{\beta}(I_0)$ be the least squares estimator of the coefficients on the covariates associated with $I_0$, and define $\widehat{\sigma}_{I_0} := \sqrt{\mathbb{E}_n[(y_i - x_i'\bar{\beta}(I_0))^2]}$.

An algorithm for estimating $\sigma$ using Lasso is as follows:

**Algorithm 1** (Estimation of $\sigma$ using Lasso iterations)**.** *For a positive number $\psi$, set $\widehat{\sigma}_0 = \psi\widehat{\sigma}_{I_0}$. Set $k = 0$, and specify a small constant $\nu \geqslant 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations. (1) Compute the Lasso estimator $\widehat{\beta}$ based on $\lambda = 2c\widehat{\sigma}_k\Lambda(1-\gamma|X)$.(2) Set $\widehat{\sigma}^2_{k+1} = \widehat{Q}(\widehat{\beta})$. (3) If $|\widehat{\sigma}_{k+1} - \widehat{\sigma}_k| \leqslant \nu$ or $k > K$, report $\widehat{\sigma} = \widehat{\sigma}_{k+1}$; otherwise set $k \leftarrow k+1$ and go to (1).*

Similarly, an algorithm for estimating $\sigma$ using Post-Lasso is as follows:

**Algorithm 2** (Estimation of $\sigma$ using Post-Lasso iterations)**.** *For a positive number $\psi$, set $\widehat{\sigma}^0 = \psi\widehat{\sigma}_{I_0}$. Set $k = 0$, and specify a small constant $\nu \geqslant 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations. (1) Compute the Post-Lasso estimator $\widetilde{\beta}$ based on $\lambda = 2c\widehat{\sigma}_k\Lambda(1-\gamma|X)$. (2) For $\widehat{s} = \|\widetilde{\beta}\|_0 = |\widehat{T}|$ set $\widehat{\sigma}^2_{k+1} = \widehat{Q}(\widetilde{\beta}) \cdot n/(n-\widehat{s})$. (3) If $|\widehat{\sigma}_{k+1} - \widehat{\sigma}_k| \leqslant \nu$ or $k > K$, report $\widehat{\sigma} = \widehat{\sigma}_{k+1}$; otherwise, set $k \leftarrow k+1$ and go to (1).*

**Comment A.1.** We note that we employ the standard degree-of-freedom correction with $\widehat{s} = \|\widetilde{\beta}\|_0 = |\widehat{T}|$ when using Post-Lasso (Algorithm 2). No additional correction is necessary when using Lasso (Algorithm 1) since the Lasso estimate is already sufficiently regularized. We note that the sequence $\widehat{\sigma}_k$, $k \geqslant 2$, produced by Algorithm 1 is monotone and that the estimates $\widehat{\sigma}_k$, $k \geqslant 1$, produced by Algorithm 2 can only assume a finite number of different values. Belloni and Chernozhukov (2011b) and Belloni and Chernozhukov (2011c) provide theoretical analysis for $\psi = 1$. In preliminary simulations with coefficients that were not well separated from zero, we found that $\psi = 0.1$ worked better than $\psi = 1$ by avoiding unnecessary overpenalization in the first iteration. □

## Appendix B. Proof of Theorem 3

Step 1. Recall that $A_i = (f(z_i), w_i')'$ and $d_i = (y_{2i}, w_i')'$ for $i = 1, \ldots, n$. Let $X = [x_1, \ldots, x_n]'$, $A = [A_1, ..., A_n]'$, $D = [d_1, ..., d_n]'$, $W = [w_1, ..., w_n]'$, $f = [f(z_1), ..., f(z_n)]'$, $Y_2 = [y_{21}, ..., y_{2n}]'$, $V = [v_1, ..., v_n]'$, and $\zeta = [\zeta_1, ..., \zeta_n]'$. We have that

$$\sqrt{n}(\widehat{\alpha}^* - \alpha) \quad = \quad [\widehat{A}'D/n]^{-1}\widehat{A}'\zeta/\sqrt{n} = [Q_n + o_P(1)]^{-1}\left(A'\zeta/\sqrt{n} + o_P(1)\right)$$

where by Steps 3 and 4 below:

