S.1. Simulations Continued. The simulation setting in this section is the same as in Section 9. We first describe the reason for using the null situation $\beta_0 = 0_p$ in the model. If $\beta_0$ is an arbitrary non-zero vector, then, for fixed covariates, $X_i Y_i$ cannot be identically distributed and hence only (asymptotically) conservative inference is possible. In simulations this conservativeness confounds with the simultaneity so that the coverage becomes close to 1 (if not 1).

In the main manuscript, we have shown plots comparing our method with Berk et al. (2013) and selective inference. We label our confidence region $\hat{R}_{n,M}^\dagger (12)$ as “UPoSI,” the projected confidence region $\hat{B}_{n,M}^\dagger (28)$ as “UPoSIBox”, and Berk et al. (2013) as “PoSI.” Tables 1, 2, and 3 show exact numbers for the comparison of our method with Berk et al. (2013). Note that size of each dot in the row plot of Figure 9 indicates the proportion of confidence regions of that volume among same-sized models. In Setting A and B, the confidence region volumes of same-sized models are the same. In Setting C, volumes of confidence regions of Berk and PoSI Box enlarge (hence smaller log(Vol)/|M|) if the last covariate is included. Tables 4 and 5 show the numbers for the comparison of our method with selective inference when the selection procedure is forward stepwise and LARS, respectively.

Sample splitting is a simple procedure that provides valid inference after selection as discussed in Section 1.3. We stress here that this is valid only for independent observations and that the model selected in the first split half could be different from the one selected in the full data. The comparison results with $n = 1000, p = 500$ and selection methods forward stepwise, LARS and BIC are summarized in Figure S.1. For sample splitting we have used the Bonferroni correction to obtain simultaneous inference for all coefficients in a model. Table 6 shows the comparison of our method with sample splitting.

---

*Supported in part by NSF Grants DMS-10-07657, DMS-1310795 and NSF567018.
†Supported in part by NSF Grant DMS-14-06563.

MSC 2010 subject classifications: 62J05, 62F40, 62F25, 62F12

Keywords and phrases: Simultaneous Inference, Multiplier Bootstrap, Uniform Consistency, High-dimensional Linear Regression, Concentration Inequalities, Orlicz Norms, Model Selection
Comparison of UPoSI with PoSI of Berk et al. (2013) in Setting A. The Gram matrix being identity implies that the coverage of Berk et al. (2013) for all model sizes is the same as the overall models. For this reason all the rows have the same values. The simulated data have size $n = 200, p = 15$, and the results are based on 100 Monte Carlo replications.

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Coverage of PoSI</th>
<th>log(Vol)/M for PoSI</th>
<th>Coverage of UPoSI</th>
<th>log(Vol)/M for UPoSI</th>
<th>Coverage of UPoSI Box</th>
<th>log(Vol)/M for UPoSI Box</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>2</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>3</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>4</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>5</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>6</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>7</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>8</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>9</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>10</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>11</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>12</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>13</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>14</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>15</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
<tr>
<td>Over all models</td>
<td>0.990</td>
<td>−0.798</td>
<td>0.970</td>
<td>−0.892</td>
<td>0.970</td>
<td>−0.892</td>
</tr>
</tbody>
</table>

Comparison of UPoSI with PoSI of Berk et al. (2013) in Setting B for $n = 200, p = 15$, based on 100 Monte Carlo replications.

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Coverage of PoSI</th>
<th>log(Vol)/M for PoSI</th>
<th>Coverage of UPoSI</th>
<th>log(Vol)/M for UPoSI</th>
<th>Coverage of UPoSI Box</th>
<th>log(Vol)/M for UPoSI Box</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>−0.653</td>
<td>0.950</td>
<td>−0.864</td>
<td>0.950</td>
<td>−0.864</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>−0.651</td>
<td>0.950</td>
<td>−0.862</td>
<td>0.970</td>
<td>−0.800</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>−0.649</td>
<td>0.950</td>
<td>−0.860</td>
<td>0.990</td>
<td>−0.731</td>
</tr>
<tr>
<td>4</td>
<td>0.990</td>
<td>−0.646</td>
<td>0.950</td>
<td>−0.858</td>
<td>0.990</td>
<td>−0.657</td>
</tr>
<tr>
<td>5</td>
<td>0.990</td>
<td>−0.643</td>
<td>0.950</td>
<td>−0.855</td>
<td>1.000</td>
<td>−0.577</td>
</tr>
<tr>
<td>6</td>
<td>0.990</td>
<td>−0.640</td>
<td>0.950</td>
<td>−0.852</td>
<td>1.000</td>
<td>−0.490</td>
</tr>
<tr>
<td>7</td>
<td>0.980</td>
<td>−0.636</td>
<td>0.950</td>
<td>−0.849</td>
<td>1.000</td>
<td>−0.394</td>
</tr>
<tr>
<td>8</td>
<td>0.980</td>
<td>−0.631</td>
<td>0.950</td>
<td>−0.845</td>
<td>1.000</td>
<td>−0.289</td>
</tr>
<tr>
<td>9</td>
<td>0.980</td>
<td>−0.624</td>
<td>0.950</td>
<td>−0.841</td>
<td>1.000</td>
<td>−0.171</td>
</tr>
<tr>
<td>10</td>
<td>0.980</td>
<td>−0.617</td>
<td>0.950</td>
<td>−0.836</td>
<td>1.000</td>
<td>−0.038</td>
</tr>
<tr>
<td>11</td>
<td>0.980</td>
<td>−0.606</td>
<td>0.950</td>
<td>−0.830</td>
<td>1.000</td>
<td>0.116</td>
</tr>
<tr>
<td>12</td>
<td>0.980</td>
<td>−0.592</td>
<td>0.950</td>
<td>−0.823</td>
<td>1.000</td>
<td>0.299</td>
</tr>
<tr>
<td>13</td>
<td>0.980</td>
<td>−0.572</td>
<td>0.950</td>
<td>−0.814</td>
<td>1.000</td>
<td>0.522</td>
</tr>
<tr>
<td>14</td>
<td>0.990</td>
<td>−0.540</td>
<td>0.950</td>
<td>−0.801</td>
<td>1.000</td>
<td>0.810</td>
</tr>
<tr>
<td>15</td>
<td>0.990</td>
<td>−0.481</td>
<td>0.950</td>
<td>−0.782</td>
<td>1.000</td>
<td>1.215</td>
</tr>
<tr>
<td>Over all models</td>
<td>0.980</td>
<td>−0.631</td>
<td>0.950</td>
<td>−0.845</td>
<td>0.950</td>
<td>−0.289</td>
</tr>
</tbody>
</table>

S.2. Implied Regions based on t-Statistics. We now show a “trivial” way to obtain other confidence regions from the proposed ones. They provide computationally cheap upper bounds on the PoSI constant of Berk et al. (2013) and, in contrast to their Scheffé-based upper bound, accounts for the specific design structure. To start, define for any random
Table 3
Comparison of UPoSI with PoSI of Berk et al. (2013) in Setting C for \( n = 200, p = 15 \), based on 100 Monte Carlo replications.

