Prediction risk for global-local shrinkage regression

Anindya Bhadra
bhadra@purdue.edu
www.stat.purdue.edu/∼bhadra

Purdue University

Overview

- Goal: To quantify the prediction risk for *global* and *global-local* shrinkage regressions.
- Stein's unbiased risk estimate (SURE) for global shrinkage regression.
- Stein's unbiased risk estimate (SURE) for global-local shrinkage regression.
- Numerical examples.
- *Joint work with Jyotishka Datta (University of Arkansas); Yunfan Li (Purdue University); Nick Polson and Brandon Willard (The University of Chicago).* *Bhadra and Polson are supported by NSF Grant DMS-1613063.*

Orthogonalization in high-dimensional regression

- Consider the high-dimensional regression model with $p > n$

$$y = X\beta + \epsilon,$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p$ and $\epsilon \sim \text{Normal}(0, \sigma^2 \mathbf{I}_n)$.

- Let $X = UDW^T$, $\text{Rank}(D) = n$ where $D = \text{diag}(d_i)$ with $d_1 \geq \ldots \geq d_n > 0$.

- Define $Z = UD$ and $\alpha = W^T\beta$.

- Then the regression problem can be reformulated as:

$$y = Z\alpha + \epsilon.$$

Shrinkage regression estimates as posterior means (Frank and Friedman, 1993)

- Define OLS estimate of α as $\hat{\alpha} = (Z^T Z)^{-1} Z^T y = D^{-1} U^T y$.

- Consider the following hierarchical model with $\sigma^2, \tau^2 > 0$:

$$\begin{aligned} (\hat{\alpha}_i \mid \alpha_i, \sigma^2) &\overset{ind}{\sim} \text{Normal}(\alpha_i, \sigma^2 d_i^{-2}), \\ (\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) &\overset{ind}{\sim} \text{Normal}(0, \sigma^2 \tau^2 \lambda_i^2). \end{aligned}$$

- Given λ_i and τ, the estimate for β, denoted by $\tilde{\beta}$ is given by:

$$\tilde{\alpha}_i = \frac{\tau^2 \lambda_i^2 d_i^2}{1 + \tau^2 \lambda_i^2 d_i^2} \hat{\alpha}_i, \quad \tilde{\beta} = \sum_{i=1}^{n} \tilde{\alpha}_i w_i,$$

where $\tilde{\alpha}_i = E(\alpha_i \mid \tau, \lambda_i^2, X, y)$, w_i is the ith column of the $p \times n$ matrix W and the term $\tau^2 \lambda_i^2 d_i^2 / (1 + \tau^2 \lambda_i^2 d_i^2) \in (0, 1)$ is the shrinkage factor.

Some examples: ridge, PCR and regression with g-prior

- For ridge regression, $\lambda_i^2 = 1$ for all i and we have $\tilde{\alpha}_i = \{\tau^2 d_i^2/(1+\tau^2 d_i^2)\}\hat{\alpha}_i$.
- For K component PCR, λ_i^2 is infinite for the first K components and then 0. Thus, $\tilde{\alpha}_i = \hat{\alpha}_i$ for $i = 1, \ldots, K$ and $\tilde{\alpha}_i = 0$ for $i = K+1, \ldots, n$.
- For regression with g-prior, $\lambda_i^2 = d_i^{-2}$ and we have $\tilde{\alpha}_i = \{\tau^2/(1+\tau^2)\}\hat{\alpha}_i$ for $i = 1, \ldots, n$.

Stein's unbiased risk estimate or SURE (Stein, 1981)

- If "prediction" is the main modeling goal, then the fitted risk is an underestimation of the prediction risk.
- Define the fit $\tilde{y} = X\tilde{\beta} = Z\tilde{\alpha}$, where $\tilde{\alpha}$ is the posterior mean of α.
- Then SURE is given by

$$R = ||y - \tilde{y}||^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \tilde{y}_i}{\partial y_i},$$

where $\sum_{i=1}^{n}(\partial \tilde{y}_i / \partial y_i)$ is the "degrees of freedom."

SURE for global shrinkage regressions

- A simple formula need not exist for the degrees of freedom!
- However, since our estimates are posterior means under certain priors, perhaps we can get some simplifications?
- According to Tweedie's formula:

$$\tilde{\alpha} = \hat{\alpha} + \sigma^2 D^{-2} \nabla_{\hat{\alpha}} \log m(\hat{\alpha}).$$

- Noting that $y = Z\hat{\alpha}$ and $\tilde{y} = Z\tilde{\alpha}$ and $\hat{\alpha}_i$s are independent:

$$R = \sigma^4 \sum_{i=1}^{n} d_i^{-2} \left\{ \frac{\partial}{\partial \hat{\alpha}_i} \log m(\hat{\alpha}_i) \right\}^2 + 2\sigma^2 \sum_{i=1}^{n} \left\{ 1 + \sigma^2 d_i^{-2} \frac{\partial^2}{\partial \hat{\alpha}_i^2} \log m(\hat{\alpha}_i) \right\}.$$

SURE for global shrinkage regressions (contd.)

