Confidence Sets Based on Shrinkage Estimators

Mikkel Plagborg-Møller

April 12, 2017

Shrinkage estimators in applied work

$$\hat{\beta}_{\text{shrink}} = \text{argmin}_{\beta} \left\{ \hat{Q}(\beta) + \lambda C(\beta) \right\}$$

- Shrinkage/penalized estimators popular in economics:
  - Random effects.
  - High-dimensional prediction.
  - Smoothing jagged functions. Shiller (1973); Barnichon & Brownlees (2017)
  - Estimating fixed effects. Chetty et al. (2014); Chamberlain (2016)
  - Shrinking toward theory. Hansen (2016); Fessler & Kasy (2017)
- Shrinkage parameter λ often data-dependent.

Challenges of shrinkage inference

- How to "calculate SEs" for shrinkage estimators?
- With data-dependent shrinkage parameter λ, asy. distribution often discontinuous in true parameters.
- Impossible to estimate CDF of $\hat{\beta}_{\text{shrink}}$ *uniformly* consistently.
  Leeb & Pötscher (2005)
- Standard bootstrap typically doesn't work. Beran (2010)
- Applied researchers often just undersmooth (i.e., SE for usual point estimator). Not always valid.

This project

- Class of generalized ridge regression estimators: Vinod (1978)

$$\hat{\beta}_{M,W}(\lambda) = \text{argmin}_{\beta \in \mathbb{R}^n} \left\{ \|\beta - \hat{\beta}\|_W^2 + \lambda \|M\beta\|^2 \right\}.$$

  Shrinkage parameter λ selected by unbiased risk estimate.

- Gaussian location model: $\hat{\beta} \sim N_n(\beta^\dagger, \Sigma)$, known Σ.
  - Conditional QLR test for linear hypothesis on $\beta^\dagger$. Exact size.
  - Conditional QLR confidence region by test inversion.
  - Simulations show favorable average length of CIs.
- Uniform asymptotic validity even when data is non-Gaussian.

Relationship to literature

- Large stats lit uses analytically convenient transformations and priors.
  Casella & Hwang (1982, 1984, 1987, 2012); Tseng & Brown (1997)
- My starting point: How to "calculate SEs" for given ridge estimator? Arbitrary correlation structure, arbitrary shrinkage hypothesis.
- CSs tied to (and always contain) meaningful point estimator.
- Tests/CSs have Empirical Bayes (random effects) interpretation. But I do *not* start from decision-theoretic first principles.
- Impossible to *uniformly* dominate expected volume of Wald ellipsoid for 1-D or 2-D problems. Stein (1962); Brown (1966); Joshi (1969)

Other related literature

- Shrinkage: Stein (1956); James & Stein (1961)
- Projection shrinkage: Bock (1975); Oman (1982); Casella & Hwang (1987)
- Unbiased risk estimate: Mallows (1973); Stein (1973, 1981); Berger (1985); Claeskens & Hjort (2003); Hansen (2010)
- Asymptotics for shrinkage: Hansen (2016)
- Uniform inference: Andrews et al. (2011); McCloskey (2015)
- Post-regularization inference: Chernozhukov et al. (2015)

Outline

1. Shrinkage estimators and Unbiased Risk Estimate
2. Testing
3. Confidence sets (and simulations)
4. Uniform asymptotic validity
5. Summary and next steps

Gaussian location model

- For now, consider finite-sample Gaussian location model

$$\hat{\beta} \sim N_n(\beta^\dagger, \Sigma).$$

- $\beta^\dagger \in \mathbb{R}^n$ unknown.
- Σ symmetric p.d. and known.
- Will later consider asymptotic framework for which the Gaussian model is the right limit experiment. Plug in consistent estimator $\hat{\Sigma}$.

General shrinkage estimator class

$$\hat{\beta}_{M,W}(\lambda) = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \|\beta - \hat{\beta}\|_W^2 + \lambda \|M\beta\|^2 \right\} = \Theta_{M,W}(\lambda)\hat{\beta},$$

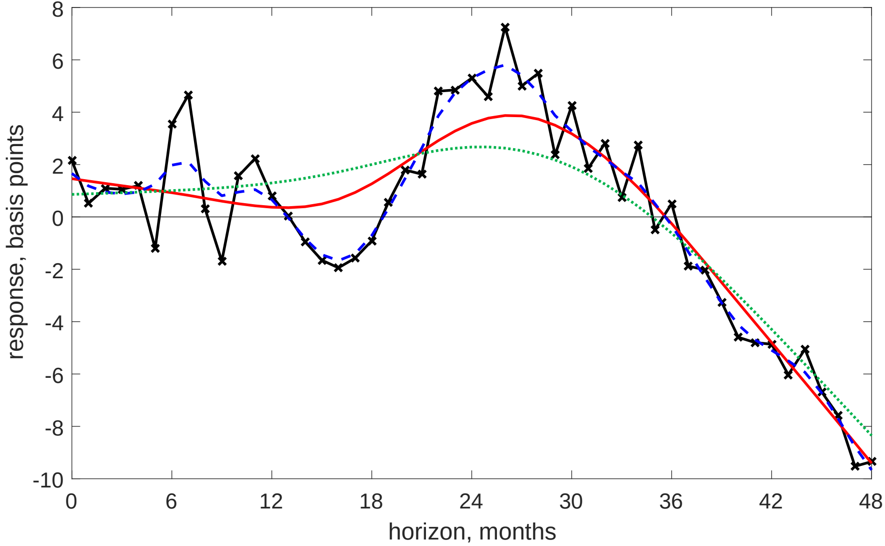
$$\Theta_{M,W}(\lambda) = (I_n + \lambda W^{-1} M' M)^{-1}.$$