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Coverage for PoSI</th>
<th>Coverage for UPoSI</th>
<th>Coverage for PoSI Box</th>
<th>Coverage for UPoSI Box</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coverage log(|M|)</td>
<td>Coverage log(|M|)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>2</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>3</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>4</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>5</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>6</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>7</td>
<td>0.990</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
<tr>
<td>8</td>
<td>0.970</td>
<td>-0.688</td>
<td>0.980</td>
<td>-0.856</td>
</tr>
<tr>
<td>9</td>
<td>0.970</td>
<td>-0.686</td>
<td>0.980</td>
<td>-0.855</td>
</tr>
<tr>
<td>10</td>
<td>0.980</td>
<td>-0.684</td>
<td>0.980</td>
<td>-0.853</td>
</tr>
<tr>
<td>11</td>
<td>0.980</td>
<td>-0.682</td>
<td>0.980</td>
<td>-0.852</td>
</tr>
<tr>
<td>12</td>
<td>0.980</td>
<td>-0.680</td>
<td>0.980</td>
<td>-0.850</td>
</tr>
<tr>
<td>13</td>
<td>0.980</td>
<td>-0.677</td>
<td>0.980</td>
<td>-0.849</td>
</tr>
<tr>
<td>14</td>
<td>0.980</td>
<td>-0.674</td>
<td>0.980</td>
<td>-0.847</td>
</tr>
<tr>
<td>15</td>
<td>1.000</td>
<td>-0.671</td>
<td>0.980</td>
<td>-0.846</td>
</tr>
<tr>
<td>Over all models</td>
<td>0.970</td>
<td>-0.727</td>
<td>0.980</td>
<td>-0.892</td>
</tr>
</tbody>
</table>

The arguments above yield that for any random model \( \hat{M} \in \mathcal{M}_p(p) \) and \( \theta \in \mathbb{R}^{\|\hat{M}\|} \) a “max-\( |t| \)” statistic:

\[
F_{n,\hat{M}}(\theta) := \max_{j \notin \hat{M}} \left| \sqrt{n} (\hat{\beta}_{n,\hat{M}}(j) - \theta(j)) / \hat{\sigma}_{n,\hat{M}}(j) \right|,
\]

where \( \hat{\sigma}_{n,\hat{M}}(j) \) represents an estimator of the standard deviation of \( \sqrt{n}(\hat{\beta}_{n,\hat{M}}(j) - \beta_{n,\hat{M}}(j)) \). We know from Theorem 4.1 that \( \mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}) \geq 1 - \alpha \). Thus with probability at least \( 1 - \alpha \), we have

\[
F_{n,\hat{M}}(\beta_{n,\hat{M}}) \leq \sup_{\theta \notin \hat{\mathcal{R}}_{n,\hat{M}}} F_{n,\hat{M}}(\theta) =: C_{\hat{M}}(\alpha).
\]

Recall that \( \hat{\mathcal{R}}_{n,\hat{M}} \) depends on \( \alpha \). Similarly one can define \( C_{\hat{M}}^\dagger(\alpha) \) by replacing \( \hat{\mathcal{R}}_{n,\hat{M}} \) with \( \hat{\mathcal{R}}_{n,\hat{M}}^\dagger \). Based on the quantities \( C_{\hat{M}}(\alpha) \) and \( C_{\hat{M}}^\dagger(\alpha) \), we get the following confidence regions that are rectangular in shape:

\[
\hat{\mathcal{I}}_{n,\hat{M}} := \{ \theta \in \mathbb{R}^{\|\hat{M}\|} : F_{n,\hat{M}}(\theta) \leq C_{\hat{M}}(\alpha) \}, \quad \text{and} \quad \hat{\mathcal{I}}_{n,\hat{M}}^\dagger := \{ \theta \in \mathbb{R}^{\|\hat{M}\|} : F_{n,\hat{M}}(\theta) \leq C_{\hat{M}}^\dagger(\alpha) \}.
\]

The arguments above yield that for any random model \( \hat{M} \)

\[
\mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{I}}_{n,\hat{M}}) \geq 1 - \alpha \quad \text{if} \quad \mathbb{P}(\hat{M} \in \mathcal{M}_p(p)) = 1,
\]

\[
\lim_{n \to \infty} \mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{I}}_{n,\hat{M}}^\dagger) \geq 1 - \alpha \quad \text{if} \quad \mathbb{P}(\hat{M} \in \mathcal{M}_p(k)) \to 1,
\]

when \( k \) satisfies \((A1)(k)\). The guarantees (E.1) hold for any function \( F_{n,\hat{M}}(\theta) \) and need not restrict to the maximum form above. The maximum statistic above is similar the “max-\( |t| \)” statistic used in Berk et al. (2013) as described in Section 6. Hence the quantities \( C_{\hat{M}}(\alpha) \)
<table>
<thead>
<tr>
<th>Step→</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage of SelInf</td>
<td>0.870</td>
<td>0.760</td>
<td>0.700</td>
<td>0.650</td>
<td>0.600</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>-0.652</td>
<td>-0.190</td>
<td>0.089</td>
<td>0.284</td>
</tr>
<tr>
<td>Prop. of Infinite Volume</td>
<td>0.100</td>
<td>0.240</td>
<td>0.430</td>
<td>0.590</td>
<td>0.740</td>
</tr>
<tr>
<td>Coverage of UPoSI Box</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>of UPoSI Box</td>
<td>-1.396</td>
<td>-1.396</td>
<td>-1.396</td>
</tr>
<tr>
<td>Coverage of UPoSI</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>of UPoSI</td>
<td>-1.396</td>
<td>-1.396</td>
<td>-1.396</td>
</tr>
<tr>
<td>Setting B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage of SelInf</td>
<td>0.850</td>
<td>0.790</td>
<td>0.740</td>
<td>0.640</td>
<td>0.540</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>-0.681</td>
<td>-0.161</td>
<td>0.087</td>
<td>0.281</td>
</tr>
<tr>
<td>Prop. of Infinite Volume</td>
<td>0.120</td>
<td>0.260</td>
<td>0.490</td>
<td>0.670</td>
<td>0.790</td>
</tr>
<tr>
<td>Coverage of UPoSI Box</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.930</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>of UPoSI Box</td>
<td>-1.395</td>
<td>-1.393</td>
<td>-1.391</td>
</tr>
<tr>
<td>Coverage of UPoSI</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>of UPoSI</td>
<td>-1.395</td>
<td>-1.395</td>
<td>-1.395</td>
</tr>
<tr>
<td>Setting C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage of SelInf</td>
<td>0.870</td>
<td>0.760</td>
<td>0.690</td>
<td>0.650</td>
<td>0.600</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>-0.652</td>
<td>-0.191</td>
<td>0.092</td>
<td>0.284</td>
</tr>
<tr>
<td>Prop. of Infinite Volume</td>
<td>0.100</td>
<td>0.240</td>
<td>0.440</td>
<td>0.590</td>
<td>0.740</td>
</tr>
<tr>
<td>Coverage of UPoSI Box</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>of UPoSI Box</td>
<td>-1.396</td>
<td>-1.396</td>
<td>-1.396</td>
</tr>
<tr>
<td>Coverage of UPoSI</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
<td>0.920</td>
</tr>
<tr>
<td>log(Vol)/</td>
<td>/M</td>
<td>of UPoSI</td>
<td>-1.396</td>
<td>-1.396</td>
<td>-1.396</td>
</tr>
</tbody>
</table>

and $C^\dagger_M(\alpha)$ provide computationally efficient, data-driven upper bounds on the quantiles of “max-$|l|$” for the selected model $M$ without having to compute the maximum over all models. It should, however, be noted that the confidence regions $\hat{\mathcal{I}}_{n,\hat{M}}$ are always larger than the corresponding regions $B_{n,\hat{M}}$. The reason simply is that $B_{n,\hat{M}}$ is the smallest rectangle that encloses the region $\hat{\mathcal{R}}_{n,\hat{M}}$ and $\hat{\mathcal{I}}_{n,\hat{M}}$ is also a rectangle that encloses the region $\hat{\mathcal{R}}_{n,\hat{M}}$.

**S.3. A Generalization to Linear Regression with Problematic Data.** A simple generalization of Theorems 4.1 and 4.2 allows us to address situations with problematic data that may suffer from outliers and/or missing values. It might then be useful to modify the estimators of $\Sigma_n$ and $\Gamma_n$ to account for these difficulties. Modified estimators for the cases of missing data, errors-in-covariates and multiplicative noise can be found in Loh and Wainwright (2012, Examples 1, 2 and 3). For the case of outliers either in the sense of the classical robustness literature or in the adversarial corruption setting, see Chen et al. (2013). For further settings of applicability, see Kuchibhotla et al. (2018, pages 11–12).

For correct use of Theorem S.3.1 below, it is crucial that sub-model regression is based on
the proper sub-matrix of $\Sigma_n^*$ and sub-vector of $\Gamma_n^*$ and their respective modified estimators.