- Thus, calculating the first two derivatives of the log marginal of the independent $\hat{\alpha}_i$s is enough to calculate SURE!

- Integrating out α_i, it is easy to see that

$$(\hat{\alpha}_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \text{Normal}(0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)).$$

- After elementary calculations, SURE is $R = \sum_{i=1}^{n} R_i$ where

$$R_i = \frac{\hat{\alpha}_i^2 d_i^2}{(1+\tau^2\lambda_i^2 d_i^2)^2} + 2\sigma^2 \frac{\tau^2\lambda_i^2 d_i^2}{(1+\tau^2\lambda_i^2 d_i^2)}.$$

Difficulties with purely global shrinkage

- Recall that in purely global shrinkage λ_i^2 are fixed and there is a single tuning parameter τ.
- If a small τ is chosen df ≈ 0 but terms with large $\hat{\alpha}_i^2 d_i^2$ make a large contribution to the risk.
- If a large τ is chosen it solves the above problem, but at the expense of a df $\approx 2\sigma^2$ for all terms!
- Maybe component-specific shrinkage will help?
- Also note the shrinkage factor $\tau^2\lambda_i^2 d_i^2/(1+\tau^2\lambda_i^2 d_i^2)$ is monotone in d_i for any given τ and fixed λ_is.

Global-local shrinkage regression

- Consider the equations

$$\begin{aligned}(\hat{\alpha}_i \mid \alpha_i, \sigma^2) &\overset{ind}{\sim} \text{Normal}(\alpha_i, \sigma^2 d_i^{-2}),\\ (\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) &\overset{ind}{\sim} \text{Normal}(0, \sigma^2\tau^2\lambda_i^2),\\ \lambda_i &\overset{ind}{\sim} p(\lambda_i).\end{aligned}$$

- The first two equations are the same as before.

- However, now we treat λ_i as random and put a half-Cauchy prior on it, i.e.,

$$p(\lambda_i) \propto \frac{1}{1+\lambda_i^2}.$$

A bit more on the choice of prior

- The induced prior on α_i on the previous slide is the so called "horseshoe prior."
- A small τ should help in shrinking the small α_i terms to zero.
- The half-Cauchy prior on λ_i has heavy tails. This should help in "not shrinking" the large α_i terms too much.
- This is what Polson and Scott (2012) did in simulations and noticed good prediction results.
- But can we rigorously show an improved prediction risk?

SURE for global-local shrinkage regression

Theorem 1

Let $m'(\hat{\alpha}_i) = (\partial/\partial\hat{\alpha}_i)m(\hat{\alpha}_i)$ *and* $m''(\hat{\alpha}_i) = (\partial^2/\partial\hat{\alpha}_i^2)m(\hat{\alpha}_i)$. *Then,*

A. *SURE for the global-local shrinkage regression model is given by* $R = \sum_{i=1}^{n} R_i$, *where*

$$R_i \;=\; 2\sigma^2 - \sigma^4 d_i^{-2}\left\{\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}\right\}^2 + 2\sigma^4 d_i^{-2}\frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}.$$

B. *Under independent standard half-Cauchy prior on* λ_i*s, for the second and third terms in Part A we have:*

$$\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathbb{E}(Z_i), \text{ and, } \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{d_i^2}{\sigma^2}\mathbb{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathbb{E}(Z_i^2),$$

where $(Z_i \mid \hat{\alpha}_i, \sigma, \tau)$ *follows a* $\mathrm{CCH}(p = 1, q = 1/2, r = 1, s = \hat{\alpha}_i^2 d_i^2/2\sigma^2, v = 1, \theta = 1/\tau^2 d_i^2)$ *distribution.*

Some remarks on Theorem 1

- The previous theorem establishes that SURE for global-local regression can be expressed by the first two moments of the compound confluent hypergeometric (CCH) distribution.
- These moments can be expressed as doubly infinite series that converge relatively fast and numerical calculations are quick (Gordy, 1998).
- An easy consequence is that now one can do a one-dimensional optimization on τ to minimize SURE.