- $M \in \mathbb{R}^{m \times n}$, $W \in \mathbb{R}^{n \times n}$ symmetric p.d.
- Example:

$$M = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{(n-2) \times n}.$$

Penalizes jaggedness.

Whittaker (1923); Shiller (1972); Hodrick & Prescott (1981); Wahba (1990)



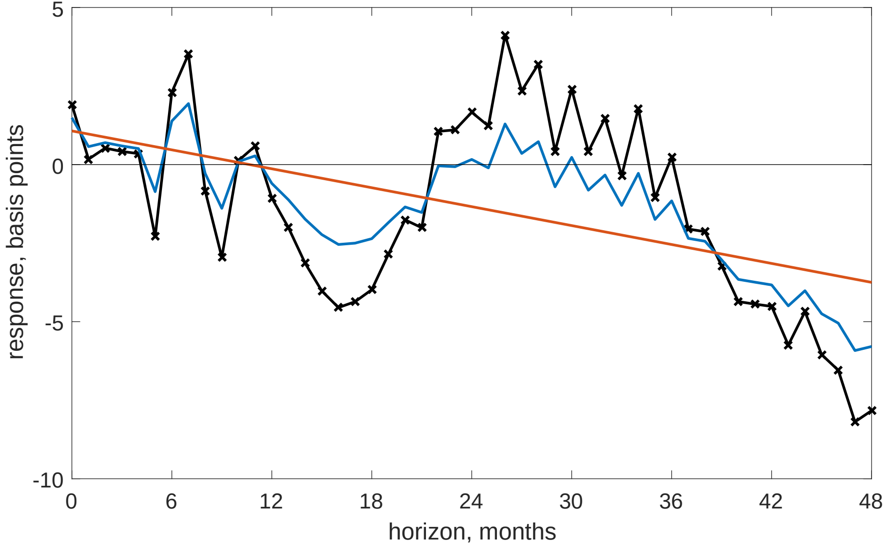
y_t: GZ excess bond premium. x_t: high-freq. FFF shock.
Controls: 2 lags of y_t, x_t, log(CPI), log(IP), 1yrTreas. Sample: 1991–2012.

Projection shrinkage

- Shrinkage particularly tractable when $W = I_n$ and $M = P \in \mathbb{R}^{n\times n}$ is orthogonal projection matrix: $P = P' = P^2$.
- Projection shrinkage towards linear subspace $\text{span}(I_n - P)$.
  Stein (1956); Oman (1982a,b); Bock (1985); Casella & Hwang (1987)

$$\hat{\beta}_P(\lambda) = \underset{\beta\in\mathbb{R}^n}{\text{argmin}} \left\{ \|\beta - \hat{\beta}\|^2 + \lambda \|P\beta\|^2 \right\}$$
$$= \frac{1}{1+\lambda} P\hat{\beta} + (I_n - P)\hat{\beta}.$$

- Example: $I_n - P =$ proj. matrix from regression onto basis functions.



y_t: GZ excess bond premium. x_t: high-freq. FFF shock.
Controls: 2 lags of y_t, x_t, log(CPI), log(IP), 1yrTreas. Sample: 1991–2012.

Unbiased Risk Estimate

- MSE risk criterion:

$$R_{M,W}(\lambda) = E\left(\|\hat{\beta}_{M,W}(\lambda) - \beta^{\dagger}\|_W^2\right).$$

- Unbiased Risk Estimate (URE): ▸ Bias/var.