For the intended generalization, consider the following setting. Let $\hat{\Sigma}_n^*, \Sigma_n^*$ be two $p \times p$ matrices and $\hat{\Gamma}_n^*, \Gamma_n^*$ be two $p$-vectors. Consider the error norms

$$D_n^{\hat{\Gamma}^*} := |\hat{\Gamma}_n^* - \Gamma_n^*|_\infty \quad \text{and} \quad D_n^{\hat{\Sigma}^*} := |\hat{\Sigma}_n^* - \Sigma_n^*|_\infty.$$
Define for every $M \in \mathcal{M}_p(p)$, the estimator and the corresponding target as

$$
\hat{\xi}_{n,M} := \arg\min_{\theta \in \mathbb{R}^M} \left\{ \theta^T \hat{\Sigma}_n^* (M) \theta - 2 \theta^T \hat{\Gamma}_n^* (M) \right\},
$$

$$
\hat{\xi}_{n,M} := \arg\min_{\theta \in \mathbb{R}^M} \left\{ \theta^T \Sigma_n^* (M) \theta - 2 \theta^T \Gamma_n^* (M) \right\}.
$$

**Fig 1.** Comparison of coverage and volume of UPoSI with sample splitting. In all cases the volume of our confidence regions are at least as good as sample splitting. The latter is slightly more conservative in coverage in some cases, but not dramatically so.
Consider the confidence regions $\hat{R}^*_{n,M}$ and $\hat{R}^{\dagger}_{n,M}$, analogues to those before, as

$$\hat{R}^*_{n,M} := \left\{ \theta \in \mathbb{R}^{[M]} : |\hat{\Sigma}_n(M)\hat{\xi}_{n,M} - \theta|_\infty \leq C_n^\Gamma^*(\alpha) + C_n^{\Sigma^*}(\alpha) \|\theta\|_1 \right\},$$

$$\hat{R}^{\dagger}_{n,M} := \left\{ \theta \in \mathbb{R}^{[M]} : |\hat{\Sigma}_n(M)\hat{\xi}_{n,M} - \theta|_\infty \leq C_n^\Gamma^*(\alpha) + C_n^{\Sigma^*}(\alpha) |\hat{\xi}_{n,M}|_1 \right\},$$

where $C_n^\Gamma^*(\alpha)$ and $C_n^{\Sigma^*}(\alpha)$ are constants (or joint quantiles) that satisfy,

$$\Pr(D_n^\Gamma^* \leq C_n^\Gamma^*(\alpha) \text{ and } D_n^{\Sigma^*} \leq C_n^{\Sigma^*}(\alpha)) \geq 1 - \alpha.$$

Finally, let $\Lambda_n^*(k) = \min\{\lambda_{\min}(\hat{\Sigma}_n(M)) : M \in M_P(k)\}$.

**Theorem S.3.1.** For the set of confidence regions $\left\{ \hat{R}^*_{n,M} : M \in M_P(p) \right\}$ we have

$$\Pr \left( \bigcap_{M \in M_P(p)} \left\{ \xi_{n,M} \in \hat{R}^*_{n,M} \right\} \right) \geq 1 - \alpha,$$

and for any $1 \leq k \leq p$ that satisfies $kD_{2n}^\star = o_p(\Lambda_n^*(k))$, we have

$$\liminf_{n \to \infty} \Pr \left( \bigcap_{M \in M_P(k)} \left\{ \xi_{n,M} \in \hat{R}^{\dagger}_{n,M} \right\} \right) \geq 1 - \alpha.$$

**Proof.** The proof is exactly the same as for Theorems 4.1 and 4.2. The reader only has to realize that we did not use any structure of $\hat{\Sigma}_n$ and $\hat{\Gamma}_n$ nor did we assume that they are unbiased estimators of $\Sigma_n$ and $\Gamma_n$, respectively.

**S.4. Connection to High-dimensional Regression and Other Confidence Regions.** The confidence regions $\hat{R}_{n,M}$ and $\hat{R}^{\dagger}_{n,M}$ have a very close connection to a well-known estimator in the high-dimensional linear regression literature called the Dantzig selector proposed by Candes and Tao (2007) and the closely related ones by Rosenbaum and Tsybakov (2010) and Chen et al. (2013). These methods are not related to post-selection inference and were proposed under a linear model assumption. The Dantzig selector estimates $\beta_0 \in \mathbb{R}^p$ using observations $(X_i^T, Y_i)$ that satisfy $Y_i = X_i^T \beta_0 + \epsilon_i$ for independent and identically distributed errors $\epsilon_i$ with a mean zero normal distribution. Candes and Tao (2007), like many others, assumed fixed covariates. In our notation, the Dantzig selector is defined by the optimization problem

$$\minimize \quad \|\beta\|_1 \quad \text{subject to} \quad |\hat{\Gamma}_n - \hat{\Sigma}_n\beta|_\infty \leq \lambda_n,$$

for some tuning parameter $\lambda_n$ that converges to zero as $n$ increases. To relate this to our confidence regions $\hat{R}^{\dagger}_{n,M}$ (in (11)), note that $\hat{\Gamma}_n - \hat{\Sigma}_n\beta_0 = \Sigma_n(\hat{\beta} - \beta_0)$ where $\hat{\beta}$ is any least
squares estimator (which is not unique if $p \geq n$). The estimator of Rosenbaum and Tsybakov (2010) and Chen et al. (2013) is defined as follows.

$$\minimize \| \beta \|_1 \quad \text{subject to} \quad |\hat{\Gamma}_n - \hat{\Sigma}_n \beta|_{\infty} \leq \lambda_n + \delta_n \| \beta \|_1,$$

for some tuning parameters $\lambda_n$ and $\delta_n$ both converging to zero as $n$ increases. This constraint set corresponds to our confidence region $\hat{R}_{n,M}$ in Theorem 4.1 if $\lambda_n = C_n^T(\alpha)$ and $\delta_n = C_n^\Sigma(\alpha)$. With these identifications, we see that these high-dimensional estimators amount to minimizers of $\| \beta \|_1$ over our confidence regions.

A similar connection can be established between confidence regions and lasso (Tibshirani (1996)) as well as sqrt-lasso (Belloni et al. (2011)). To this end, define for every $M \in \mathcal{M}_p(p)$ the following confidence regions corresponding to lasso:

$$\hat{R}_{n,M}^{\hat{\beta}} := \left\{ \theta \in \mathbb{R}^{|M|} : \hat{R}_n(\theta; M) \leq \hat{R}_n(\hat{\beta}_{n,M}; M) + 4C_n^T(\alpha) |\hat{\beta}_{n,M}|_1 + 2C_n^\Sigma(\alpha) |\hat{\beta}_{n,M}|^2_1 \right\},$$

$$\hat{R}_{n,M} := \left\{ \theta \in \mathbb{R}^{|M|} : \hat{R}_n(\theta; M) \leq \hat{R}_n(\hat{\beta}_{n,M}; M) + 2C_n^T(\alpha) \left[ |\hat{\beta}_{n,M}|_1 + \| \theta \|_1 \right] + C_n^\Sigma(\alpha) \left[ |\hat{\beta}_{n,M}|^2_1 + \| \theta \|^2_1 \right] \right\}.$$

Also, define the following confidence regions corresponding to sqrt-lasso

$$\tilde{R}_{n,M}^{\hat{\beta}} := \left\{ \theta \in \mathbb{R}^{|M|} : \tilde{R}_n^{1/2}(\theta; M) \leq \tilde{R}_n^{1/2}(\hat{\beta}_{n,M}; M) + 2C_n^{\Sigma}(\alpha) \left( 1 + |\hat{\beta}_{n,M}|_1 \right) \right\},$$

$$\tilde{R}_{n,M} := \left\{ \theta \in \mathbb{R}^{|M|} : \tilde{R}_n^{1/2}(\theta; M) \leq \tilde{R}_n^{1/2}(\hat{\beta}_{n,M}; M) + C_n^{\Sigma}(\alpha) \left( 1 + \| \theta \|_1 \right) + C_n^{1/2}(\alpha) \left( 1 + |\hat{\beta}_{n,M}|_1 \right) \right\},$$

where $\tilde{R}_n(\cdot; M)$ is the empirical least squares objective function defined in Equation (2). Here $C_n(\alpha)$ is the $(1 - \alpha)$-upper quantile of $\max\{D_n^T, D_n^\Sigma\}$. Importantly note that, if the covariates are fixed we have $C_n^\Sigma(\alpha) = 0$ and hence the quantities on the right hand side inequalities correspond to the objective functions of lasso and sqrt-lasso, respectively.