SURE when $\hat{\alpha}_i^2 d_i^2$ is large and when it is small

Theorem 2

Define $s_i = \hat{\alpha}_i^2 d_i^2 / 2\sigma^2$. When $s_i \gg 1$, both $m''(\hat{\alpha}_i)/m(\hat{\alpha}_i)$ and $[m'(\hat{\alpha}_i)/m(\hat{\alpha}_i)]^2$ are $O(1/\hat{\alpha}_i^2)$ and therefore, the contributions of the second and the third terms to R_i is $O(1/\hat{\alpha}_i^2 d_i^2)$. Consequently, the component-wise SURE $R_i \approx 2\sigma^2$.

Theorem 3

Define $s_i = \hat{\alpha}_i^2 d_i^2 / 2\sigma^2$. Then the following statements are true.

A. *The component-wise SURE R_i is an increasing function of s_i in the interval $[0, 1]$ for any fixed τ.*

B. *When $s_i = 0$, the component-wise SURE R_i is a monotone increasing function of τ, and is bounded in the interval $(0, 2\sigma^2/3]$ when $\tau^2 d_i^2 \in (0, 1]$.*

Some remarks on Theorems 2 and 3

- Recall the risk for pure global regression $R = \sum_{i=1}^{n} R_i$ where

$$R_i = \frac{\hat{\alpha}_i^2 d_i^2}{(1+\tau^2\lambda_i^2 d_i^2)^2} + 2\sigma^2 \frac{\tau^2\lambda_i^2 d_i^2}{(1+\tau^2\lambda_i^2 d_i^2)}.$$

- For global-local regression, Theorem 2 establishes that the terms with $s_i = \hat{\alpha}_i^2 d_i^2 / 2\sigma^2 \gg 1$ will contribute $2\sigma^2$ to the risk.

- For global-local regression, Theorem 3 establishes that terms with $s_i = 0$ contribute less than $2\sigma^2/3$ to the risk, provided τ is chosen sufficiently small, i.e., $\tau^2 \leq d_i^{-2}$.

- *Simultaneously controlling* the risk in these two situations (i.e., $s_i \gg 1$ and $s_i = 0$) is not possible with a single τ.

Numerical examples

Table : The true orthgonalized regression coefficients α_{0i}, their OLS estimates $\hat{\alpha}_i$, and singular values d_i of X, for $n = 100$ and $p = 500$.

i	α_{0i}	$\hat{\alpha}_i$	d_i	$\hat{\alpha}_i d_i$
1	0.10	0.10	635.10	62.13
2	-0.44	-0.32	3.16	-1.00
...	...	...	...	...
5	-0.13	0.30	3.05	0.91
6	10.07	10.22	3.02	30.88
...	...	...	...	...
29	0.46	0.60	2.53	1.53
30	10.47	11.07	2.51	27.76
...	...	...	...	...
56	0.35	0.57	2.07	1.18
57	10.23	10.66	2.07	22.05
...	...	...	...	...
66	-0.00	-0.35	1.90	-0.66
67	11.14	11.52	1.88	21.70
...	...	...	...	...
95	-0.82	-0.56	1.42	-0.79
96	9.60	10.21	1.40	14.26
...	...	...	...	...
100	0.61	0.91	1.27	1.15

Numerical examples (contd.)