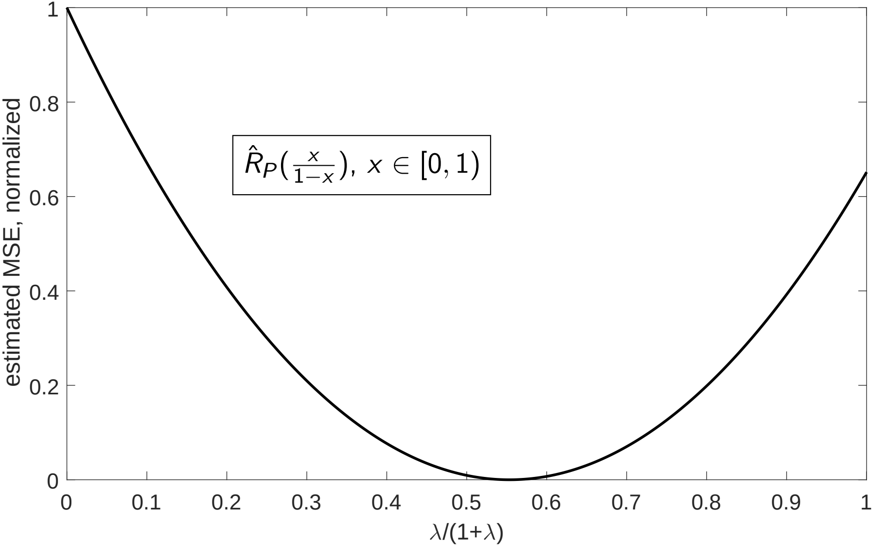
  Mallows (1973); Stein (1973, 1981); Berger (1985); Hansen (2010)

$$\hat{R}_{M,W}(\lambda) = \|\hat{\beta}_{M,W}(\lambda) - \hat{\beta}\|_W^2 + 2\,\mathrm{tr}\{W\Theta_{M,W}(\lambda)\Sigma\}.$$

- If $\mathrm{rk}(M) = m$ or $M = P$, URE is strictly convex in $\frac{\lambda}{1+\lambda}$.
- Define

$$\hat{\lambda}_{M,W} = \operatorname*{argmin}_{\lambda \geq 0} \hat{R}_{M,W}(\lambda).$$

  May equal ∞. $\lim_{\lambda\to\infty} \hat{\beta}_{M,W}(\lambda)$ well defined if M full rank or proj.



y_t: GZ excess bond premium. x_t: high-freq. FFF shock.
Controls: 2 lags of y_t, x_t, log(CPI), log(IP), 1yrTreas. Sample: 1991–2012.

Optimal projection shrinkage

- For projection shrinkage, can minimize URE in closed form:

$$\hat{\beta}_P(\hat{\lambda}_P) = \left(1 - \frac{\operatorname{tr}(\Sigma_P)}{\|P\hat{\beta}\|^2}\right)_+ P\hat{\beta} + (I_n - P)\hat{\beta}, \quad \Sigma_P = P\Sigma P.$$

- James-Stein shrinkage towards linear subspace.
  Stein (1956); James & Stein (1961); Oman (1982a,b); Bock (1985)

Optimal projection shrinkage

- For projection shrinkage, can minimize URE in closed form:

$$\hat{\beta}_P(\hat{\lambda}_P) = \left(1 - \frac{\operatorname{tr}(\Sigma_P)}{\|P\hat{\beta}\|^2}\right)_+ P\hat{\beta} + (I_n - P)\hat{\beta}, \quad \Sigma_P = P\Sigma P.$$

- James-Stein shrinkage towards linear subspace.
  Stein (1956); James & Stein (1961); Oman (1982a,b); Bock (1985)

- Proposition (Hansen, 2016): If $\operatorname{tr}(\Sigma_P) > 4\rho(\Sigma_P)$,

$$E_{\beta^\dagger}\left(\|\hat{\beta}_P(\hat{\lambda}_P) - \beta^\dagger\|^2\right) < E_{\beta^\dagger}\left(\|\hat{\beta} - \beta^\dagger\|^2\right) \quad \text{for all } \beta^\dagger.$$

- Necessary cond'n: $\operatorname{rk}(P) > 4$. E.g., if $I_n - P$ is projection onto p basis functions, then need $n > p + 4$.

Outline

1. Shrinkage estimators and Unbiased Risk Estimate
2. Testing
3. Confidence sets (and simulations)
4. Uniform asymptotic validity
5. Summary and next steps

Hypothesis testing in shrinkage applications

$$H_0 \colon R\beta^\dagger = b, \quad H_1 \colon R\beta^\dagger \neq b.$$

- $R \in \mathbb{R}^{r \times n}$ full row rank.
- No UMP test exists.
- Usual Wald test is UMP unbiased/invariant and admissible.
- If we're already using shrinkage point estimator, might want hypothesis test tied to this estimator as well. Obtain CS by inversion.
- My proposed test is biased+noninvariant, so *may* achieve higher power than usual Wald test for *some* DGPs.