**Theorem S.4.1.** For any $n \geq 1, p \geq 1$, the following simultaneous inference guarantee holds:

\begin{align*}
\text{(E.2)} & \quad \mathbb{P} \left( \bigcap_{M \in \mathcal{M}_p(p)} \{ \beta_{n,M} \in \hat{R}_{n,M} \} \right) \geq 1 - \alpha, \\
\text{(E.3)} & \quad \mathbb{P} \left( \bigcap_{M \in \mathcal{M}_p(p)} \{ \beta_{n,M} \in \tilde{R}_{n,M} \} \right) \geq 1 - \alpha,
\end{align*}
and for any $1 \leq k \leq p$ satisfying $\text{(A1)}(k)$, we have

\begin{equation}
\liminf_{n \to \infty} \mathbb{P} \left( \bigcap_{M \in M_p(p)} \{ \beta_{n,M} \in \hat{R}_{n,M}^\dagger \} \right) \geq 1 - \alpha,
\end{equation}

\begin{equation}
\liminf_{n \to \infty} \mathbb{P} \left( \bigcap_{M \in M_p(p)} \{ \beta_{n,M} \in \hat{R}_{n,M}^\dagger \} \right) \geq 1 - \alpha,
\end{equation}

**Proof.** See Appendix S.9 for a proof.

The connection established above between the post-selection confidence regions and high-dimensional linear regression methods suggest an interesting question: Does there exists a deeper connection between post-selection inference and high-dimensional estimation? Other than the illustrative results above, we do not have a more general answer.

**Remark S.4.1** (Intersection of Confidence Regions) As mentioned earlier, all our confidence regions are based on deterministic inequalities. This implies that the intersection of the confidence regions $\hat{R}_{n,M}, \hat{R}_{n,M}^\dagger$ and $\hat{R}_{n,M}$ provides valid simultaneous and hence post-selection inference. For any $1 \leq k \leq p$ satisfying $\text{(A1)}(k)$, we have

\begin{equation}
\liminf_{n \to \infty} \mathbb{P} \left( \bigcap_{M \in M_p(p)} \{ \hat{R}_{n,M} \cap \hat{R}_{n,M}^\dagger \cap \hat{R}_{n,M} \} \right) \geq 1 - \alpha.
\end{equation}

To prove this, let $\hat{C}_{n,M}, \hat{C}_{n,M}^\dagger$ and $\hat{C}_{n,M}$ represent the confidence sets $\hat{R}_{n,M}, \hat{R}_{n,M}^\dagger$ and $\hat{R}_{n,M}$ with $(C_n^\Gamma(\alpha), C_n^{\Sigma}(\alpha))$ replaced by $(\hat{D}_n^\Gamma, \hat{D}_n^{\Sigma})$. From the proofs of Theorems 4.1, 4.2 and S.4.1, it follows that

\begin{equation}
\liminf_{n \to \infty} \mathbb{P} \left( \bigcap_{M \in M_p(p)} \{ \hat{C}_{n,M} \cap \hat{C}_{n,M}^\dagger \cap \hat{C}_{n,M} \} \right) = 1.
\end{equation}

Hence, by the definition (13) of $(C_n^\Gamma(\alpha), C_n^{\Sigma}(\alpha))$, the result (E.6) follows. The intersection of the confidence regions is provably smaller. By a similar argument it is possible to further intersect with the confidence regions $\hat{R}_{n,M}^\dagger, \hat{R}_{n,M}$, and $\hat{R}_{n,M}$.

**Remark S.4.2** (Usefulness of Lasso-based Regions) The confidence regions discussed in this section are given solely for the illustrative purposes and for making a solid connection between post-selection inference and high-dimensional linear regression. These confidence regions are ellipsoidal in shape and have larger volume than the confidence region $\hat{R}_{n,M}^\dagger$ in terms of the rate. The claim about larger volume is not presented here but is not difficult to prove. This rate comparison is only asymptotic and the intersection argument presented in Remark S.4.1 might still be useful in finite samples.
S.5. Proof of Lemma 4.1. Fix $M \in \mathcal{M}_p(k)$ with $kD_n^\Sigma \leq \Lambda_n(k)$. Observe that the least squares estimator satisfies

$$\hat{\beta}_{n,M} - \beta_{n,M} = \left(\hat{\Sigma}_n(M)\right)^{-1}\left\{\hat{\Gamma}_n(M) - \Gamma_n(M)\right\} = \left\{\hat{\Sigma}_n(M) - \Sigma_n(M)\right\} \beta_{n,M},$$

and for all $M \in \mathcal{M}_p(k)$,

$$(E.7) \quad \left\| \hat{\Sigma}_n(M) - \Sigma_n(M) \right\|_{op} \leq \sup_{\|\theta\|_2 \leq 1} \left\| \theta^T \left(\hat{\Sigma}_n - \Sigma_n\right) \theta \right\| \leq k \left\| \hat{\Sigma}_n - \Sigma_n \right\|_{\infty} = kD_n^\Sigma.$$ 

Thus, for all $M \in \mathcal{M}_p(k)$,

$$\left\| \left(\hat{\Sigma}_n(M)\right)^{-1} \right\|_{op} \leq (\Lambda_n(k) - kD_n^\Sigma)^{-1}.  
$$

Hence, for $k$ satisfying $kD_n^\Sigma \leq \Lambda_n(k)$,

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 \leq \frac{|\hat{\Gamma}_n(M) - \Gamma_n(M)|_2 + |[\hat{\Sigma}_n(M) - \Sigma_n(M)] \beta_{n,M}|_2}{\Lambda_n(k) - kD_n^\Sigma} \leq \frac{|M|^{1/2} |\hat{\Gamma}_n(M) - \Gamma_n(M)|_{\infty} + |[\hat{\Sigma}_n(M) - \Sigma_n(M)] \beta_{n,M}|_{\infty}}{\Lambda_n(k) - kD_n^\Sigma} \leq \frac{|M|^{1/2} (D_n^\Gamma + D_n^\Sigma \|\beta_{n,M}\|_1)}{\Lambda_n(k) - kD_n^\Sigma}.  
$$

Now applying

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_1 \leq |M|^{1/2} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2,$$

the result follows.

S.6. Proof of Theorem 4.2. The starting point of this proof is Equation (17). Under assumption (A1)(k), Lemma 4.1 (inequality (18)) implies that for all $M \in \mathcal{M}_p(k)$,

$$\frac{D_n^\Gamma + D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \|\beta_{n,M}\|_1 - 1} \leq \frac{D_n^\Sigma \|\hat{\beta}_{n,M} - \beta_{n,M}\|_1}{D_n^\Gamma + D_n^\Sigma \|\beta_{n,M}\|_1} \leq \frac{D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \|\beta_{n,M}\|_1} \cdot \frac{|M| \left\{D_n^\Gamma + D_n^\Sigma \|\beta_{n,M}\|_1\right\}}{\Lambda_n(k) - |M|D_n^\Sigma} \leq \frac{D_n^\Sigma}{\Lambda_n(k) - kD_n^\Sigma}.  
$$

Therefore, for $1 \leq k \leq p$ satisfying assumption (A1)(k),

$$\sup_{M \in \mathcal{M}_p(k)} \frac{D_n^\Gamma + D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \|\beta_{n,M}\|_1 - 1} \leq \frac{kD_n^\Sigma/\Lambda_n(k)}{1 - (kD_n^\Sigma/\Lambda_n(k))} = o_p(1).$$
Hence,
\[
\liminf_{n \to \infty} \mathbb{P}\left( \bigcap_{M \in \mathcal{M}_p(k)} \left\{ \left\| \Sigma_n(M) \{ \beta_{n,M} - \hat{\beta}_{n,M} \} \right\|_\infty \leq D_n^\Gamma + D_n^\Sigma \left\| \beta_{n,M} \right\|_1 \right\} \right) = 1.
\]

The definition of \((C_n^\Gamma(\alpha), C_n^\Sigma(\alpha))\) in (13) proves the required result.