$$\widehat{A}'D/n = A'D/n + o_P(1) = Q_n + o_P(1) \tag{B.55}$$

$$\widehat{A}'\zeta/\sqrt{n} = A'\zeta/\sqrt{n} + o_P(1). \tag{B.56}$$

Moreover, by the assumption on $\sigma_\zeta$ and $Q_n$, $\mathrm{Var}(A'\zeta/\sqrt{n}) = \sigma_\zeta^2 Q_n$ has eigenvalues bounded away from zero and bounded from above, uniformly in $n$. Therefore, $\sqrt{n}(\widehat{\alpha}^* - \alpha_0) = Q_n^{-1}A'\zeta/\sqrt{n} + o_P(1)$, and $Q_n^{-1}A'\zeta/\sqrt{n}$ is a vector distributed as normal with mean zero and covariance $\sigma_\zeta^2 Q_n^{-1}$. This verifies the main claim of the theorem.

Step 2. This is an auxiliary step where we note that conditions of the theorem imply by Markov inequality:

$$\begin{aligned}
&f'f/n + \mathrm{tr}(W'W/n) = \mathrm{tr}(A'A/n) = \mathrm{tr}(Q_n) \lesssim 1, \\
&\|D'\zeta/n\| \leqslant |V'\zeta/n| + \|A'\zeta/n\| \lesssim_P \sigma_{\zeta v} + 1/\sqrt{n}, \\
&\|A'V/n\|^2 = |f'V/n|^2 + \|W'V/n\|^2 \lesssim_P 1/n, \\
&\|D/\sqrt{n}\| \leqslant \|V/\sqrt{n}\| + \|A/\sqrt{n}\| \lesssim_P 1.
\end{aligned}$$

Step 3. To show (B.55), note that $\widehat{A} - A = (\widehat{f}' - f', 0')'$. Thus,

$$\|\widehat{A}'D/n - A'D/n\| = |(\widehat{f} - f)'Y_2/n| \quad \leqslant \quad \sqrt{(\widehat{f} - f)'(\widehat{f} - f)/n}\sqrt{Y_2'Y_2/n} = o_P(1)$$

since $\sqrt{Y_2'Y_2/n} \lesssim_P 1$ by Markov inequality, and $\sqrt{(\widehat{f} - f)'(\widehat{f} - f)/n} = o_P(1)$ by Theorems 1 or 2. Next, since $f'V/n = o_P(1)$ and $W'V/n = o_P(1)$ by Step 2, note that $A'D/n = A'A/n + o_P(1) = Q_n + o_P(1)$.

Step 4. To show (B.56), note that

$$\begin{aligned}
\|(\widehat{A} - A)'\zeta/\sqrt{n}\| \quad &= |(\widehat{f} - f)'\zeta/\sqrt{n}| = |(X(\widehat{\beta} - \beta_0))'\zeta/\sqrt{n} + (f - X\beta_0)'\zeta/\sqrt{n}| \\
&\leqslant \|X'\zeta/\sqrt{n}\|_\infty \|\widehat{\beta} - \beta_0\|_1 + |(f - X\beta_0)'\zeta/\sqrt{n}| \to_P 0.
\end{aligned}$$

This follows because the first term is of order $\sqrt{\log(p \vee n)}\sqrt{s^2 \log(p \vee n)/n} \to 0$ by conditions of the theorem; the order follows because $\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p \vee n)}$ by (3.11), and $\|\widehat{\beta} - \beta_0\|_1 \lesssim_P \sqrt{[s^2 \log(p \vee n)]/n}$ by Theorems 1 and 2 since $\|\widehat{\beta} - \beta_0\|_1 \leqslant \sqrt{s + \widehat{s}}\|\widehat{\beta} - \beta_0\| \lesssim_P \sqrt{[s^2 \log(p \vee n)]/n}$ under condition SE and $\widehat{s} \lesssim_P s$. On the other hand, the second term converges to zero in probability by Markov inequality, because the expectation of $|(f - X\beta_0)'\zeta/\sqrt{n}|^2$ is of order $\sigma_\zeta^2 c_s^2 \to 0$.

