Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems. ArXiv: 1304.0282

Victor Chernozhukov
MIT, Economics + Center for Statistics

Co-authors:
Alexandre Belloni (Duke) + Kengo Kato (Tokyo)

August 12, 2015

The presentation is based on:

"Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems"
Oberwolfach, 2012; ArXiv, 2013; published by Biometrika, 2014

1. Develop uniformly valid confidence regions for a target regression coefficient in a high-dimensional sparse median regression model (extends our earlier work in ArXiv 2010, 2011).

The presentation is based on:

"Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems"

Oberwolfach, 2012; ArXiv, 2013; published by Biometrika, 2014

1. Develop uniformly valid confidence regions for a target regression coefficient in a high-dimensional sparse median regression model (extends our earlier work in ArXiv 2010, 2011).
2. New methods are based on Z-estimation using scores that are Neyman-orthogonalized with respect to perturbations of nuisance parameters.

The presentation is based on:

"Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems"

Oberwolfach, 2012; ArXiv, 2013; published by Biometrika, 2014

1. Develop uniformly valid confidence regions for a target regression coefficient in a high-dimensional sparse median regression model (extends our earlier work in ArXiv 2010, 2011).
2. New methods are based on Z-estimation using scores that are Neyman-orthogonalized with respect to perturbations of nuisance parameters.
3. The estimator of a target regression coefficient is root-n consistent and asymptotically normal, **uniformly** with respect to the underlying sparse model, and is semi-parametrically efficient.

The presentation is based on:

"Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems"

Oberwolfach, 2012; ArXiv, 2013; published by Biometrika, 2014

1. Develop uniformly valid confidence regions for a target regression coefficient in a high-dimensional sparse median regression model (extends our earlier work in ArXiv 2010, 2011).
2. New methods are based on Z-estimation using scores that are Neyman-orthogonalized with respect to perturbations of nuisance parameters.
3. The estimator of a target regression coefficient is root-n consistent and asymptotically normal, **uniformly** with respect to the underlying sparse model, and is semi-parametrically efficient.
4. Extend methods and results to **general** Z-estimation problems with orthogonal scores and many target parameters $p_1 \gg n$, and construct *joint confidence rectangles* on all target coefficients and control *Family-Wise Error Rate*.

Outline

1. Z-problems like mean, median, logistic regressions and the associated scores
2. Problems with naive plug-in inference (where we plug-in regularized or post-selection estimators)
3. Problems can be fixed by using Neyman-orthogonal scores, which differ from original scores in most problems
4. Generalization to many target coefficients
5. Literature: orthogonal scores vs. debiasing
6. Conclusion

1. Z-problems

- Consider examples with y_i response, d_i the target regressor, and x_i covariates, with $p = dim(x_i) \gg n$
- Least squares projection:

$$\mathbb{E}[(y_i - d_i\alpha_0 - x_i'\beta_0)(d_i, x_i')'] = 0$$

- LAD regression:

$$\mathbb{E}[\{1(y_i \leq d_i\alpha_0 + x_i'\beta_0) - 1/2\}(d_i, x_i')'] = 0$$

- Logistic Regression:

$$\mathbb{E}[\{y_i - \Lambda(d_i\alpha_0 + x_i'\beta_0)\}w_i(d_i, x_i')'] = 0,$$

where $\Lambda(t) = \exp(t)/\{1 + \exp(t)\}$, $w_i = 1/\Lambda_i(1 - \Lambda_i)$, and $\Lambda_i = \Lambda(d_i\alpha_0 + x_i'\beta_0)$.

1. Z-problems

- In all cases have the Z-problem (focusing on a subset of equations that identify α_0 given β_0):

$$\mathbb{E}[\varphi(\underbrace{W}_{data}, \underbrace{\alpha_0}_{target\ \ parameter}, \underbrace{\beta_0}_{high-dim\ \ nuisance\ \ parameter})] = 0$$

with non-orthogonal scores (check!):

$$\partial_\beta \mathbb{E}[\varphi(W, \alpha_0, \beta)]\Big|_{\beta=\beta_0} \neq 0.$$

- Can we use plug-in estimators $\hat{\beta}$, based on regularization via penalization or selection, to form Z-estimators of α_0?

$$\mathbb{E}_n[\varphi(W, \hat{\alpha}, \hat{\beta})] = 0$$

1. Z-problems

- In all cases have the Z-problem (focusing on a subset of equations that identify α_0 given β_0):

$$\mathbb{E}[\varphi(\underbrace{W}_{data}, \underbrace{\alpha_0}_{target\ parameter}, \underbrace{\beta_0}_{high-dim\ nuisance\ parameter})] = 0$$

with non-orthogonal scores (check!):

$$\partial_\beta \mathbb{E}[\varphi(W, \alpha_0, \beta)]\Big|_{\beta=\beta_0} \neq 0.$$

- Can we use plug-in estimators $\hat{\beta}$, based on regularization via penalization or selection, to form Z-estimators of α_0?

$$\mathbb{E}_n[\varphi(W, \hat{\alpha}, \hat{\beta})] = 0$$

- The answer is NO!

