The Effect of Estimating Residual Variance on Post-Selection Inference of LASSO in High-Dimensional Settings

Jinyuan Liu, MS, Brian Kwan, BS, Ellen Lee, MD, Tian Chen, PhD, Xin M. Tu, PhD, Loki Natarajan, PhD

Department of Family Medicine and Public Health, UC San Diego

INTRODUCTION

The LASSO procedure is widely used nowadays in a preponderance of studies involving high-dimensional data, such as large-scale analysis of genomic and genetic data, etc. By adding L_1 penalizations to the optimization object thus shrinking estimated coefficients towards zero, it serves the role of both parameter fitting and variable selection, and achieves high prediction accuracy in various applications.

However, it is only recently that attention has been focused on **the inferential aspects of the LASSO approach**, i.e., **testing the significance** of the included predictor variables, in the sequence of models visited along the LASSO solution path. **The adaptive nature** of the estimation procedure makes this problem difficult. Some "traditional" methods include the **Bayesian methods and the bootstrap**, with more recent approaches proposed like the **covariance and spacing tests**, among others.

A MOTIVATING EXAMPLE

Patients with Type II diabetes are at high risk for mortality and comorbidities, particularly chronic kidney disease (CKD). Metabolomics may reveal novel features that identify Type II diabetics at high-risk for CKD. Here we would like to test the accuracy of a published 13- metabolite signature to predict kidney disease progression over and above clinical variables, with the primary outcome of eGFR slope, a continuous slope obtained from Generalized Linear Mixed Model (GLMM). Using a sample of 1000 Type 2 diabetic patients, we applied the penalized model selection method, LASSO, to identify new multivariate metabolite sets that are significantly associated with CKD after adjusting for clinical variables.

OBJECTIVES

We will focus on the **covariance test**, which consists of the procedure of estimating residual variance and then plugging it into the test statistics to obtain p-values. In the high dimensional settings with $N < p$, **estimating residual variance ($\boldsymbol{\sigma^2}$)** is not as straightforward, but several variance estimators have been proposed in the literature, all with **accompanying consistency** and **asymptotic normality** under certain assumptions. These methods include:

- **LASSO-CV estimation**
- **SCAD-CV estimation** (Fan, J. and Li, R, 2001)
- **Refitted Cross-Validation (RCV)** (Fan et al., 2012)
- **Scaled LASSO estimation** (Sun and Zhang, 2012)
- **Moment-based method** (Dicker, L. H., 2014)
- **De-biased LASSO estimation** (Javanmard et al., 2014)

We will explore the role of **estimating error variance ($\boldsymbol{\sigma^2}$)** with an extensive simulation study comparing various methods for estimating residual variance in high-dimensional settings, and comparing various methods by varying **the sparseness of predictors**, **signal-to-noise ratio (SNR)** and **the correlation structures in the design matrix** inform the choice of the optimal method to achieve higher statistical power.

We will also apply post-selection inference to the clinical dataset in predicting kidney disease progression to make valid statistical inference after model selection from LASSO.

METHODS: Covariance Test

The Covariance test (Lockhart et. al, 2014) is an idea on making inference after selection by the LASSO, which assigns **p-values** to predictors as they are **successively entered by the LASSO**.

Assume $\boldsymbol{\sigma^2}$ is know for now. And under usual LM setup, instead of comparing the reduction in residual sum of squares (RSS) of 2 nested models with $\chi^2(1)$ distribution, as in traditional inference, covariance test takes the adaptive fitting procedure into account by **balancing the two opposing procedures of adaptivity and shrinkage**.