Step 5. This step establishes consistency of the variance estimator. Since $\sigma_\zeta^2$ and the eigenvalues of $Q_n$ are bounded away from zero and from above uniformly in $n$, it suffices to show $\widehat{\sigma}_\zeta^2 - \sigma_\zeta^2 \to_P 0$ and $\widehat{A}'\widehat{A}/n - Q_n \to_P 0$. Indeed, $\widehat{\sigma}_\zeta^2 = \|\zeta - D(\widehat{\alpha}^* - \alpha_0)\|^2/n = \|\zeta\|^2/n + 2\zeta'D(\alpha_0 - \widehat{\alpha}^*)/n + \|D(\alpha_0 - \widehat{\alpha}^*)\|^2/n$ so that $\|\zeta\|^2/n - \sigma_\zeta^2 \to_P 0$ by Chebyshev inequality since $\max_i \mathrm{E}[\zeta_i^4]$ is bounded uniformly in $n$, and the remaining terms converge to zero in probability since $\widehat{\alpha}^* - \alpha_0 \to_P 0$, $\|D'\zeta/n\| \lesssim_P 1$ by Step 2. Next, note that

$$\|\widehat{A}'\widehat{A}/n - A'A/n\| = \|A'(\widehat{A} - A)/n + (\widehat{A} - A)'A/n + (\widehat{A} - A)'(\widehat{A} - A)/n\|$$

which is bounded up to a constant by $(\|\widehat{A} - A\|/\sqrt{n})(\|A\|/\sqrt{n}) + \|\widehat{A} - A\|^2/n \to_P 0$ since $\|\widehat{A} - A\|^2/n = \|\widehat{f} - f\|^2/n = o_P(1)$ by Theorems 1 or 2, and $\|A\|^2/n \lesssim_P 1$ holding by Step 2. □

## Appendix C. Proof of Theorem 4

Step 1. When $a = \alpha_1$ we have that

$$\Lambda_{\alpha_1} = \max_{1 \leqslant j \leqslant p} \frac{n|\mathbb{E}_n[\tilde{\epsilon}_i \tilde{x}_{ij}]|}{\sqrt{\mathbb{E}_n[\tilde{\epsilon}_i^2 \tilde{x}_{ij}^2]}} = \max_{1 \leqslant j \leqslant p} \frac{n|\mathbb{E}_n[\tilde{g}_i \tilde{x}_{ij}]|}{\sqrt{\mathbb{E}_n[\tilde{g}_i^2 \tilde{x}_{ij}^2]}}$$

so claim (1) follows from the definition of quantile and from the continuity of the distribution of $\Lambda_{\alpha_1}$.

Step 2. To establish claim (2), we note that

$$n\mathbb{E}_n[\tilde{g}_i \tilde{x}_{ij}] = n\mathbb{E}_n[g_i \tilde{x}_{ij}] = \sqrt{n}\mathcal{N}_j\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]} = \sqrt{n}\mathcal{N}_j,$$

where $\mathcal{N}_j \sim N(0,1)$ for each $j$. Since for $\widehat{\mu}_g = (\mathbb{E}_n[w_i w_i'])^{-1}\mathbb{E}_n[w_i g_i]$ we have $\|\widehat{\mu}_g\| \lesssim_P 1/\sqrt{n}$ by the assumed boundedness of $\|(\mathbb{E}_n[w_i w_i'])^{-1}\|$ and boundedness of $\|w_i\|$, we conclude that $\max_{i\leqslant n}|w_i'\widehat{\mu}_g| \lesssim_P 1/\sqrt{n}$, so that

$$|\sqrt{\mathbb{E}_n[\tilde{g}_i^2\tilde{x}_{ij}^2]} - \sqrt{\mathbb{E}_n[g_i^2\tilde{x}_{ij}^2]}| \leqslant \sqrt{\mathbb{E}_n[(w_i'\widehat{\mu}_g)^2\tilde{x}_{ij}^2]} \lesssim_P n^{-1/2}\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]},$$

uniformly in $j \in \{1,...,p\}$, using the triangular inequality and the decomposition $\tilde{g}_i = g_i - w_i'\widehat{\mu}_g$. Moreover, using the Bernstein-type inequality in Lemma 5.15 of van de Geer (2000), we can conclude that

$$|\mathbb{E}_n[g_i^2\tilde{x}_{ij}^2] - \mathbb{E}_n[\tilde{x}_{ij}^2]| \lesssim_P \sqrt{(\log p)/n},$$

uniformly in $j \in \{1,...,p\}$. Hence since $\mathbb{E}_n[\tilde{x}_{ij}^2] = 1$ by the normalization assumption, we conclude that with probability approaching 1,

$$\Lambda_{\alpha_1} \leqslant \max_{1\leqslant j\leqslant p} cn|\mathbb{E}_n[g_i\tilde{x}_{ij}^2]|/\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]} = \max_{1\leqslant j\leqslant p} c\sqrt{n}|\mathcal{N}_j|$$

and the claim (2) follows by the union bound and standard tail properties of $N(0,1)$.