2. Problems with naive plug-in inference: MC Example

- In this simulation we used: $p = 200, \quad n = 100, \quad \alpha_0 = .5$

$$y_i = d_i \alpha_0 + x_i' \beta_0 + \zeta_i, \quad \zeta_i \sim N(0,1)$$

$$d_i = x_i' \gamma_0 + v_i, \quad v_i \sim N(0,1)$$

- **approximately sparse model**

$$|\beta_{0j}| \propto 1/j^2, |\gamma_{0j}| \propto 1/j^2$$

$\rightarrow$ so can use L1-penalization

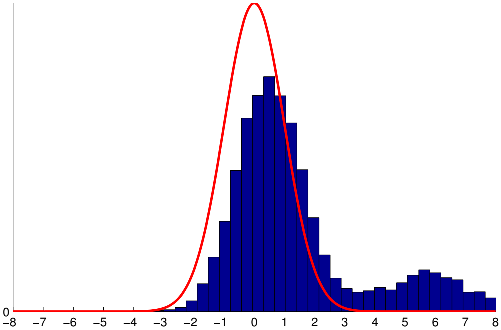
- $R^2 = .5$ **in each equation**
- regressors are correlated Gaussians:

$$x \sim N(0, \Sigma), \quad \Sigma_{kj} = (0.5)^{|j-k|}.$$

2.a. Distribution of The Naive Plug-in Z-Estimator

$$p = 200 \text{ and } n = 100$$

(the picture is roughly the same for median and mean problems)



$\Longrightarrow$ *badly biased, misleading confidence intervals;*
predicted by "impossibility theorems" in Leeb and Pötscher (2009)

2.b. Regularization Bias of The Naive Plug-in Z-Estimator

- $\hat\beta$ is a plug-in for β_0; bias in estimating equations:

$$\sqrt{n}\mathbb{E}\varphi(W,\alpha_0,\beta)\big|_{\beta=\hat\beta} = \overbrace{\sqrt{n}\mathbb{E}\varphi(W,\alpha_0,\beta_0)}^{=0}$$
$$+\underbrace{\partial_\beta \mathbb{E}\varphi(W,\alpha_0,\beta)\Big|_{\beta=\beta_0}\sqrt{n}(\hat\beta-\beta_0)}_{=:I\to\infty} + \underbrace{O(\sqrt{n}\|\hat\beta-\beta_0\|^2)}_{=:II\to 0}$$

- $II \to 0$ under sparsity conditions

$$\|\beta_0\|_0 \le s = o(\sqrt{n/\log p})$$

or approximate sparsity (more generally) since

$$\sqrt{n}\|\hat\beta-\beta_0\|^2 \lesssim \sqrt{n}(s/n)\log p = o(1).$$

- $I \to \infty$ generally, since

$$\sqrt{n}(\hat\beta-\beta_0) \sim \sqrt{s\log p} \to \infty,$$

  - due to non-regularity of $\hat\beta$, arising due to regularization via penalization or selection.

3. Solution: Solve Z-problems with Orthogonal Scores

- In all cases, it is possible to construct Z-problems

$$\mathbb{E}[\psi(\underbrace{W}_{data}, \underbrace{\alpha_0}_{target\ \ parameter}, \underbrace{\eta_0}_{high-dim\ \ nuisance\ \ parameter})] = 0$$

with Neyman-orthogonal (or "immunized") scores ψ:

$$\partial_\eta \mathbb{E}[\psi(W, \alpha_0, \eta)]\Big|_{\eta=\eta_0} = 0.$$

- Then we can simply use plug-in estimators $\hat{\eta}$, based on regularization via penalization or selection, to form Z-estimators of α_0:

$$\mathbb{E}_n[\psi(W, \check{\alpha}, \hat{\eta})] = 0.$$

3. Solution: Solve Z-problems with Orthogonal Scores

- In all cases, it is possible to construct Z-problems

$$\mathbb{E}[\psi(\underbrace{W}_{data}, \underbrace{\alpha_0}_{target\ \ parameter}, \underbrace{\eta_0}_{high-dim\ \ nuisance\ \ parameter})] = 0$$