With the **assumption** that columns of **X are in general position**. The LASSO path is a **continuous and piecewise linear** function of λ, with **knots** at values $\lambda_1 > \lambda_2 > \cdots > \lambda_K$. To define a test statistics **at the $\boldsymbol{k_{th}}$ step of the path**, consider:

- $\boldsymbol{A}$: active set just before λ_k; predictor k enters at λ_k
- $\boldsymbol{\hat{\beta}\ (\lambda_{k+1})}$: solution at the next knot, using predictor $\boldsymbol{A} \cup \{k\}$, i.e.

$$\hat{\beta}\ (\lambda_{k+1}) = argmin_{\beta \in R^{|A|+1}} \frac{1}{2}\ ||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2 + \lambda_{k+1}||\beta||_1$$

- $\boldsymbol{\widehat{\beta_A}\ (\lambda_{k+1})}$: solution using only active predictors at $\boldsymbol{A}$, but with $\lambda = \lambda_{k+1}$, i.e.

$$\widehat{\beta_A}(\lambda_{k+1}) = argmin_{\beta \in R^{|A|}} \frac{1}{2}\ ||\boldsymbol{y} - \boldsymbol{X}_A\beta_A||_2^2 + \lambda_{k+1}||\beta_A||_1$$

The **covariance test** statistics has been defined as:

$$T_k = \frac{< \boldsymbol{y}, \boldsymbol{X}\hat{\beta}(\lambda_{k+1}) > - < \boldsymbol{y}, \boldsymbol{X}_A\widehat{\beta_A}(\lambda_{k+1}) >}{\sigma^2}$$

Under the **null hypothesis** that **all truly active predictors are contained in the current active set**, it can be shown that

$$T_k \to_{H_0}^{d} \boldsymbol{Exp(1)}\ \ as\ N, p \to \infty$$

However, $\boldsymbol{\sigma^2}$ **is unknown** in practice. If $N > p$, $\hat{\sigma}^2 = \frac{1}{N-p}RSS_P$

$$\frac{\hat{\sigma}^2}{\sigma^2} \to_{H_0}^{d} \frac{1}{N-p}\chi^2(N-p)$$

And using $\boldsymbol{F_k} = \frac{<y, X\hat{\beta}(\lambda_{k+1})> - <y, X_A\widehat{\beta_A}(\lambda_{k+1})>}{\hat{\sigma}^2} = \frac{T_k \cdot \sigma^2}{\hat{\sigma}^2} \to_{H_0}^{d} \boldsymbol{F_{2,N-p}}$ yields p-values, but when $N < p$, **estimating residual variance ($\boldsymbol{\sigma^2}$)** is not as straightforward, several variance estimators have been proposed in the literature.

METHODS: Residual Variance Estimation

Existing methods (except for the moment-based method which assumes $\boldsymbol{X}$ has multivariate normal distribution and derives $\hat{\sigma}^2$) have similar ideas of emulating the sum of squares estimator of OLS in high dimensional settings, with the general form of $\hat{\sigma}^2 = \frac{1}{N - s_{\hat{\lambda}}}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_\lambda}||_2^2$, where $\widehat{\boldsymbol{\beta}_\lambda}$ is some estimator with regularization parameter λ, $\hat{\lambda}$ is selected via cross-validation, and $s_{\hat{\lambda}}$ is the number of non-zero element in $\widehat{\boldsymbol{\beta}_\lambda}$.

Thus, the methods of estimating $\hat{\sigma}^2$ depends on different estimating procedure for $\widehat{\boldsymbol{\beta}_\lambda}$, including: **LASSO-CV** estimation, **SCAD-CV** estimation, **Refitted Cross-Validation (RCV)** method**, Scaled LASSO** estimation, **Moment-based method, De-biased LASSO** estimation etc.

SIMULATION SETTINGS

In the study we control the **sparsity** of the underlying true β vector as well as its **signal-to-noise ratio (SNR).**

All simulations were conducted at a sample size of n = 100, and total number of predictors p = 100. Elements of the predictor matrix $\boldsymbol{X}$ were generated randomly as **orthonormal.**

The true β was generated from Poisson(1), and then the elements were scaled such that the signal-to-noise ratio, defined as $\beta'\boldsymbol{\Sigma}\beta/\sigma^2$, was some predetermined value snr. At each setting, B = 1000 replications were obtained.

SIMULATION RESULTS

Sizes of the tests:

- All methods achieved the nominal level of 5%.