Step 3. To show claim (3) we note that using triangular and other elementary inequalities:

$$\begin{aligned}\Lambda_a &= \max_{1\leqslant j\leqslant p}\left|\frac{n|\mathbb{E}_n[(\tilde{\epsilon}_i - (a-\alpha_1)\tilde{y}_{2i})\tilde{x}_{ij}]}{\sqrt{\mathbb{E}_n[(\tilde{\epsilon}_i - (a-\alpha_1)\tilde{y}_{2i})^2\tilde{x}_{ij}^2]}}\right| \\ &\geqslant \max_{1\leqslant j\leqslant p}\left|\frac{|a-\alpha_1|n|\mathbb{E}_n[\tilde{y}_{2i}\tilde{x}_{ij}]|}{\sqrt{\mathbb{E}_n[\tilde{\epsilon}_i^2\tilde{x}_{ij}^2]} + |a-\alpha_1|\sqrt{\mathbb{E}_n[\tilde{y}_{2i}^2\tilde{x}_{ij}^2]}}\right| - \Lambda_{\alpha_1}\end{aligned}$$

The first term is bounded below by, with probability approaching 1,

$$c^{-1}\max_{1\leqslant j\leqslant p}\frac{|a-\alpha_1||n\mathbb{E}_n[\tilde{y}_{2i}\tilde{x}_{ij}]|}{\sigma_\zeta\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]} + |a-\alpha_1|\sqrt{\mathbb{E}_n[\tilde{y}_{2i}^2\tilde{x}_{ij}^2]}},$$

by Step 2 for some $c > 1$, and $\Lambda_{\alpha_1} \lesssim_P \sqrt{n\log p}$ by Step 2. Hence for any constant $C$, with probability converging to 1, $\Lambda_a - C\sqrt{n\log p} \to +\infty$, so that Claim (3) immediately follows, since by Step 2 $\Lambda(1-\gamma|X,W) \lesssim \Lambda(1-\gamma) \lesssim \sqrt{n\log p}$, since $\gamma \in (0,1)$ is fixed by assumption. □

## Appendix D. Proof of Theorem 5

Let me prepare some notation. I will use the standard matrix notation, namely $Y_1 = [y_{11},...,y_{1n}]'$, $X = [x_1,...,x_n]'$, $D = [d_1,...,d_n]'$, $V = [v_1,...,v_n]'$, $\zeta = [\zeta_1,...,\zeta_n]'$, $m = [m_1,...,m_n]'$ for $m_i = m(z_i)$, $R_m = [r_{m1},...,r_{mn}]'$, $g = [g_1,...,g_n]'$ for $g_i = g(z_i)$, $R_g = [r_{g1},...,r_{gn}]'$, and so on. Let $\phi_{\min}(\widehat{s}) = \phi_{\min}(\widehat{s})[\mathbb{E}_n[x_i x_i']]$. For $A \subset \{1,...,p\}$, let $X[A] = \{X_j, j \in A\}$, where $\{X_j, j = 1,...,p\}$ are the columns of $X$. Let

$$\mathcal{P}_A = X[A](X[A]'X[A])^- X[A]'$$

be the projection operator sending vectors in $\mathbb{R}^n$ onto span$[X[A]]$, and let $\mathcal{M}_A = \mathrm{I}_n - \mathcal{P}_A$ be the projection onto the subspace that is orthogonal to span$[X[A]]$. For a vector $Z \in \mathbb{R}^n$, let

$$\tilde{\beta}_Z(A) := \arg\min_{b\in\mathbb{R}^p}\|Z - X'b\|^2 : \ b_j = 0, \ \forall j \notin A,$$

be the coefficient of linear projection of $Z$ onto span$[X[A]]$. If $A = \varnothing$, interpret $\mathcal{P}_A = 0_n$, and $\tilde{\beta}_Z = 0_p$.