**S.7. Proof of Proposition 5.1.** For any fixed model \(M\), the Lebesgue measure of the confidence region is given by

\[(E.8) \quad \text{Leb}(R^1_{n,M}) = |\hat{\Sigma}_n(M)|^{-1} \left( 2C_n^\Gamma(\alpha) + 2C_n^\Sigma(\alpha) |\hat{\beta}_{n,M}|_1 \right)^{|M|},\]

which converges to zero as \(n\) tends to infinity. Here for any matrix \(A \in \mathbb{R}^{p \times p}\), \(|A|\) denotes the determinant of \(A\). This equality follows since the confidence region \(R^1_M\) can be written as

\[R^1_M = \left\{ \hat{\beta}_{n,M} + (\hat{\Sigma}_n(M))^{-1} \theta : \|\theta\|_\infty \leq \left( C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha) |\hat{\beta}_{n,M}|_1 \right) \right\}.
\]

By inequality (E.7), for all \(M \in \mathcal{M}_p(k)\)
\n\[|\hat{\Sigma}_n(M)|^{-1} \leq (\Lambda_n(k) - kD_n^\Sigma)^{-|M|}.
\]

The result now follows from equation (E.8) and uniform consistency of \(\hat{\beta}_{n,M}\) in the \(\|\cdot\|_1\)-norm as shown in Lemma 4.1 under \((A1)(k)\).

To prove the second result, first note that from Lemma 5.1,
\n\[\max\{C_n^\Gamma(\alpha), C_n^\Sigma(\alpha)\} = O\left( \sqrt{\frac{\log p}{n}} \right),\]

since the second term in the expectation bound in Lemma 5.1 is of lower order than the first term under the assumption \((\log p)^{2/\alpha}(\log n)^{2/\alpha-1/2} = o(n^{1/2})\) of Lemma 5.1. The result is now proved if we prove that for all \(M \in \mathcal{M}_p(k)\),
\n\[(E.9) \quad \|\beta_{n,M}\|_1^2 \leq \frac{|M|}{\Lambda_n(k)} \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E} [Y_i^2] \right).
\]

By definition of \(\beta_{n,M}\) it follows that
\n\[0 \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} [Y_i^2] - \frac{1}{n} \sum_{i=1}^n \hat{\beta}_{n,M}^T \mathbb{E} \left[ X_i(M)X_i^T(M) \right] \beta_{n,M} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ (Y_i - X_i^T(M)\beta_{n,M})^2 \right].
\]

Therefore, by definition of \(\Lambda_n(k)\),
\n\[\Lambda_n(k) \|\beta_{n,M}\|_2^2 \leq \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E} [Y_i^2] \right).
\]

Now using the inequality \(\|\beta_{n,M}\|_1 \leq \sqrt{|M|} \|\beta_{n,M}\|_2\), inequality (E.9) follows.
S.8. Proof of Proposition 6.1. From the proof of Theorem 4.1, we know that for all $M \in \mathcal{M}_p(k)$,

$$\|\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M})\|_\infty \leq D_n^\Gamma.$$  

Observe that

$$\|\hat{\beta}_{n,M} - \beta_{n,M}\| \leq \|\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M})\|_\infty + \|\hat{\Sigma}_n(M) - I|M\| (\hat{\beta}_{n,M} - \beta_{n,M})\|_2 \leq \|\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M})\|_\infty + \delta |\hat{\beta}_{n,M} - \beta_{n,M}|_2$$

(E.10)

From Remark 4.3 of Kuchibhotla et al. (2018), we get that

$$\sup_{M \in \mathcal{M}_p(k)} \|\hat{\beta}_{n,M} - \beta_{n,M}\|_2 \leq \frac{\sqrt{k}D_n^\Gamma}{\Lambda_n(k)}$$

(Note that in the notation of Kuchibhotla et al. (2018), $D_n^\Gamma$ is different and as shown in Proposition 3.1 there the bound above holds.) Therefore, combining (E.10) and (6.1), we get that for all $M \in \mathcal{M}_p(k)$,

$$\|\hat{\beta}_{n,M} - \beta_{n,M}\|_\infty \leq D_n^\Gamma \left(1 + \frac{\delta \sqrt{k}}{\Lambda_n(k)}\right).$$

From the RIP property (25), $\Lambda_n(k) \geq 1 - \delta$ and so, for all $M \in \mathcal{M}_p(k)$,

$$\|\hat{\beta}_{n,M} - \beta_{n,M}\|_\infty \leq D_n^\Gamma \left(1 + \frac{\delta \sqrt{k}}{(1 - \delta)}\right).$$

Therefore, under $\delta \sqrt{k} \to 0$ and using the definition of $C_n^\Gamma(\alpha)$,

$$\liminf_{n \to \infty} \mathbb{P}\left(\bigcap_{M \in \mathcal{M}_p(k)} \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}^{\RIP}\} \right) \geq 1 - \alpha.$$  

To prove (26), first note that the Hausdorff distance is shift invariant and hence

$$d_H\left(\hat{\mathcal{R}}_{n,M}^{\RIP}, \hat{\mathcal{R}}_{n,M}\right) = d_H\left(\hat{\mathcal{R}}_{n,M}^{\RIP} - \hat{\beta}_{n,M}, \hat{\mathcal{R}}_{n,M} - \hat{\beta}_{n,M}\right).$$

Now observe that

$$A_1 := \hat{\mathcal{R}}_{n,M}^{\RIP} - \hat{\beta}_{n,M} = \left\{\Delta \in \mathbb{R}^{[M]} : \|\Delta\|_\infty \leq C_n^\Gamma(\alpha)\right\},$$

$$A_2 := \hat{\mathcal{R}}_{n,M} - \hat{\beta}_{n,M} = \left\{\Delta \in \mathbb{R}^{[M]} : \|\hat{\Sigma}_n(M)\Delta\|_\infty \leq C_n^\Gamma(\alpha)\right\}.$$
This implies that if \( \Delta \in A_1 \) then \((\hat{\Sigma}_n(M))^{-1}\Delta \in A_2\) and if \( \Delta \in A_2 \) then \(\Sigma_n(M)\Delta \in A_1\). Hence, we get

\[
\sup_{\Delta \in A_1} \inf_{\Delta' \in A_2} \| \Delta - \Delta' \|_2 \leq \sup_{\Delta \in A_1} \| \Delta - (\hat{\Sigma}_n(M))^{-1}\Delta \|_2 \\
\leq \sup_{\Delta \in A_1} \| I_{|M|} - (\hat{\Sigma}_n(M))^{-1} \|_{op} \| \Delta \|_2 \\
\leq \sup_{\Delta \in A_1} \sqrt{|M|} \| \Delta \|_\alpha \| I_{|M|} - (\hat{\Sigma}_n(M))^{-1} \|_{op} \leq C_n^\Gamma(\alpha) \left( \frac{|M|^{1/2} \delta}{1 - \delta} \right),
\]

where the last inequality follows from the RIP condition (25) and the fact that \(\| \Delta \|_\infty \leq C_n^\Gamma(\alpha)\) for \(\Delta \in A_1\). Similarly,

\[
\sup_{\Delta \in A_2} \inf_{\Delta' \in A_1} \| \Delta - \Delta' \|_2 \leq \sup_{\Delta \in A_2} \| \Delta - \hat{\Sigma}_n(M)\Delta \|_2 \\
\leq \sup_{\Delta \in A_2} \| (\hat{\Sigma}_n(M))^{-1} - I_{|M|} \|_{op} \| \hat{\Sigma}_n(M)\Delta \|_2 \leq C_n^\Gamma(\alpha) \left( \frac{|M|^{1/2} \delta}{1 - \delta} \right).\]

Combining both the cases, inequality (26) follows.

**S.9. Proof of Theorem S.4.1.** Only the proof of (E.2) and (E.4) is provided and the steps to prove (E.3) and (E.5) are sketched since the proof is similar.