Power comparisons:

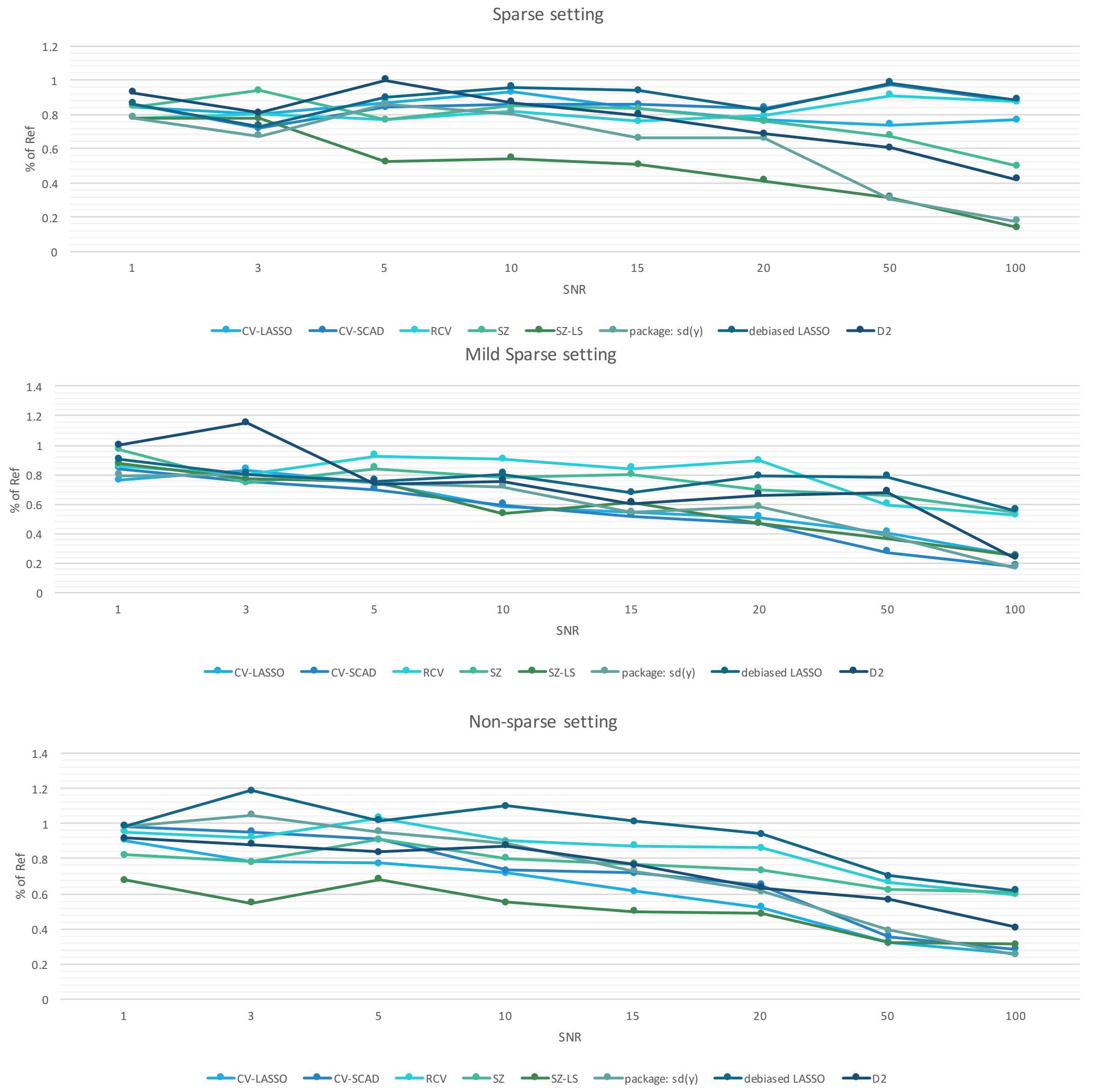
1. Effect of **sparseness**:

Table 2: Power comparisons under difference sparseness setting of $\boldsymbol{\beta}$