Step 1.(Main) Write $\check{\alpha} = \left[D'\mathcal{M}_{\widehat{I}}D/n\right]^{-1}\left[D'\mathcal{M}_{\widehat{I}}Y_1/n\right]$ so that

$$\sqrt{n}(\check{\alpha}-\alpha_0) = \left[D'\mathcal{M}_{\widehat{I}}D/n\right]^{-1}\left[D'\mathcal{M}_{\widehat{I}}(g+\zeta)/\sqrt{n}\right] =: ii^{-1}\cdot i.$$

By Steps 2 and 3, $ii = V'V/n + o_P(1)$ and $i = V'\zeta/\sqrt{n} + o_P(1)$. Since $V'V/n = \sigma_v^2 + o_P(1)$ by Chebyshev inequality, and $\sigma_\zeta^2$ and $\sigma_v^2$ are bounded from above and away from zero by assumption, and

$$V'\zeta/\sqrt{n} = [\sigma_\zeta\sqrt{V'V/n}]N(0,1)$$

conclude that

$$\sigma_\zeta^{-1}(V'V/n)^{1/2}\sqrt{n}(\check{\alpha}-\alpha_0) = N(0,1) + o_P(1).$$

Step 2. (Behavior of $i$.) Decompose

$$i = V'\zeta/\sqrt{n} + m'\mathcal{M}_{\widehat{I}}g/\sqrt{n} + m'\mathcal{M}_{\widehat{I}}\zeta/\sqrt{n} + V'\mathcal{M}_{\widehat{I}}g/\sqrt{n} - V'\mathcal{P}_{\widehat{I}}\zeta/\sqrt{n}. \tag{D.57}$$
$$\underset{=:i_a}{\phantom{x}} \qquad \underset{=:i_b}{\phantom{x}} \qquad \underset{=:i_c}{\phantom{x}} \qquad \underset{=:i_d}{\phantom{x}}$$

First, note that by Steps 4 and 5 and by the growth condition $s^2\log^2(p\vee n) = o(n)$

$$|i_a| \leqslant \sqrt{n}\|m'\mathcal{M}_{\widehat{I}}/\sqrt{n}\|\|g'\mathcal{M}_{\widehat{I}}/\sqrt{n}\| \lesssim_P \sqrt{n}\sqrt{[s\log(p\vee n)]^2/n^2} = o_P(1).$$

Second, using decomposition $m = X\beta_{m0} + R_m$, bound

$$|i_b| \leqslant |R_m'\zeta/\sqrt{n}| + |(\tilde{\beta}_m(\widehat{I}) - \beta_{m0})'X'\zeta/\sqrt{n}| \lesssim_P \sqrt{[s\log(p\vee n)]^2/n} = o_P(1),$$

where $|R_m'\zeta/\sqrt{n}| \lesssim_P \sqrt{R_m'R_m/n} \lesssim \sqrt{s/n}$ by Chebyshev inequality and by assumption ASTE, and

$$|(\tilde{\beta}_m(\widehat{I}) - \beta_{m0})'X'\zeta/\sqrt{n}| \leqslant \|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\|_1\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{[s^2\log(p\vee n)]/n}\sqrt{\log(p\vee n)},$$

$\|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\|_1 \leqslant \sqrt{\widehat{s}}\|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\| \lesssim_P \sqrt{[s^2\log(p\vee n)]/n}$ by Step 4, using that $\widehat{s} \lesssim_P s$ by Theorem 2, $\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)}$ by the Gaussian maximal inequality (3.11) and normalization condition on $X$. Third, using similar reasoning, decomposition $g = X\beta_{g0} + R_g$, and Step 5, conclude

$$|i_c| \leqslant |R_g'\zeta| + |(\tilde{\beta}_g(\widehat{I}) - \beta_{g0})'X'V/\sqrt{n}| \lesssim_P \sqrt{[s\log(p\vee n)]^2/n} = o_P(1).$$