Empirical Bayes quasi-likelihood ratio test

- Base hypothesis test on (negative) quasi-log-likelihood

$$\hat{L}_{M,W}(\beta) = \|\beta - \hat{\beta}\|_W^2 + \hat{\lambda}_{M,W}\|M\beta\|^2.$$

- Empirical Bayes (random effects) interpretation:

$$\beta^\dagger \mid \text{data} \sim N\left(\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}), (W + \hat{\lambda}_{M,W}M'M)^{-1}\right).$$

- QLR test statistic of $R\beta^\dagger = b$:

$$\begin{aligned}\min_{\beta:\, R\beta=b} \hat{L}_{M,W}(\beta) - \min_{\beta} \hat{L}_{M,W}(\beta) \\ = \|R\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - b\|^2_{(R(W+\hat{\lambda}_{M,W}M'M)^{-1}R')^{-1}}\end{aligned}$$

Null distribution impractical

$$\widehat{LR}_{M,W}(b) = \|R\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - b\|^2_{(R(W+\hat{\lambda}_{M,W}M'M)^{-1}R')^{-1}}$$

- Assume $\text{Var}(RZ \mid MZ)$ nonsingular, $Z \sim N_n(0, W^{-1})$. Then $\widehat{LR}$ well defined even when $\hat{\lambda}_{M,W} = \infty$. Holds in many cases.
- If $\text{Var}(RZ \mid MZ)$ singular, can use *ad hoc* LR statistic

$$\widetilde{LR}_{M,W}(b) = \|R\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - b\|^2_{(RW^{-1}R')^{-1}}.$$

Null distribution impractical

$$\widehat{LR}_{M,W}(b) = \|R\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - b\|^2_{(R(W+\hat{\lambda}_{M,W}M'M)^{-1}R')^{-1}}$$

- Assume $\text{Var}(RZ \mid MZ)$ nonsingular, $Z \sim N_n(0, W^{-1})$. Then $\widehat{LR}$ well defined even when $\hat{\lambda}_{M,W} = \infty$. Holds in many cases.
- If $\text{Var}(RZ \mid MZ)$ singular, can use *ad hoc* LR statistic

$$\widetilde{LR}_{M,W}(b) = \|R\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - b\|^2_{(RW^{-1}R')^{-1}}.$$

- Practical problem: Null distribution of LR statistic depends on entire n-dimensional parameter vector $\beta^\dagger$.
- Proposed solution: Condition on sufficient statistic for $n - r$ nuisance parameters. Andrews & Mikusheva (2016)

Sufficient statistic for nuisance parameters

- Define $\zeta = \Sigma R'(R\Sigma R')^{-1} \in \mathbb{R}^{n\times r}$ and $\tilde{P} = \zeta R \in \mathbb{R}^{n\times n}$.
- Statistic $\hat{\nu} = (I_n - \tilde{P})\hat{\beta}$ is "S-ancillary" wrt. $R\beta^\dagger$:

$$\hat{\beta} \mid \hat{\nu} \sim F_{R\beta^\dagger,\Sigma}, \quad \hat{\nu} \sim F_{(I_n-\tilde{P})\beta^\dagger,\Sigma}.$$

- It would be uncontroversial to condition on $\hat{\nu}$ in the *absence* of prior information linking $R\beta^\dagger$ and $(I_n - \tilde{P})\beta^\dagger$.
- In practice, the prior information $\|M\beta^\dagger\| \ll 1$ may not substantially constrain the relationship between $R\beta^\dagger$ and $(I_n - \tilde{P})\beta^\dagger$. Then conditioning wastes little information. Severini (1995)
- I condition on $\hat{\nu}$.

Critical value by simulation

- Conditional QLR test rejects H_0 if

$$\widehat{LR}_{M,W}(b) > q_{1-\alpha,M,W}(b,\hat{\nu}).$$

- Conditional critical value given $\hat{\nu} = \nu$:

$$q_{1-\alpha,M,W}(b,\nu) = \text{quantile}_{1-\alpha}\left(\|R\tilde{\beta}(\tilde{\lambda};U) - b\|^2_{(R(W+\tilde{\lambda}(U)M'M)^{-1}R')^{-1}}\right),$$

where

$$U \sim N_r(b, R\Sigma R'),$$

$$\tilde{\beta}(\lambda;U) = \Theta_{M,W}(\lambda)(\zeta U + \nu) \quad \text{for all } \lambda \geq 0,$$

$$\tilde{\lambda}(U) = \underset{\lambda \geq 0}{\text{argmin}}\left\{\|\tilde{\beta}(\lambda;U) - (\zeta U + \nu)\|^2_W + 2\,\text{tr}(W\Theta_{M,W}(\lambda)\Sigma)\right\}.$$

- By design, conditional (and thus unconditional) size $= 1 - \alpha$.

Outline

1. Shrinkage estimators and Unbiased Risk Estimate
2. Testing
3. Confidence sets (and simulations)
4. Uniform asymptotic validity
5. Summary and next steps

Confidence set by test inversion

- Invert CQLR test to obtain CS for $b^\dagger = R\beta^\dagger$:

$$\widehat{C}_{M,W} = \left\{ b \in \mathbb{R}^r \colon \widehat{LR}_{M,W}(b) \leq q_{1-\alpha,M,W}(b, \hat{\nu}) \right\}.$$

- Do this by grid search. Simulate quantile at each point. Feasible in one or two dimensions (proj. shrinkage fast). ▸ Uniform band
- If M full rank or proj., can compute simple, finite upper bound on critical value. ▸ More
  - $\widehat{C}_{M,W}$ contained in bounded ellipsoid centered at $R\hat{\beta}_{M,W}(\hat{\lambda}_{M,W})$.
  - Limits grid search.