with Neyman-orthogonal (or "immunized") scores ψ:

$$\partial_\eta \mathbb{E}[\psi(W, \alpha_0, \eta)]\Big|_{\eta=\eta_0} = 0.$$

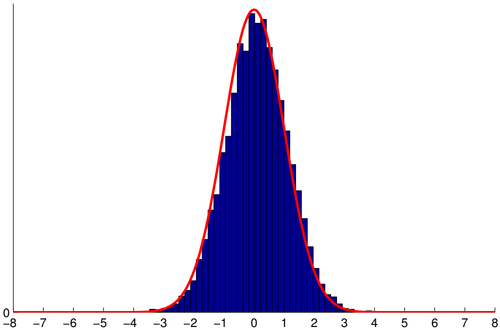
- Then we can simply use plug-in estimators $\hat{\eta}$, based on regularization via penalization or selection, to form Z-estimators of α_0:

$$\mathbb{E}_n[\psi(W, \check{\alpha}, \hat{\eta})] = 0.$$

- Note that $\varphi \neq \psi$ + extra nuisance parameters!

3.a. Distribution of the Z-Estimator with Orthogonal Scores

$$p = 200, \; n = 100$$



$\Longrightarrow$ *low bias, accurate confidence intervals*

obtained in a series of our papers, ArXiv, 2010, 2011, ...

3.b. Regularization Bias of The Orthogonal Plug-in Z-Estimator

- Expand the bias in estimating equations:

$$
\begin{aligned}
\sqrt{n}\mathbb{E}\psi(W,\alpha_0,\eta)\big|_{\eta=\hat{\eta}} &= \overbrace{\sqrt{n}\mathbb{E}\psi(W,\alpha_0,\eta_0)}^{=0} \\
&+ \underbrace{\partial_\eta \mathbb{E}\psi(W,\alpha_0,\eta)\Big|_{\eta=\eta_0}\sqrt{n}(\hat{\eta}-\eta_0)}_{=:I=0} + \underbrace{O(\sqrt{n}\|\hat{\eta}-\eta_0\|^2)}_{=:II\to 0}
\end{aligned}
$$

- $II \to 0$ under sparsity conditions

$$\|\eta_0\|_0 \leq s = o(\sqrt{n/\log p})$$

or approximate sparsity (more generally) since

$$\sqrt{n}\|\hat{\eta}-\eta_0\|^2 \lesssim \sqrt{n}(s/n)\log p = o(1).$$

- $I = 0$ by Neyman orthogonality.

3c. Theoretical result I

APPROXIMATE SPARSITY: after sorting absolute values of components of η_0 decay fast enough:

$$|\eta_0|_{(j)} \leq Aj^{-a}, \quad a > 1.$$

Theorem (BCK, Informal Statement)

Uniformly within a class of approximately sparse models with restricted isometry conditions

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0, 1),$$

where σ_n^2 is conventional variance formula for Z-estimators assuming η_0 is known. If the orthogonal score is efficient score, then $\check{\alpha}$ is semi-parametrically efficient.

3.d. Neyman-Orthogonal Scores

- In low-dimensional parametric settings, it was used by Neyman (56, 79) to deal with crudely estimated nuisance parameters. Frisch-Waugh-Lovell partialling out goes back to the 30s.
- Newey (1990, 1994), Van der Vaart (1990), Andrews (1994), Robins and Rotnitzky (1995), and Linton (1996) used orthogonality in semi parametric problems.
- For $p \gg n$ settings, Belloni, Chernozhukov, and Hansen (ArXiv 2010a,b) first used Neyman-orthogonality in the context of IV models. The η_0 was the parameter of the optimal instrument function, estimated by Lasso and OLS-post-Lasso methods

3.f. Examples of Orthogonal Scores: Least Squares

- Least squares:

$$\psi(W_i, \alpha, \eta_0) = \{\tilde{y}_i - \tilde{d}_i \alpha\} \tilde{d}_i,$$

$$\begin{aligned} y_i &= x_i' \eta_{10} + \tilde{y}_i, \quad \mathbb{E}[\tilde{y}_i x_i] = 0, \\ d_i &= x_i' \eta_{20} + \tilde{d}_i, \quad \mathbb{E}[\tilde{d}_i x_i] = 0. \end{aligned}$$

  Thus the orthogonal score is constructed by Frisch-Waugh partialling out from y_i and d_i. Here

$$\eta_0 := (\eta_{10}', \eta_{20}')'$$

  can be estimated by sparsity based methods, e.g. OLS-post-Lasso.
  Semi-parametrically efficient under homoscedasticity.
- Reference: Belloni, Chernozhukov, Hansen (ArXiv, 2011a,b).