Fourth, using that $\widehat{s} \lesssim_P s$ by Theorem 2 so that $1/\phi_{\min}(\widehat{s}) \lesssim_P 1$ by condition SE, conclude,

$$|i_d| \leqslant |\tilde{\beta}_V(\widehat{I})'X'\zeta/\sqrt{n}| \leqslant \|\tilde{\beta}_V(\widehat{I})\|_1\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{n}\sqrt{[s\log(p\vee n)]^2/n^2} = o_P(1),$$

since $\|\tilde{\beta}_V(\widehat{I})\|_1 \leqslant \sqrt{\widehat{s}}\|\tilde{\beta}_V(\widehat{I})\| \leqslant \sqrt{\widehat{s}}\|(X[\widehat{I}]'X[\widehat{I}])^{-1}X[\widehat{I}]'V/n\| \leqslant \sqrt{\widehat{s}}\phi_{\min}^{-1}(\widehat{s})\ \sqrt{\widehat{s}}\|X'V/\sqrt{n}\|_\infty/\sqrt{n} \lesssim_P s\sqrt{[\log(p\vee n)]/n}$.

Step 3. (Behavior of $ii$.) Decompose

$$ii = (m+V)'\mathcal{M}_{\widehat{I}}(m+V)/n = V'V/n + m'\mathcal{M}_{\widehat{I}}m/n + 2m'\mathcal{M}_{\widehat{I}}V/n - V'\mathcal{P}_{\widehat{I}}V/n.$$
$$\underset{=:ii_a}{\phantom{x}} \qquad \underset{=:ii_b}{\phantom{x}} \qquad \underset{=:ii_c}{\phantom{x}}$$

Then $|ii_a| \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ by Step 4, $|ii_b| \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ by reasoning similar to deriving the bound for $|i_b|$, and $|ii_c| \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ by reasoning similar to deriving the bound for $|i_d|$.

Step 4. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}}m\|$ and related quantities.) Observe that

$$\sqrt{[s\log(p\vee n)]/n} \underset{(1)}{\gtrsim_P} \|\mathcal{M}_{\widehat{I}_1}m/\sqrt{n}\| \underset{(2)}{\gtrsim_P} \|\mathcal{M}_{\widehat{I}}m/\sqrt{n}\| \underset{(3)}{\gtrsim_P} |\|X(\tilde{\beta}_m(\widehat{I}) - \beta_{m0})/\sqrt{n}\| - \|R_m/\sqrt{n}\||$$

where inequality (1) holds since by Theorem 2 $\|\mathcal{M}_{\widehat{I}_1} m/\sqrt{n}\| \leqslant \|(X\tilde{\beta}_D(\widehat{I}_1) - m)/\sqrt{n}\| \lesssim_P \sqrt{[s\log(p\vee n)]/n}$, (2) holds by $\widehat{I}_1 \subseteq \widehat{I}$, and (3) by the triangle inequality. Since $\|R_m/\sqrt{n}\| \lesssim \sqrt{s/n}$ by assumption ASTE, conclude that wp $\to 1$,

$$\begin{aligned}\sqrt{[s\log(p\vee n)]/n} &\gtrsim_P \|X(\tilde{\beta}_m(\widehat{I}) - \beta_{m0})/\sqrt{n}\| \\ &\geqslant \sqrt{\phi_{\min}(\widehat{s})}\|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\| \gtrsim_P \|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\|,\end{aligned}$$

since $\widehat{s} \lesssim_P s$ by Theorem 2 so that $1/\phi_{\min}(\widehat{s}) \lesssim_P 1$ by condition SE.

Step 5. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}} g\|$ and related quantities.) Observe that

$$\begin{aligned}\sqrt{[s\log(p\vee n)]/n} &\underset{(1)}{\gtrsim_P} \|\mathcal{M}_{\widehat{I}_2}(\alpha_0 m + g)/\sqrt{n}\| \\ &\underset{(2)}{\gtrsim_P} \|\mathcal{M}_{\widehat{I}}(\alpha_0 m + g)/\sqrt{n}\| \underset{(3)}{\gtrsim_P} |\|\mathcal{M}_{\widehat{I}} g/\sqrt{n}\| - \|\mathcal{M}_{\widehat{I}}\alpha_0 m/\sqrt{n}\||\end{aligned}$$