It is easy to verify that for any \(M \subseteq \mathcal{M}_p(p)\) and \(\theta \in \mathbb{R}^{|M|}\)

\[
\begin{equation}
\theta^T \hat{\Sigma}_n(M)\theta - 2\theta^T \hat{\Gamma}_n(M) - \theta^T \Sigma_n(M)\theta + 2\theta^T \Gamma_n(M) \leq \| \theta \|^2 \mathcal{D}_n^\Sigma + 2 \| \theta \|_1 \mathcal{D}_n^\Gamma.
\end{equation}
\]

Therefore, for every \(M \in \mathcal{M}_p(p)\),

\[
\beta_{n,M} \hat{\Sigma}_n(M)\beta_{n,M} - 2\beta_{n,M}^T \hat{\Gamma}_n(M) \leq \beta_{n,M} \Sigma_n(M)\beta_{n,M} - 2\beta_{n,M}^T \Gamma_n(M) + 2\mathcal{D}_n^\Sigma \| \beta_{n,M} \|_1 + \mathcal{D}_n^\Sigma \| \beta_{n,M} \|^2_1 \\
\leq \beta_{n,M} \Sigma_n(M)\beta_{n,M} - 2\beta_{n,M}^T \Gamma_n(M) + 2\mathcal{D}_n^\Gamma \| \beta_{n,M} \|_1 + \mathcal{D}_n^\Sigma \| \beta_{n,M} \|^2_1 \\
\leq \beta_{n,M} \hat{\Sigma}_n(M)\beta_{n,M} - 2\beta_{n,M}^T \Gamma_n(M) + 2\mathcal{D}_n^\Gamma \| \beta_{n,M} \|_1 + \| \beta_{n,M} \|_1 \\
+ \mathcal{D}_n^\Sigma \| \beta_{n,M} \|_1^2 + \| \beta_{n,M} \|^2_1.
\]

Here the first inequality follows from inequality (E.11) with \(\theta = \beta_{n,M}\), the second inequality follows from the definition of \(\beta_{n,M}\) (see Equation (??)) and the third inequality follows from inequality (E.11) with \(\theta = \hat{\beta}_{n,M}\). Adding the sample average of \(\{Y_i^2 : 1 \leq i \leq n\}\) on both sides, we get for all \(M \in \mathcal{M}_p(p)\),

\[
\hat{R}_n(\beta_{n,M}; M) \leq \hat{R}_n(\hat{\beta}_{n,M}; M) + 2\mathcal{D}_n^\Gamma \| \hat{\beta}_{n,M} \|_1 + \| \beta_{n,M} \|_1 + \mathcal{D}_n^\Sigma \| \hat{\beta}_{n,M} \|^2_1 + \| \beta_{n,M} \|^2_1.
\]
This implies the first result (E.2). To prove the second result (E.4), note that
\[
\left| \left( \frac{D_n^\Gamma + D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \| \beta_{0,M} \|_1} \right)^2 - 1 \right| \leq \left( \frac{D_n^\Gamma + D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \| \beta_{0,M} \|_1} \right)^2 \left( \frac{D_n^\Gamma + D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \| \beta_{0,M} \|_1} \right)^2 - 1 = 2 \left( \frac{D_n^\Gamma + D_n^\Sigma}{D_n^\Gamma + D_n^\Sigma \| \beta_{0,M} \|_1} \right)^2 - 1,
\]
which converges to zero under assumption (A1)(k), following the proof of Theorem 4.2. This implies that the error
\[
\left[ 2D_n^\Gamma \| \hat{\beta}_{n,M} \|_1 + D_n^\Sigma \| \beta_{n,M} \|_1^2 \right] - \left[ 2D_n^\Gamma \| \beta_{n,M} \|_1 + D_n^\Sigma \| \beta_{n,M} \|_1^2 \right],
\]
is of smaller order than each of the terms uniformly in \( M \in M_p(k) \). The second result (E.4) then follows trivially by substituting the estimated parameters for the targets in inequality (E.12) and using the definition of \( (C_n^\Gamma(\alpha), C_n^\Sigma(\alpha)) \).

To prove the results with square-root lasso based regions, note that from inequality (E.12)
\[
\hat{R}_n^1/2(\beta_{n,M}; M) \leq \hat{R}_n^1/2(\hat{\beta}_{n,M}; M) + \max\{D_n^\Gamma, D_n^\Sigma\}^{1/2} \left( 1 + \| \hat{\beta}_{n,M} \|_1 \right) + \max\{D_n^\Gamma, D_n^\Sigma\}^{1/2} \left( 1 + \| \beta_{n,M} \|_1 \right).
\]

S.10. High-dimensional CLT and Bootstrap Consistency. Suppose \( W_i, 1 \leq i \leq n \) are independent random vectors in \( \mathbb{R}^q \) with finite second moment. Let \( G_i, 1 \leq i \leq n \) be independent Gaussian random vectors in \( \mathbb{R}^q \) with mean zero satisfying
\[
\mathbb{E} [G_i G_i^T] = \mathbb{E} [W_i W_i^T] \quad \text{for all} \quad 1 \leq i \leq n.
\]
Set
\[
S_n^W := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{W_i - \mathbb{E} [W_i]\} \quad \text{and} \quad S_n^G := \frac{1}{\sqrt{n}} \sum_{i=1}^n G_i.
\]
Before deriving the exact rate under the assumption (19) of Lemma 5.1, we prove that a central limit theorem for \( S_n^W \) implies a CLT for \( (D_n^\Gamma, D_n^\Sigma) \). Observe that for any \( t_1, t_2 \in \mathbb{R}^+ \cup \{0\} \),
\[
\{D_n^\Gamma \leq t_1, \ D_n^\Sigma \leq t_2\} = \left\{ -t_1 \leq \frac{1}{n} \sum_{i=1}^n \{X_i(j)Y_i - \mathbb{E} [X_i(j)Y_i]\} \leq t_1 \quad \text{for all} \quad 1 \leq j \leq p \right\} \cap \left\{ -t_2 \leq \frac{1}{n} \sum_{i=1}^n \{X_i(l)X_i(m) - \mathbb{E} [X_i(l)X_i(m)]\} \leq t_2 \quad \text{for all} \quad 1 \leq l \leq m \leq p \right\}.
\]
The right hand side here is a rectangle in terms of the vector \( S_n^W \) with vectors \( W_i \) containing
\[
(X_i(j)Y_i, 1 \leq j \leq p; X_i(l)X_i(m), 1 \leq l \leq m \leq p).
\]
Note that $W_i$’s are vectors in $\mathbb{R}^q$ with
\[ q = 2p + \frac{p(p - 1)}{2}. \]

Let $\mathcal{A}^{re}$ denote the set of all rectangles in $\mathbb{R}^q$, that is, $\mathcal{A}^{re}$ consists of all sets $A$ of the form
\[ A = \{ z \in \mathbb{R}^q : a(j) \leq z(j) \leq b(j) \text{ for all } 1 \leq j \leq q \}, \]
for some vectors $a, b \in \mathbb{R}^q$. Define
\[ L_{n,q} := \max_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ |W_i(j) - \mathbb{E}[W_i(j)]|^2 \right]. \]

Finally, set for any class $A$ of (Borel) sets in $\mathbb{R}^q$,
\[ \rho_n(A) := \sup_{A \in \mathcal{A}} |\mathbb{P}(S_n^W \in A) - \mathbb{P}(S_n^G \in A)|. \]

The following theorem proved in Section 6 of Kuchibhotla and Chakrabortty (2018) provides a central limit theorem for $S_n^W$ over all rectangles. The proof there is based on Theorem 2.1 of Chernozhukov et al. (2017).

**Theorem S.10.1.** Suppose $W_1, \ldots, W_n$ are independent mean zero random vectors in $\mathbb{R}^q$ satisfying for some $\gamma, B, K_{n,q} > 0$,
\[ \min_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^{n} \text{Var}[W_i^2(j)] \geq B \quad \text{and} \quad \max_{1 \leq i \leq n} \max_{1 \leq j \leq q} \|W_i(j)\|_{\psi,\gamma} \leq K_{n,q}. \]

Assume further that for some constant $K_2 > 0$ (depending only on $B$),
\[ \frac{1}{8K_2K_{n,q}^2} \left( \frac{nL_{n,q}}{\log q} \right)^{1/3} \geq \max \{1, 2^{1/\gamma - 1}\} \left\{ (\log q)^{1/\gamma} + (6/\gamma)^{1/\gamma} + 1 \right\}. \]

Then there exist constants $K_1 > 0$ depending only on $B$, and $C_{\gamma,B} > 0$ depending only on $B, \gamma$ such that
\[ \rho_n(\mathcal{A}^{re}) \leq K_1 \left( \frac{L_{n,q}^2 \log^7 q}{n} \right)^{1/6} + C_{\gamma,B} \frac{K_{n,q}^6 \log q}{n}. \]

Based on (E.13), it is clear that Theorem S.10.1 implies a CLT for $(D_n^T, D_n^G)$. This does not require the observations to be identically distributed or equal expectations for the $W_i$ vectors defined in (E.14).
S.10.1. Bootstrap Consistency. In this subsection, we consider the consistency of the multiplier bootstrap based on Section 4.1 of Chernozhukov et al. (2017). It is also possible to consider the empirical bootstrap in high dimensions and prove its consistency based on the proof of Proposition 4.3 of Chernozhukov et al. (2017). We do not prove it here as the proof techniques are the same.