Properties of shrinkage confidence set

1. $\widehat{C}_{M,W}$ always contains shrinkage point estimate.
2. Generally not symmetric around point estimate.
3. Empirical Bayes intuition: CS should have small volume for DGPs where shrinkage estimator has low MSE.
4. Appears to not always be convex in simulations.
5. Converges a.s. to usual Wald ellipsoid as $\|M\beta^{\dagger}\| \to \infty$, M fixed.

- Appears difficult to characterize expected volume. Even for projection shrinkage, *conditional* power of CQLR test depends on 6 parameters.

Simulation study of confidence intervals

$$\hat{\beta} \sim N_n(\beta^\dagger, \Sigma),$$

$$\beta_i^\dagger = \begin{cases} 1 - \frac{i-1}{n-1} & \text{if } K = 0, \\ \sin \frac{2\pi K(i-1)}{n-1} & \text{if } K > 0, \end{cases}$$

$$\Sigma_{ij} = \sigma_i \sigma_j \kappa^{|i-j|}, \quad \sigma_i = \sigma_0 \left(1 + (i-1)\frac{\varphi - 1}{n-1}\right).$$

- Consider projection shrinkage toward quadratic polynomial.
- Lower bound on expected length relative to Wald CI: Pratt (1961)

$$\frac{(1-\alpha)\Phi^{-1}(1-\alpha) + (2\pi)^{-1/2} e^{-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2}}{\Phi^{-1}(1-\alpha/2)} \approx 0.808 \text{ for } \alpha = 0.1.$$

					MSE $\hat{\beta}(\hat{\lambda})$			Length $\hat{C}$	
n	K	κ	σ_0	φ	Tot	1st	Mid	1st	Mid
10	0.5	0.5	0.25	1	0.63	0.95	0.56	0.97	0.85
25	0.5	0.5	0.25	1	0.34	0.69	0.29	0.88	0.86
50	0.5	0.5	0.25	1	0.19	0.46	0.16	0.83	0.88
25	0	0.5	0.25	1	0.34	0.68	0.29	0.87	0.86
25	1	0.5	0.25	1	0.93	1.29	0.77	1.10	0.88
25	2	0.5	0.25	1	0.96	0.93	0.86	0.98	0.90
25	0.5	0	0.25	1	0.16	0.35	0.13	0.83	0.84
25	0.5	0.9	0.25	1	0.81	1.11	0.76	1.05	0.91
25	0.5	0.5	0.5	1	0.34	0.66	0.28	0.88	0.86
25	0.5	0.5	0.25	3	0.35	1.19	0.30	0.96	0.85

MSE relative to $\hat{\beta}$, average length relative to Wald.
Level $=$ 90%. 1st $= \beta_1$, Mid $= \beta_{1+[n/2]}$.

Takeaways from simulation

- $\beta_{1+[n/2]}$: Expected length of CI close to performance limit.
- β_1: Expected length competitive with Wald CI, but sometimes slightly wider. Intuition: Fewer relevant parameters to average across.
- Shrinkage works less well when...

1. n is small.
2. Shrinkage hypothesis $M\beta^\dagger = 0$ is neither approximately true nor dramatically false.
3. Correlations are high.
4. Variance of MLE of nuisance parameters large relative to variance of MLE of parameter of interest.

Empirical Bayes HPD set

$$\hat{L}_{M,W}(\beta) = \|\beta - \hat{\beta}\|_W^2 + \hat{\lambda}_{M,W}\|M\beta\|^2,$$

$$\beta^\dagger \mid \text{data} \sim N\left(\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}), (W + \hat{\lambda}_{M,W}M'M)^{-1}\right).$$

- "Empirical Bayes" $1-\alpha$ Highest Posterior Density set for $R\beta^\dagger$:

$$\widehat{C}_{EB} = \left\{b \in \mathbb{R}^r \colon \widehat{LR}_{M,W}(b) \leq \chi^2_{r,1-\alpha}\right\}.$$

- Doesn't control frequentist coverage.
- Like shrinkage CS, but non-random critical value.

Minimum coverage discrepancy with EB HPD set

- Symmetric set difference: $A \triangle B = (A \cup B) \backslash (A \cap B)$.