where inequality (1) holds since by Theorem 2 $\|\mathcal{M}_{\widehat{I}_2}(\alpha_0 m + g)/\sqrt{n}\| \leqslant \|(X\tilde{\beta}_{Y_1}(\widehat{I}_2) - \alpha_0 m - g)/\sqrt{n}\| \lesssim_P \sqrt{[s\log(p\vee n)]/n}$, (2) holds by $\widehat{I}_2 \subseteq \widehat{I}$, and (3) by the triangle inequality. Since $\|\alpha_0\|$ is bounded uniformly in $n$ by assumption, by Step 4, $\|\mathcal{M}_{\widehat{I}}\alpha_0 m/\sqrt{n}\| \lesssim_P \sqrt{[s\log(p\vee n)]/n}$. Hence conclude that

$$\sqrt{[s\log(p\vee n)]/n} \gtrsim_P \|\mathcal{M}_{\widehat{I}} g/\sqrt{n}\| \geqslant |\|X(\tilde{\beta}_g(\widehat{I}) - \beta_{g0})/\sqrt{n}\| - \|R_g/\sqrt{n}\||$$

where $\|R_g/\sqrt{n}\| \lesssim \sqrt{s/n}$ by condition ASTE. Then conclude similarly to Step 4 that wp $\to 1$,

$$\sqrt{[s\log(p\vee n)]/n} \gtrsim_P \|X(\tilde{\beta}_g(\widehat{I}) - \beta_{g0})/\sqrt{n}\| \geqslant \sqrt{\phi_{\min}(\widehat{s})}\|\tilde{\beta}_g(\widehat{I}) - \beta_{g0}\| \gtrsim_P \|\tilde{\beta}_g(\widehat{I}) - \beta_{g0}\|.$$

Step 6. (Variance Estimation.) Since $\widehat{s} \lesssim_P s = o(n)$, $(n - \widehat{s} - 1)/n = 1 + o_P(1)$. Hence consider

$$\widehat{\sigma}^2_\zeta = \|(Y_1 - \check{\alpha}D)'\mathcal{M}_{\widehat{I}}\|^2/n = \|(\zeta + (\alpha_0 - \check{\alpha})'D + g)'\mathcal{M}_{\widehat{I}}\|^2/n.$$

Then by Steps 1, 3, and 5

$$|\widehat{\sigma} - \|\zeta'\mathcal{M}_{\widehat{I}}\|/\sqrt{n}\| \leqslant \|g'\mathcal{M}_{\widehat{I}}\|/\sqrt{n} + \|\check{\alpha} - \alpha_0\|\|D'\mathcal{M}_{\widehat{I}}\|/\sqrt{n} \lesssim_P \sqrt{[s\log(p\vee n)]/n} + n^{-1/2} = o_P(1).$$

Moreover,

$$\|\zeta'\mathcal{M}_{\widehat{I}}\|^2/n = \zeta'\zeta/n - \zeta'\mathcal{P}_{\widehat{I}}\zeta/n = \sigma^2_\zeta + o_P(1),$$

where $\zeta'\zeta/n = \sigma^2_\zeta + O_P(n^{-1/2})$ by Chebyshev inequality and $\zeta'\mathcal{P}_{\widehat{I}}\zeta/n \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ by the argument similar to that used to bound $|i_d|$. □

## References


Akaike, H. (1974): "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Anderson, T. W., and H. Rubin (1949): "Estimation of the Parameters of Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.

Angrist, J., V. Chernozhukov, and I. Fernandez-Val (2006): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74(2), 539–563.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

——— (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13(2), 225–235.

Barro, R. J., and J.-W. Lee (1994): "Data set for a panel of 139 countries," *NBER,*.

Barro, R. J., and X. Sala-i-Martin (1995): *Economic Growth*. McGraw-Hill, New York.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 63, 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Preprint, ArXiv.

Belloni, A., and V. Chernozhukov (2011a): "$\ell_1$-Penalized Quantile Regression for High Dimensional Sparse Models," *Annals of Statistics*, 39(1), 82–130.

——— (2011b): "High Dimensional Sparse Econometric Models: An Introduction," *Inverse problems and high dimensional estimation - Stats in the Château summer school in econometrics and statistics, 2009, Springer Lecture Notes in Statistics - Proceedings*, pp. 121–156.