3.f. Examples of Orthogonal Scores: LAD regression

- LAD regression:

$$\psi(W_i, \alpha, \eta_0) = \{1(y_i \leq d_i\alpha + x_i'\beta_0) - 1/2\}\tilde{d}_i,$$

where

$$f_i d_i = f_i x_i'\gamma_0 + \tilde{d}_i, \qquad \mathbb{E}[\tilde{d}_i f_i x_i] = 0,$$
$$f_i := f_{y_i|d_i,x_i}(d_i\alpha_0 + x_i'\beta_0 \mid d_i, x_i).$$

Here

$$\eta_0 := (f_{y_i|d_i,x_i}(\cdot), \alpha_0', \beta_0', \gamma_0')'$$

can be estimated by sparsity based methods, by L1-penalized LAD and by OLS-post-Lasso. Semi-parametrically efficient.

- Reference: Belloni, Chernozhukov, Kato (ArXiv, 2013a,b).

3.f. Examples of Orthogonal Scores: Logistic regression

- Logistic regression,

$$\psi(W_i, \alpha, \eta_0) = \{y_i - \Lambda(d_i\alpha + x_i'\beta_0)\}\tilde{d}_i/\sqrt{w_i},$$

$$\sqrt{w_i}d_i = \sqrt{w_i}x_i'\gamma_0 + \tilde{d}_i, \quad \mathbb{E}[\sqrt{w_i}\tilde{d}_i x_i] = 0,$$

$$w_i = \Lambda(d_i\alpha_0 + x_i'\beta_0)(1 - \Lambda(d_i\alpha_0 + x_i'\beta_0))$$

  Here

$$\eta_0 := (\alpha_0', \beta_0', \gamma_0')'$$

  can be estimated by sparsity based methods, by L1-penalized logistic regression and by OLS-post-Lasso.
  Semi-parametrically efficient.

- Reference: Belloni, Chernozhukov, Ying (ArXiv, 2013).

4. Generalization: Many Target Parameters

- Consider many Z-problems

$$\mathbb{E}[\psi_j(\underbrace{W_j}_{data}, \underbrace{\alpha_{j0}}_{target\ \ parameter}, \underbrace{\eta_{j0}}_{high-dim\ \ nuisance\ \ parameter})] = 0$$

with Neyman-orthogonal (or "immunized") scores:

$$\partial_{\eta_j} \mathbb{E}[\psi_j(W, \alpha_{j0}, \eta_j)]\Big|_{\eta_j = \eta_{j0}} = 0$$

$$j = 1, ..., p_1 \gg n.$$

- The can simply use plug-in estimators $\hat{\eta}_j$, based on regularization via penalization or selection, to form Z-estimators of α_{j0}:

$$\mathbb{E}_n[\psi_j(W, \check{\alpha}_j, \hat{\eta}_j)] = 0, \quad j = 1, ..., p_1.$$

4. Generalization: Many Target Parameters

Theorem (BCK, Informal Statement)

Uniformly within a class of approximately sparse models with restricted isometry conditions holding uniformly in $j = 1, ..., p_1$ and $(\log p_1)^7 = o(n)$,

$$\sup_{R \in \text{rectangles in } \mathbb{R}^{p_1}} |\mathrm{P}(\{\sigma_{jn}^{-1}\sqrt{n}(\check{\alpha}_j - \alpha_{j0})\}_{j=1}^{p_1} \in R) - \mathrm{P}(\mathcal{N} \in R)| \to 0,$$

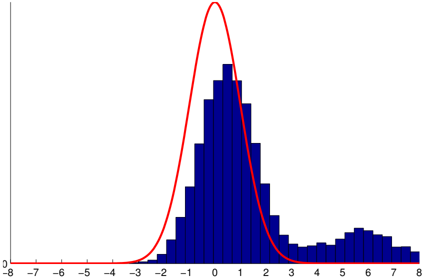
where σ_{jn}^2 is conventional variance formula for Z-estimators assuming η_{j0} is known, and $\mathcal{N}$ is the normal random p_1-vector that has mean zero and matches the large sample covariance function of $\{\sigma_{jn}^{-1}\sqrt{n}(\check{\alpha}_j - \alpha_{j0})\}_{j=1}^{p_1}$. Moreover, we can estimate $\mathrm{P}(\mathcal{N} \in R)$ by **Multiplier Bootstrap**.