Proposition (following Andrews & Mikusheva, 2016)

Let $\widetilde{C}$ be any similar confidence set for $R\beta^\dagger$ (like $\widehat{C}_{M,W}$):

$$P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C}\right) = 1 - \alpha \text{ for all } \beta^\dagger \in \mathbb{R}^n.$$

Then

$$P_{\beta^\dagger}\left(R\beta^\dagger \in \widehat{C}_{M,W} \triangle \widehat{C}_{EB}\right) \leq P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \triangle \widehat{C}_{EB}\right) \text{ for all } \beta^\dagger \in \mathbb{R}^n.$$

▸ Proof

Outline

1. Shrinkage estimators and Unbiased Risk Estimate
2. Testing
3. Confidence sets (and simulations)
4. Uniform asymptotic validity
5. Summary and next steps

Uniform asymptotic size control

- CQLR test achieves uniform asymptotic size control, provided $\hat{\beta}$ is uniformly asy. normal, and $\hat{\Sigma}$ is uniformly consistent for Σ.
- Uniform frequentist validity stands in stark contrast to other approaches.
  - Undersmoothing: Pretend λ is "small", ignore bias of shrinkage estimator as well as variability in λ.
  - Switching rule: Use Wald SE if $\|M\hat{\beta}\| > c$, otherwise use asymptotics under assumption $M\beta^\dagger = 0$.
  - Random effects: Treat random effects assumption as part of the DGP rather than just a prior. Size control wrt. random effects distribution.

Assumption: Preliminary estimator well-behaved

Assumption

Define $\mathcal{S} = \{A \in \mathbb{S}^n_+ : \underline{c} \leq 1/\rho(A^{-1}) \leq \rho(A) \leq \overline{c}\}$ *for fixed* $\underline{c}, \overline{c} > 0$.
The distribution of the data F_T *for sample size* T *is indexed by three parameters* $\beta \in \mathcal{B} \subset \mathbb{R}^n$, $\Sigma \in \mathcal{S}$, *and* $\gamma \in \Gamma$.

The estimators $(\hat{\beta}, \hat{\Sigma}) \in \mathbb{R}^n \times \mathbb{S}^n_+$ *satisfy the following:*

For all sequences $\{\beta_T, \Sigma_T, \gamma_T\}_{T \geq 1} \in \mathcal{B} \times \mathcal{S} \times \Gamma$ *and all subsequences* $\{k_T\}_{T \geq 1}$ *of* $\{T\}_{T \geq 1}$,

$$\sqrt{k_T}\hat{\Sigma}^{-1/2}(\hat{\beta} - \beta_{k_T}) \underset{F_{k_T}(\beta_{k_T}, \Sigma_{k_T}, \gamma_{k_T})}{\overset{d}{\longrightarrow}} N_n(0, I_n),$$

$$(\hat{\Sigma} - \Sigma_{k_T}) \underset{F_{k_T}(\beta_{k_T}, \Sigma_{k_T}, \gamma_{k_T})}{\overset{p}{\longrightarrow}} 0, \quad \text{as } T \to \infty.$$

$\mathbb{S}^n$ = set of symmetric positive definite $n \times n$ matrices.

Shrinkage test is uniformly valid

- Let $\widehat{LR}$ and $\hat{q}_{1-\alpha}$ denote CQLR test statistic and quantile obtained by plugging in $T^{-1}\hat{\Sigma}$ in place of Σ. (Suppress M, W.)

Proposition

Let the previous assumption hold. Assume either $\mathrm{rk}(M) = m$ *or* $M = P$. *Assume also* $\mathrm{Var}(RZ \mid MZ)$ *is nonsingular,* $Z \sim N_n(0, W^{-1})$. *Then*

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\left(\widehat{LR}(R\beta) \leq \hat{q}_{1-\alpha}(R\beta, \hat{\nu})\right) = 1 - \alpha.$$

Shrinkage test is uniformly valid

- Let $\widehat{LR}$ and $\hat{q}_{1-\alpha}$ denote CQLR test statistic and quantile obtained by plugging in $T^{-1}\hat{\Sigma}$ in place of Σ. (Suppress M, W.)

Proposition

Let the previous assumption hold. Assume either $\text{rk}(M) = m$ *or* $M = P$*. Assume also* $\text{Var}(RZ \mid MZ)$ *is nonsingular,* $Z \sim N_n(0, W^{-1})$*. Then*

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\left(\widehat{LR}(R\beta) \le \hat{q}_{1-\alpha}(R\beta,\hat{\nu})\right) = 1-\alpha.$$

- *Caveat*: I have only written down the full proof for proj. shrinkage. I believe I have the arguments worked out for the general case.
- Proof idea: Consider drifting parameters β_T...
  1. If $\|\sqrt{T}M\beta_T\| \to \infty$, we converge to non-shrinkage case.
  2. If $\sqrt{T}M\beta_T$ is bounded, we're in the Gaussian model in the limit.

Outline

1. Shrinkage estimators and Unbiased Risk Estimate
2. Testing
3. Confidence sets (and simulations)
4. Uniform asymptotic validity
5. Summary and next steps

Summary

- Considered setting where generalized ridge regression point estimator is of interest: smoothing, shrinking toward average, penalization, etc.
- Proposed conditional QLR test based on same quasi-log-likelihood as shrinkage point estimator.
  - Exact conditional size in Gaussian location model.
  - Asymptotic uniform size control more generally.
- Shrinkage confidence set by test inversion.
  - Contains shrinkage point estimate.
  - Minimum "coverage discrepancy" with EB HPD set among similar CSs.
  - Computationally feasible in 1–2 dimensions. Proj. shrinkage fast.
  - Promising simulation evidence.