Power	idp X, orthogonal X, at first step k=1		
$n = p$ $= 100$	**Sparse** **# non-zero $\boldsymbol{\beta}$ = 1**	**Mild sparse** **# non-zero $\boldsymbol{\beta}$ =$\boldsymbol{p/2}$**	**Non- sparse** **# non-zero $\boldsymbol{\beta = p}$**
Reference: σ=1	% of ref	% of ref	% of ref
Package: sd(y)	0.1721	0.0013	0.0000
CV-LASSO	**0.7861**	0.0126	0.0032
CV-SCAD	**0.8605**	0.0151	0.0032
RCV	**0.9116**	**0.4384**	**0.6645**
SZ	**0.4419**	**0.4096**	**0.6966**
SZ-LS	0.1721	0.0013	0.0021
D2	0.1721	0.0013	0.0598
De-biased LASSO	**0.9535**	**0.4812**	**0.4402**

2. Effect of **SNR**:

Figure 1: Power comparisons under different sparseness + SNR combinations



REAL DATA ANALYSES

Table 1: Covariance test results evaluated at λ_{min} from CV.

	Coefficients	Z-score	P-value
AGE_INTEGER	0.019	0.687	0.482
race.fOther	0.034	1.229	0.292
race.fWhite	**0.176**	**5.635**	**0.04**
SEXFemale	**-0.053**	**-1.817**	**0.037**
SMOKE100Smoker	-0.008	-0.317	0.751
BMI	0.034	1.291	0.397
HEMOGLOBIN_A1C	**-0.071**	**-2.452**	**0.024**
MAP	**-0.123**	**-4.367**	**0.026**
alb.f>300	**-0.557**	**-15.66**	**0**
alb.f30-300	**-0.184**	**-5.982**	**0**
EGFR_CKD_EPI	0.012	0.428	0.418
Ion.122	0.043	1.048	0.266
Ion.1269	**0.084**	**2.683**	**0.007**
Ion.1468	0.013	0.443	0.658
Ion.1715	**0.054**	**1.759**	**0.049**
Ion.226	-0.087	-1.752	0.816
Ion.282	0.019	0.657	0.494
Ion.333	-0.004	-0.113	0.91
Ion.353	0.002	0.055	0.957
Ion.54	-0.05	-1.106	0.148
Ion.928	-0.066	-2.357	0.117
Ion.986	-0.022	-0.665	0.53

CONCLUSIONS

Simulation results show that all residual variance estimation methods will lead to covariance test statistics **achieving correct coverage probability**. However, **the sparseness of predictors, SNR and the correlation structures in the design matrix** are all important factors that can affect the statistical **power** of the covariance test.

- When **predictors are sparse, CV-LASSO, CV-SCAD, RCV** and **De-biased LASSO** performed the best**.** When **predictors are less sparse, RCV, SZ,** and **De-biased LASSO** performed better**.**
- In all sparseness settings, increasing **SNR** seems to induce decreasing trend of the power;
  - ❖In sparse setting, **CV-SCAD, RCV** and **De-biased LASSO** perform better, **scaled LASSO (SZ)** and **Moment method (D2)** perform good with small SNR.
  - ❖In mild sparse setting, **RCV** and and **De-biased LASSO** perform better, **Moment method (D2)** perform good with small SNR.
  - ❖In non-sparse setting, **De-biased LASSO** performs better.

Therefore, in order to achieve higher statistical power, it is crucial to assess the condition of the data before choosing the post-selection inference procedure.

REFERENCES

1. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A Significance Test for the LASSO. *Ann Stat. 2014 Apr*;42(2):413-468.

2. Hastie et. al Statistical Learning with Sparsity. 2015 CRC Press.

3. Adel Javanmard and Andrea Montanari. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* 15, 1 (January 2014), 2869-2909.

4. Stephen Reid, Robert Tibshirani and Jerome Friedman A Study of Error Variance Estimation in LASSO Regression *Statistica Sinica* 26 (2016), 35-67 doi:http://dx.doi.org/10.5705/ss.2014.042

5. Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*. doi: 10.1093/biomet/ast065.

6. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96, 1348-1360.

7. Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Roy. Statist. Soc. Ser. B* 75, 37-65.

8. Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* 99, 879-898.

CONTACT

Email: jil1168@ucsd.edu Tel: +1 (585)-285-9437