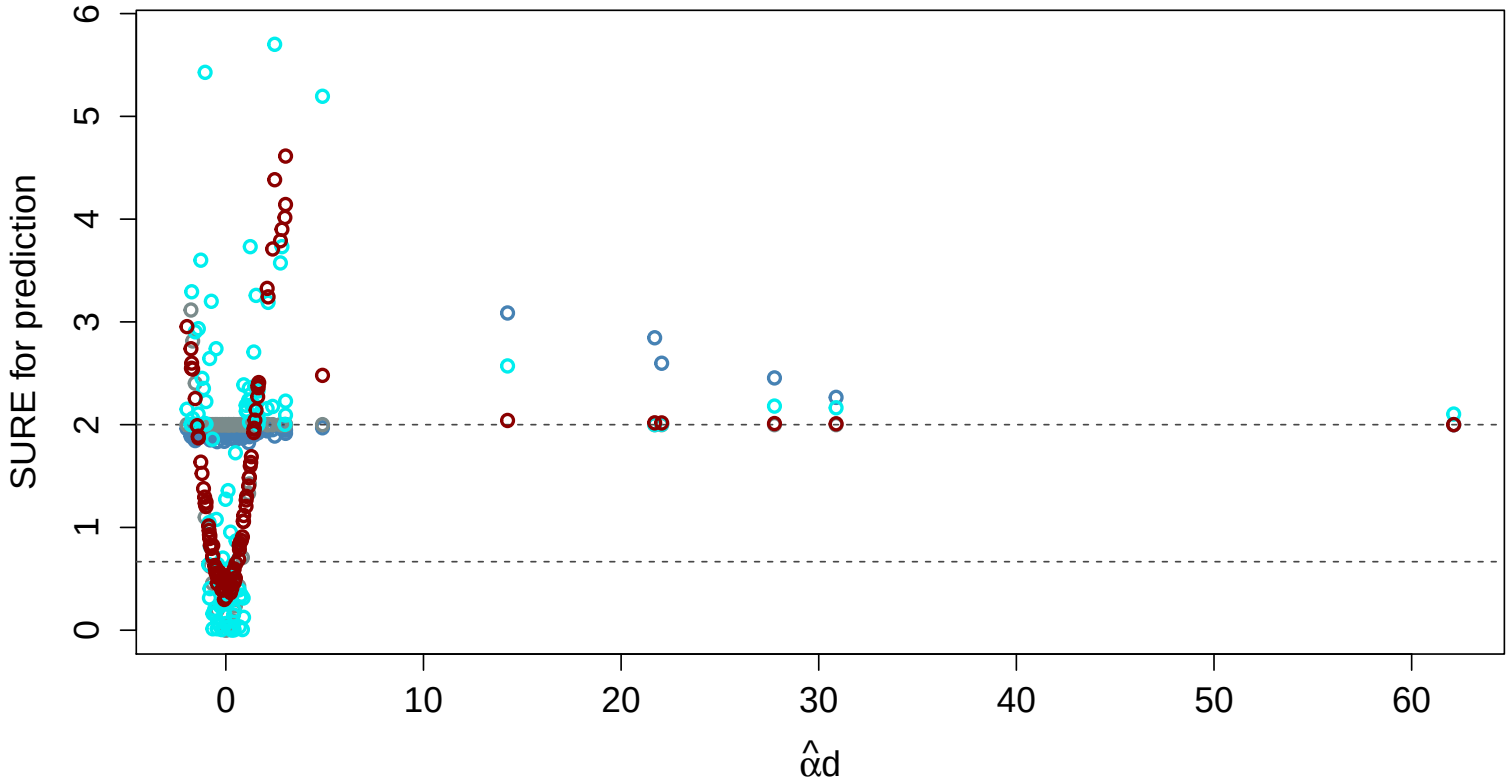

Figure : SURE for ridge (blue), PCR (gray), lasso (cyan) and horseshoe regression (red), versus $\hat{\alpha}d$, where $\hat{\alpha}$ is the OLS estimate of the orthogonalized regression coefficient, and d is the singular value, for $n = 100$ and $p = 500$. Dashed horizontal lines are at $2\sigma^2 = 2$ and $2\sigma^2/3 = 0.67$.

Numerical examples (contd.)

Table : SURE and average out of sample prediction SSE (standard deviation of SSE) on one training set and 200 testing sets for the competing methods for $n = 100$. The lowest SURE in each row is in blue and the lowest average prediction SSE is in red.

	RR		LASSO		A_LASSO	PCR		HS	
p	SURE	SSE	SURE	SSE	SSE	SURE	SSE	SURE	SSE
100	159.02	168.24	125.37	128.98	127.22	162.23	179.81	120.59	126.33
		(23.87)		(18.80)	(18.10)		(25.51)		(18.77)
200	187.38	174.92	140.99	132.46	151.89	213.90	191.33	139.32	126.99
		(21.13)		(18.38)	(20.47)		(22.62)		(17.29)
300	192.78	191.91	147.83	145.04	153.64	260.65	253.00	151.24	136.67
		(22.95)		(19.89)	(21.19)		(26.58)		(18.73)
400	195.02	182.55	148.56	165.63	178.98	346.19	292.02	147.69	143.91
		(22.70)		(21.55)	(20.12)		(28.98)		(18.41)
500	196.11	188.78	159.95	159.56	186.23	386.50	366.88	144.97	160.11
		(22.33)		(19.94)	(23.50)		(39.38)		(20.29)

References

- **Bhadra, A.**, Datta, J., Li, Y., Polson, N. G. and Willard, B. (2016). Prediction risk for global-local shrinkage regression. *(submitted)*. [arXiv:1605.04796]

- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

- Gordy, M. B. (1998). A generalization of generalized beta distributions. In *Finance and Economics Discussion Series*. Division of Research and Statistics, Federal Reserve Board.

- Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, Lévy processes, and regularized regression. *Journal of the Royal Statistical Society: Series B* **74**, 287-311.

- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9**, 1135-1151.