Next steps

- More simulation evidence.
- Comparison of 2-D ellipse with infeasible optimum.
- Empirics: impulse responses, MIDAS, exchangeable parameters, ...?
- Analytical/low-dimensional power/volume comparisons. Probably only feasible for special cases, e.g., $\Sigma = I_n$.

Thank you

URE captures bias/variance tradeoff

- $W = I_n$ for simplicity.
- Risk decomposition: Claeskens & Hjort (2003)

$$R_{M,I_n}(\lambda) = \underbrace{\text{tr}\left\{[I_n - \Theta_{M,I_n}(\lambda)]^2 \beta^\dagger \beta^{\dagger\prime}\right\}}_{\text{bias squared}} + \underbrace{\text{tr}\left\{\Theta_{M,I_n}(\lambda)^2 \Sigma\right\}}_{\text{variance}}.$$

- Unbiased estimate:

$$\beta^\dagger \beta^{\dagger\prime} = E(\hat{\beta}\hat{\beta}') - \Sigma.$$

- Plug in:

$$\begin{aligned}\tilde{R}_{M,I_n}(\lambda) &= \text{tr}\left\{[I_n - \Theta_{M,I_n}(\lambda)]^2 (\hat{\beta}\hat{\beta}' - \Sigma)\right\} + \text{tr}\left\{\Theta_{M,I_n}(\lambda)^2 \Sigma\right\} \\ &= \hat{R}_{M,I_n}(\lambda) - \text{tr}(\Sigma).\end{aligned}$$

Back

Bound on critical value

- Triangle inequality:

$$\sqrt{\widehat{LR}_{M,W}(R\beta)} \leq \|R(\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - \hat{\beta})\|_{V(\hat{\lambda})^{-1}} + \|R(\hat{\beta} - \beta)\|_{V(\hat{\lambda})^{-1}}.$$

- Let $Z \sim N_n(0, W^{-1})$. For any $\beta \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ symm. p.d.,

$$\|R(\beta - \hat{\beta})\|^2_{V(\hat{\lambda})^{-1}} \leq \|\beta - \hat{\beta}\|^2_A \times \rho\left(RA^{-1}R' \operatorname{Var}(RZ \mid MZ)^{-1}\right).$$

- Since $\hat{R}_{M,W}(\hat{\lambda}_{M,W}) \leq \hat{R}_{M,W}(0)$,

$$\|\hat{\beta}_{M,W}(\hat{\lambda}_{M,W}) - \hat{\beta}\|^2_W \leq 2 \operatorname{tr}\left\{M\Sigma M'(MW^{-1}M')^{-1}\right\}.$$

- Under the null $H_0\colon R\beta^\dagger = R\beta$,

$$\|R(\hat{\beta} - \beta)\|^2_{(R\Sigma R')^{-1}} \sim \chi^2(r).$$

Back

Uniform confidence band

- Supremum test statistic of $H_0\colon \beta_i^\dagger = \beta_i$, $i = 1, \dots, n$:

$$\widehat{SLR}_{M,W}(\beta) = \sup_{i=1,\dots,n} \left| \frac{\hat{\beta}_{i,M,W}(\hat{\lambda}_{M,W}) - \beta_i}{\sqrt{e_i'(W^{-1} + \hat{\lambda}_{M,W} M'M)^{-1} e_i}} \right|.$$

- Simulate null critical value $\tilde{q}_{1-\alpha,M,W}(\beta)$ for any β.

- Simultaneous confidence band: rectangular envelope of inverted test.

$$\widetilde{C}_{M,W} = \bigtimes_{i=1}^{n} \left[\inf_{\beta\colon \widehat{SLR}(\beta) \le \tilde{q}_{1-\alpha}(\beta)} \beta_i, \sup_{\beta\colon \widehat{SLR}(\beta) \le \tilde{q}_{1-\alpha}(\beta)} \beta_i \right].$$

- Computationally challenging. Can sample from band.
  Inoue & Kilian (2016)

Back

Coverage discrepancy: proof sketch

- Proof reinterprets Andrews & Mikusheva (2016) result on conditional testing.

$$
\begin{aligned}
P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \triangle \widehat{C}_{EB}\right) = {} & \overbrace{E_{\beta^\dagger}\left[\mathbb{1}(R\beta^\dagger \in \widetilde{C})\right]}^{=1-\alpha} + E_{\beta^\dagger}\left[\mathbb{1}(R\beta^\dagger \in \widehat{C}_{EB})\right] \\
& - 2E_{\beta^\dagger}\left[\mathbb{1}(R\beta^\dagger \in \widetilde{C})\mathbb{1}(R\beta^\dagger \in \widehat{C}_{EB})\right]
\end{aligned}
$$