——— (2011c): "Least Squares After Model Selection in High-dimensional Sparse Models," *forthcoming Bernoulli*.

Belloni, A., V. Chernozhukov, and C. Hansen (2010): "LASSO Methods for Gaussian Instrumental Variables Models," Preprint, ArXiv.

——— (2011): "Estimation of Treatment Effects with High-Dimensional Controls," Preprint, ArXiv.

Belloni, A., V. Chernozhukov, and L. Wang (2010): "Square-Root-LASSO: Pivotal Recovery of Nonparametric Regression Functions via Conic Programming," Preprint, ArXiv.

——— (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98(4), 791–806.

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

Candes, E., and T. Tao (2007): "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Statist.*, 35(6), 2313–2351.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5559–5632.

Chen, X., D. Ge, Z. Wang, and Y. Ye (2011): "Complexity of Unconstrained $L_2$-$L_p$ Minimization," *Preprint, ArXiv*.

Chernozhukov, V. (2009): "High-Dimensional Sparse Econometric Models," (Lecture notes) Stats in the Château,.

Chernozhukov, V., and C. Hansen (2008a): "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142, 379–398.

——— (2008b): "The Reduced Form: A Simple Approach to Inference with Weak Instruments," *Economics Letters*, 100, 68–71.

Fan, J., S. Guo, and N. Hao (2011): "Variance estimation using refitted cross-validation in ultrahigh dimensional regression," *forthcoming Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Fan, J., and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of American Statistical Association*, 96(456), 1348–1360.

Fan, J., and Y. Liao (2011): "Ultra-High Dimensional Covariate Selection with Endogenous Regressors," Preprint, Princeton University.

Fuller, W. A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–954.

Gautier, E., and A. Tsybakov (2011): "High-dimensional Instrumental Variables Rergession and Confidence Sets," Preprint, ArXiv.

Ge, D., X. Jiang, and Y. Ye (2011): "A Note on Complexity of $L_p$ Minimization," *to appear Mathematical Programming*.

Giné, E., and R. Nickl (2010): "Confidence bands in density estimation," *Ann. Statist.*, 38(2), 1122–1170.

Hahn, J., J. A. Hausman, and G. M. Kuersteiner (2004): "Estimation with Weak Instruments: Accuracy of Higher-order Bias and MSE Approximations," *Econometrics Journal*, 7(1), 272–306.

Hansen, B. E. (2005): "Challenges for Econometric Model Selection," *Econometric Theory*, 21, 60–68.

Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26, 398–422.

Heckman, J., R. LaLonde, and J. Smith (1999): "The economics and econometrics of active labor market programs," *Handbook of labor economics*, 3, 1865–2097.

Huang, J., J. L. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36(2), 587613.

Imbens, G. (2004): "Nonparamtric Estimation of Average Treatment Effects under Exogeneity: a Review," *Rev. Econ. Stat.*, 86(1), 4–29.

Kato, K. (2011): "Group Lasso for high dimensional sparse quantile regression models," Preprint, ArXiv.

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46(1), 33–50.

Koenker, R., and J. Machado (1999): "Goodness of fit and related inference process for quantile regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Ledoux, M., and M. Talagrand (1991): *Probability in Banach Spaces (Isoperimetry and processes)*. Ergebnisse der Mathematik undihrer Grenzgebiete, Springer-Verlag.

Levine, R., and D. Renelt (1992): "A Sensitivity Analysis of Cross-Country Growth Regressions," *The American Economic Review*, 82(4), 942–963.

Meinshausen, N., and B. Yu (2009): "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, 37(1), 2246–2270.

Natarajan, B. K. (1995): "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, 24, 227–234.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Potscher, B. (2009): "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhya*, 71-A, 1–18.

Sala-i-Martin, X. (1997): "I Just Ran Two Million Regressions," *The American Economic Review*, 87(2), 178–183.

Sala-i-Martin, X., G. Doppelhofer, and R. I. Miller (2004): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *The American Economic Review*, 94(4), 813–835.

Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, 6(2), 461–464.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Tibshirani, R. (1996): "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

van de Geer, S. A. (2000): *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36(2), 614–645.

Zhang, C.-H. (2010): "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, 38(2), 894–942.