- These results allow construction of simultaneous confidence rectangles on all target coefficients as well as control of the family-wise-error rate (FWER) in hypothesis testing.
- Rely on Gaussian Approximation Results and Multiplier Bootstrap proposed in Chernozhukov, Chetverikov, Kato (ArXiv 2012, 2013).

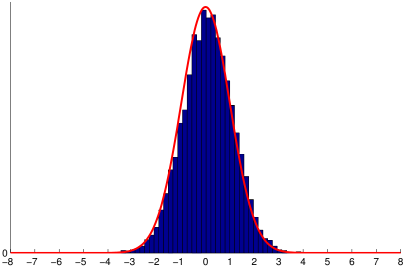
5. Literature: Neyman-Orthogonal Scores vs. Debiasing

- ArXiv 2010-2011 – use of orthogonal scores linear models
    a. Belloni, Chernozhukov, Hansen (ArXiv, 2010a, 2010b ,2011a, 2011b): use OLS-post-Lasso methods to estimate nuisance parameters in instrumental and mean regression;
    b. Zhang and Zhang (ArXiv, 2011): introduces debiasing + use Lasso methods to estimate nuisance parameters in mean regression;
- ArXiv 2013-2014 – non-linear models
    c. Belloni, Chernozhukov, Kato (ArXiv, 2013), Belloni, Chernozhukov, Wang(ArXiv, 2013);
    d. Javanmard and Montanari (ArXiv, 2013 a,b); van de Geer and co-authors (ArXiv, 2013);
    e. Han Liu and co-authors (ArXiv 2014)
- [b,d] introduce de-biasing of an initial estimator $\hat{\alpha}$. We can interpret "debiased" estimators= Bickel's "one-step" correction of an initial estimator in Z-problems with Neyman-orthogonal scores. They are first-order-equivalent to our estimators.

Conclusion



Without Orthogonalization



With Orthogonalization

Bibliography

- Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain, Alexandre Belloni, Daniel Chen, Victor Chernozhukov, Christian Hansen (arXiv 2010, Econometrica)
- LASSO Methods for Gaussian Instrumental Variables Models, Alexandre Belloni, Victor Chernozhukov, Christian Hansen (arXiv 2010)
- Inference for High-Dimensional Sparse Econometric Models, Alexandre Belloni, Victor Chernozhukov, Christian Hansen (arXiv 2011, World Congress of Econometric Society, 2010)
- Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models, Cun-Hui Zhang, Stephanie S. Zhang (arXiv 2011, JRSS(b))
- Inference on Treatment Effects After Selection Amongst High-Dimensional Controls, Alexandre Belloni, Victor Chernozhukov, Christian Hansen (arXiv 2011, Review of Economic Studies)

Bibliography

- Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors, Victor Chernozhukov, Denis Chetverikov, Kengo Kato (arXiv 2012, Annals of Statistics)
- Comparison and anti-concentration bounds for maxima of Gaussian random vectors Victor Chernozhukov, Denis Chetverikov, Kengo Kato (arXiv 2013, Probability Theory and Related Fields)
- Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems, Alexandre Belloni, Victor Chernozhukov, Kengo Kato (arXiv 2013, oberwolfach 2012, Biometrika)
- On asymptotically optimal confidence regions and tests for high-dimensional models, Sara van de Geer, Peter Bhlmann, Ya'acov Ritov, Ruben Dezeure (arXiv 2013, Annals of Statistics)

Bibliography

- Honest Confidence Regions for a Regression Parameter in Logistic Regression with a Large Number of Controls, Alexandre Belloni, Victor Chernozhukov, Ying Wei (ArXiv 2013)
- Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models, Alexandre Belloni, Victor Chernozhukov, Kengo Kato (ArXiv 2013)
- Confidence Intervals and Hypothesis Testing for High-Dimensional Regression, Adel Javanmard and Andrea Montanari (arXiv 2013, J. Mach. Learn. Res.)
- Pivotal estimation via square-root Lasso in nonparametric regression Alexandre Belloni, Victor Chernozhukov, Lie Wang, (arXiv 2013, Annals of Statistics)

Bibliography

- Program Evaluation with High-Dimensional Data, Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, Chris Hansen (arXiv 2013)
- A General Framework for Robust Testing and Confidence Regions in High-Dimensional Quantile Regression, Tianqi Zhao, Mladen Kolar and Han Liu (arXiv 2014)
- A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models, Yang Ning and Han Liu (arXiv 2014)
- Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach (Annual Review of Economics 2015), Victor Chernozhukov, Christian Hansen, Martin Spindler