Coverage discrepancy: proof sketch

- Proof reinterprets Andrews & Mikusheva (2016) result on conditional testing.

$$P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \triangle \widehat{C}_{EB}\right) = \overbrace{E_{\beta^\dagger}\left[\mathbb{1}(R\beta^\dagger \in \widetilde{C})\right]}^{=1-\alpha} + E_{\beta^\dagger}\left[\mathbb{1}(R\beta^\dagger \in \widehat{C}_{EB})\right] - 2E_{\beta^\dagger}\left[\mathbb{1}(R\beta^\dagger \in \widetilde{C})\mathbb{1}(R\beta^\dagger \in \widehat{C}_{EB})\right]$$

$$\begin{aligned} &P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \triangle \widehat{C}_{EB}\right) - P_{\beta^\dagger}\left(R\beta^\dagger \in \widehat{C}_{M,W} \triangle \widehat{C}_{EB}\right) \\ &= 2E_{\beta^\dagger}\left[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\}\mathbb{1}(R\beta^\dagger \in \widehat{C}_{EB})\right] \\ &= 2E_{\beta^\dagger}\left[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\}\mathbb{1}\left(\widehat{LR}_{M,W}(R\beta^\dagger) \le \chi^2_{r,1-\alpha}\right)\right] \end{aligned}$$

- Similarity of $\widetilde{C}$ and completeness of the Gaussian family imply conditional similarity (like $\widehat{C}_{M,W}$):

$$P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \;\middle|\; \hat{\nu}\right) = 1 - \alpha.$$

- By law of iterated expectations,

$$E_{\beta^\dagger}\left[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\}\mathbb{1}\Big(q_{1-\alpha,M,W}(R\beta^\dagger, \hat{\nu}) \leq \chi^2_{r,1-\alpha}\Big)\right] = 0.$$

- Similarity of $\widetilde{C}$ and completeness of the Gaussian family imply conditional similarity (like $\widehat{C}_{M,W}$):

$$P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \,\middle|\, \hat{\nu}\right) = 1 - \alpha.$$

- By law of iterated expectations,

$$E_{\beta^\dagger}\left[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\}\mathbb{1}\Big(q_{1-\alpha,M,W}(R\beta^\dagger, \hat{\nu}) \le \chi^2_{r,1-\alpha}\Big)\right] = 0.$$

$$\begin{aligned}
&P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \triangle \widehat{C}_{EB}\right) - P_{\beta^\dagger}\left(R\beta^\dagger \in \widehat{C}_{M,W} \triangle \widehat{C}_{EB}\right) \\
&= 2E_{\beta^\dagger}\Big[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\} \\
&\qquad \times \left\{\mathbb{1}\Big(\widehat{LR}_{M,W}(R\beta^\dagger) \le \chi^2_{r,1-\alpha}\Big) - \mathbb{1}\Big(q_{1-\alpha,M,W}(R\beta^\dagger, \hat{\nu}) \le \chi^2_{r,1-\alpha}\Big)\right\}\Big]
\end{aligned}$$

- Variable inside the expectation is a.s. nonnegative by def'n of $\widehat{C}_{M,W}$.

- Similarity of $\widetilde{C}$ and completeness of the Gaussian family imply conditional similarity (like $\widehat{C}_{M,W}$):

$$P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \;\middle|\; \hat{\nu}\right) = 1 - \alpha.$$

- By law of iterated expectations,

$$E_{\beta^\dagger}\left[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\}\mathbb{1}\left(q_{1-\alpha,M,W}(R\beta^\dagger, \hat{\nu}) \leq \chi^2_{r,1-\alpha}\right)\right] = 0.$$

$$\begin{aligned} &P_{\beta^\dagger}\left(R\beta^\dagger \in \widetilde{C} \triangle \widehat{C}_{EB}\right) - P_{\beta^\dagger}\left(R\beta^\dagger \in \widehat{C}_{M,W} \triangle \widehat{C}_{EB}\right) \\ &= 2E_{\beta^\dagger}\Big[\left\{\mathbb{1}(R\beta^\dagger \in \widehat{C}_{M,W}) - \mathbb{1}(R\beta^\dagger \in \widetilde{C})\right\} \\ &\qquad \times \left\{\mathbb{1}\left(\widehat{LR}_{M,W}(R\beta^\dagger) \leq \chi^2_{r,1-\alpha}\right) - \mathbb{1}\left(q_{1-\alpha,M,W}(R\beta^\dagger, \hat{\nu}) \leq \chi^2_{r,1-\alpha}\right)\right\}\Big] \end{aligned}$$

- Variable inside the expectation is a.s. nonnegative by def'n of $\widehat{C}_{M,W}$.
- Crucial: EB set inverts same test stat., but non-random crit. val.

Back