Let $e_1, \ldots, e_n$ be a sequence of independent standard normal random variables independent of $W_n := \{W_1, \ldots, W_n\}$. Set

$$W_n := \frac{1}{n} \sum_{i=1}^{n} W_i \in \mathbb{R}^q,$$

and consider the normalized sum

$$S_{n,e,W} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i (W_i - W_n).$$

Note that

$$S_{n,e,W} \mid W_n \sim N \left( 0, \frac{1}{n} \sum_{i=1}^{n} (W_i - W_n) (W_i - W_n)^\top \right) \in \mathbb{R}^q.$$

To prove consistency of multiplier bootstrap, we bound a quantity similar to $\rho_n(A^{re})$, defined as

$$\rho_{n,MB} (A^{re}) := \sup_{A \in A^{re}} \mathbb{P} \left( S_{n,e,W} \in A \mid W_n \right) - \mathbb{P} \left( S_{n,G^*} \in A \right),$$

where

$$S_{n,G^*} \sim N \left( 0, \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ (W_i - \bar{\mu}_n) (W_i - \bar{\mu}_n)^\top \right] \right), \quad \text{with} \quad \bar{\mu}_n := \mathbb{E} \left[ W_n \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ W_i \right].$$

Note that $\operatorname{Var}(S_{n,W}) \neq \operatorname{Var}(S_{n,G^*})$ unless $\mathbb{E}[W_1] = \mathbb{E}[W_2] = \cdots = \mathbb{E}[W_n]$. Define

$$\Delta_{n,q} := \left\| \frac{1}{n} \sum_{i=1}^{n} (W_i - W_n) (W_i - W_n)^\top - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ (W_i - \bar{\mu}_n) (W_i - \bar{\mu}_n)^\top \right] \right\|_\infty.$$

Based on Theorem 4.1 and Remark 4.1 of Chernozhukov et al. (2017), we prove the following theorem under assumption (E.15).

**Theorem S.10.2.** If $W_i, 1 \leq i \leq n$ are independent mean zero random vectors, then under assumption (E.15),

$$\mathbb{E} \left[ \rho_{n,MB} (A^{re}) \right] \leq C \log^{2/3} q \left[ A_{n,q}^{1/3} \left( \frac{\log q}{n} \right)^{1/6} + K_{n,q}^{2/3} \left( \frac{\log q \log n}{n^{1/3}} \right)^{2/7} \right],$$

for some constant $C$ depending only on $\gamma, B$. Here

$$A_{n,q} := \max_{1 \leq l \leq m \leq q} \frac{1}{n} \sum_{i=1}^{n} \operatorname{Var}(W_i(l)W_i(m)).$$
Proof. As proved in Remark 4.1 of Chernozhukov et al. (2017), we have
\[ \rho_n^{\text{MB}}(A^{re}) \leq C \Delta_{n,q}^{1/3} \log^{2/3} q. \]
So, to prove the result, all we need is to prove
\[ \mathbb{E} \left[ \Delta_{n,q}^{1/3} \right] \leq M_q \left[ A_{n,q} \sqrt{\frac{\log q}{n}} + K_{n,q}^2 (\log q \log n)^{2/3} n^{-1} \right]^{1/3}, \]
for some constant \( M_q \). This follows from Theorem 4.2 of Kuchibhotla and Chakrabortty (2018).

Remark S.10.1 (Inconsistency under unknown unequal means) Since \( \text{Var}(S_n^W) \) and \( \text{Var}(S_n^{G*}) \) are not equal (in general), Theorem S.10.2 does not prove that
\[ \sup_{A \in A^{re}} \left| \mathbb{P} (S_n^{e,W} \in A|W_n) - \mathbb{P} (S_n^G \in A) \right| \to 0 \quad \text{in probability}. \]
It was proved in Kuchibhotla et al. (2018) that the variance of an average of non-identically distributed random variables cannot be consistently estimated if the (unequal) expectations are unknown and the same comment applies to the high-dimensional multiplier bootstrap. When \( \mathbb{E}[W_i] \) are not all the same for all \( 1 \leq i \leq n \), then the variance of \( S_n^W \) cannot be consistently estimated and so the distribution of \( S_n^W \) cannot be estimated consistently using the bootstrap. However, Theorem S.10.2 implies conservative inference. Observe that
\[ \text{Var} \left( S_n^W \right) = \frac{1}{n} \sum_{i=1}^{n} \text{Var} (W_i) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ (W_i - \mathbb{E}[W_n]) (W_i - \mathbb{E}[W_n])^T \right]. \]
Hence by Anderson’s lemma (Corollary 3 of Anderson (1955)), for all \( A \in A^{re} \),
\[ \mathbb{P} \left( S_n^{G*} \in A \right) \leq \mathbb{P} \left( S_n^G \in A \right). \]
Here \( A^{re} \) represents the set of all rectangles that are symmetric around zero. Thus, we get
\[ \liminf_{n \to \infty} \inf_{A \in A^{re}} (\mathbb{P} \left( S_n^G \in A \right) - \mathbb{P} \left( S_n^{e,W} \in A \big| W_n \right)) \geq 0. \]
Observe that the sets in (E.13) are centrally convex symmetric and hence Anderson’s lemma applies. Therefore, the multiplier bootstrap provides asymptotically conservative inference for \( (D_{n}^\Gamma, D_{n}^{\Sigma}) \) in general.

S.11. Rate Bounds on \( D_{n}^\Gamma \) and \( D_{n}^{\Sigma} \) under Dependence. In this section we derive rates of convergence of \( |\hat{\Omega}_n - \Omega_n| \) under dependence. We first describe some classical notions of dependence that include well-known dependent processes as special cases. The description is essentially taken from Pötscher and Prucha (1997). Let \( \{\xi_t : t \in \mathbb{Z}\} \) be a
stochastic process on some measure space. Let $F_{m,n}$ (for $m < n$) be the $\sigma$-field generated by $\{\xi_i : m \leq i \leq n\}$ where possibly $m = -\infty$ and $n = \infty$ are included. Define

$$\alpha(j) := \sup_{k \in \mathbb{Z}} \sup \{ |P(A \cap B) - P(A)P(B)| : A \in F_{-\infty,j}, B \in F_{k+j,\infty} \},$$

$$\phi(j) := \sup_{k \in \mathbb{Z}} \sup \{ |P(B|A) - P(B)| : A \in F_{-\infty,j}, B \in F_{k+j,\infty}, P(A) > 0 \}.$$

If $\alpha(j)$ (or correspondingly $\phi(j)$) converges to zero as $j$ approaches infinity then the process $\{\xi_t : t \in \mathbb{Z}\}$ is called $\alpha$-mixing (or correspondingly $\phi$-mixing). It is clearly seen that every $\phi$-mixing process is $\alpha$-mixing since for any event $A$ with $P(A) > 0$,

$$|P(A \cap B) - P(A)P(B)| \leq P(A)[P(B|A) - P(B)].$$

A process $\{\xi_t : t \in \mathbb{Z}\}$ is said to be $m$-dependent if $\alpha(j) = 0$ for all $j \geq m$. Evidently, $m$-dependent processes are $\phi$-mixing for any $m$ and so $\alpha$-mixing too. One very useful feature of $\alpha$-mixing processes is that measurable functions of finitely many elements of the process are themselves $\alpha$-mixing.

The dependence notion used in this section is the one called functional dependence, introduced by Wu (2005). It is possible to derive the results under classical dependence notions such as $\alpha$- or $\rho$-mixing, too. However, verifying the mixing assumptions can often be hard and many well-known processes do not satisfy them. See Wu (2005) for more details. It has also been shown that many econometric time series can be studied under the notion of functional dependence; see Wu and Mielniczuk (2010), Liu et al. (2013) and Wu and Wu (2016).

The dependence notion of Wu (2005) is written in terms of an input-output process that is easy to analyze in many settings. The process is defined as follows. Let $\{\varepsilon_i, \varepsilon'_i : i \in \mathbb{Z}\}$ denote a sequence of independent and identically distributed random variables on some measurable space $(\mathcal{E}, \mathcal{B})$. Let the $q$-dimensional (stochastic) process $W_i$ has a causal representation as

$$W_i = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) \in \mathbb{R}^q,$$

for some vector-valued function $G_i(\cdot) = (g_{i1}(\cdot), \ldots, g_{iq}(\cdot))$. By Wold’s representation theorem for stationary processes, this causal representation holds in many cases.

Define the non-decreasing filtration $F_i := \sigma(\ldots, \varepsilon_{i-1}, \varepsilon_i)$. Using this filtration, we also use the notation $W_i = G_i(F_i)$. To measure the strength of dependence, define for $r \geq 1$ and $1 \leq j \leq q$, the functional dependence measure

$$\delta_{s,r,j} := \max_{1 \leq i \leq n} \|W_i(j) - W_{i,s}(j)\|_r, \quad \text{and} \quad \Delta_{m,r,j} := \sum_{s=m}^{\infty} \delta_{s,r,j},$$

where

$$W_{i,s}(j) := g_{ij}(F_{i,i-s}) \quad \text{with} \quad F_{i,i-s} := \sigma(\ldots, \varepsilon_{i-s-1}, \varepsilon'_{i-s}, \varepsilon_{i-s+1}, \ldots, \varepsilon_{i-1}, \varepsilon_i).$$
The $\sigma$-field $F_{i,i-s}$ represents a coupled version of $F_i$. The quantity $\delta_{s,r,j}$ measures the dependence using the distance in terms of $\|\cdot\|_r$-norm between $g_{ij}(F_i)$ and $g_{ij}(F_{i,i-s})$. In other words, it is quantifying the impact of changing input $\varepsilon_{i-s}$ on the output $g_{ij}(F_i)$; see Definition 1 of Wu (2005). The dependence adjusted norm for the $j$-th coordinate is given by

$$\|\{W(j)\}\|_{r,\nu} := \sup_{m \geq 0} (m + 1)^\nu \Delta_{m,r,j}, \quad \nu \geq 0.$$ 

To summarize these measures for the vector-valued process, define

$$\|\{W\}\|_{r,\nu} := \max_{1 \leq j \leq q} \|\{W(j)\}\|_{r,\nu} \quad \text{and} \quad \|\{W\}\|_{\psi,\nu} := \sup_{r \geq 2} r^{-1/\beta} \|\{W\}\|_{r,\nu}.$$ 

**Remark S.11.1** (Independent Sequences) Any notion of dependence should at least include independent random variables. It might be helpful to understand how independent random variables fit into this framework of dependence. For independent random vectors $W_i$, the causal representation reduces to

$$W_i = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) = G_i(\varepsilon_i) \in \mathbb{R}^q.$$ 

Hence $W_i$ is not a function of any of the previous $\varepsilon_j, j < i$. This implies by definition (E.17) that

$$W_{i,s} = \begin{cases} G_i(\varepsilon_i) = W_i, & \text{if } s \geq 1, \\ G_i(\varepsilon_i') =: W_i', & \text{if } s = 0. \end{cases}$$

Here $W_i'$ represents an independent and identically distributed copy of $W_i$. Hence,

$$\delta_{s,r,j} = \begin{cases} 0, & \text{if } s \geq 1, \\ \|W_i(j) - W_i'(j)\|_r \leq 2 \|W_i(j)\|_r, & \text{if } s = 0. \end{cases}$$

It is now clear that for any $\nu > 0$,

$$\|\{W\}\|_{r,\nu} = \sup_{m \geq 0} (m + 1)^\nu \Delta_{m,r} = \Delta_{0,r} \leq 2 \max_{1 \leq j \leq q} \|W_i(j)\|_r.$$ 

Clearly, if the independent sequence $W_i$ satisfies assumption (19), then $\|\{W\}\|_{\psi,\nu} < \infty$ for all $\nu > 0$, in particular for $\nu = \infty$. Therefore, independence corresponds to $\nu = \infty$. As $\nu$ decreases to zero, the random vectors become more and more dependent.

Recall that

$$\|\hat{\Omega}_n - \Omega_n\|_\infty \leq \max_{1 \leq i, k \leq p + 1} \left\{\frac{1}{n} \sum_{i=1}^n \left( Z_i(j)Z_i(k) - \mathbb{E}[Z_i(j)Z_i(k)] \right) \right\},$$

which is a maximum of $(p + 1)^2$ many averages. To derive a bound on the quantity above, consider the following assumption:
(DEP) Assume that there exist \( n \) vector-valued functions \( G_i \) and an iid sequence \( \{ \varepsilon_i : i \in \mathbb{Z} \} \) such that
\[
Z_i := (X_i, Y_i) = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) \in \mathbb{R}^{p+1}.
\]

Also, for some \( \nu, \beta > 0 \),
\[
\|\{Z\}\|_{\psi_{\beta, \nu}} \leq K_{n,p} \quad \text{and} \quad \max_{1 \leq i \leq n} \max_{1 \leq j \leq p+1} \left| \mathbb{E}[Z_i(j)] \right| \leq K_{n,p}.
\]

Based on Remark 4.4, Assumption (DEP) is equivalent to the assumption of Lemma 5.1 for independent data. For independent random variables, the second part of Assumption (DEP) about the expectations follows from the \( \psi_{\beta, \nu} \)-bound assumption. The reason for this expectation bound in the assumption here is that the functional dependence measure \( \delta_{s, r} \) does not have any information about the expectation since
\[
\|W_i(j) - W_{i,s}(j)\|_p = \|\{W_i(j) - \mathbb{E}[W_i(j)]\} - \{W_{i,s}(j) - \mathbb{E}[W_{i,s}(j)]\}\|_p.
\]

The coupled random variable \( W_{i,s} \) has the same expectation as \( W_i \). Since the quantities we need to bound involve products of random variables, such a bound on the expectations is needed for our analysis.

The following result proves a bound on \( \hat{\Omega}_n - \Omega_n \|_{\infty} \) under assumption (DEP). Define
\[
\Upsilon_{4,p} := \max_{1 \leq i \leq p+1} \left( \|\{Z(j)\}\|_{4,0} + \max_{1 \leq i \leq n} \left| \mathbb{E}[Z_i(j)] \right| \right) \|\{Z(j)\}\|_{4,\nu}.
\]

**Theorem S.11.1.** Fix \( n, k \geq 1 \) and let \( t \geq 0 \) be any real number. Then under assumption (DEP), with probability at least \( 1 - 8e^{-t} \),
\[
\|\hat{\Omega}_n - \Omega_n\|_{\infty} \leq 2eB_{\nu} \sqrt{\Upsilon_{4,p}(t + \log(4p))} \frac{1}{n} + C_{\beta}K_{n,p}^2 \left( \log n \right)^{1/\beta/2} \kappa_n(\nu)(t + \log(4p))^{1/T_i(\beta/2)} \frac{1}{n},
\]
where \( T_i(\lambda) = \min\{\lambda, 1\} \), \( s(\lambda) = (1/2 + 1/\lambda)^{-1} \) and
\[
\kappa_n(\nu) = 2^\nu \times \begin{cases} 5(\nu - 1/2)^{-3}, & \text{if } \nu > 1/2, \\ 2(\log n)^{5/2}, & \text{if } \nu = 1/2, \\ 5(2n)^{1/2-\nu}(1/2 - \nu)^{-3}, & \text{if } \nu < 1/2. \end{cases}
\]

Here \( B_{\nu} \) and \( C_{\beta} \) are positive constants depending only on \( \nu \) and \( \beta \).

**Proof.** The proof follows from Lemma B.4 and Theorem 5.1 (or Theorem B.1) of Kuchibhotla et al. (2018).

**Remark S.11.2** (Rate of Convergence under Dependence) Theorem S.11.1 readily implies bounds on \( C_n^\Sigma(\alpha) \) and \( C_n^\Gamma(\alpha) \) along with rate bounds on \( \mathcal{D}_n^\Gamma \) and \( \mathcal{D}_n^\Sigma \). ⊙
References.


Department of Statistics, The Wharton School
University of Pennsylvania
3730 Walnut Street,
Jon M. Huntsman Hall
Philadelphia, PA 19104, USA.
E-mail: arunku@wharton.upenn.edu