Variable selection and post-selection inference under endogeneity

David Rice Justice Nii-Ayitey R. Alex Thompson Mahedi Hasan Dipesh Baral
Washington State University

2022/05/07

Contents

1 Introduction

Variable selection is a central axis of research in high-dimensional statistics. Technological innovation has expanded data collection in almost every field. The implication of this is that in some cases, the number of variables in a dataset may far exceed the number of observations. This makes the goal of variable selection increasingly difficult since some variables can be represented as linear combinations of other variables. This may trigger spurious correlations between variables that, in reality, contribute little or nothing to the model. In addition, traditional variable selection techniques on linear models, such as the F-test, are no longer feasible since there exist multiple solutions to the residual minimization problem (Dezeure et al. (2015)). This has led the field toward developing thresholding algorithms that a priori filter variables based on their marginal contribution to the model. The most popular of these methods is LASSO regression, first introduced by Tibshirani (1996).

The rise of LASSO regression by Tibshirani (1996) has spurred widespread interest in methods that improve variable selection. This interest in LASSO has risen largely due to it's simplicity and easy implementation. In essence, LASSO works by soft-thresholding the coefficients of each parameter and setting any $\beta < \lambda$ strictly to zero. As a consequence, small $\beta's$ are excluded from the model. As a consequence of the regularization parameter, large $\beta's$ are also downward biased: a key limitation of the LASSO method (Dezeure et al. (2015)).

In addition to the downward bias imposed by LASSO, LASSO is also a sparse estimator. Zhang and Zhang (2014) provide an alternative approach that effectively de-sparsifies the original LASSO regression by providing an unbiased estimator. A key element of the bias-corrected and desparsified LASSO is the relative negligence of the error term, assuming the design matrix meets several assumptions (Dezeure et al. (2015)).

Although LASSO (and the de-sparsified LASSO) is an effective tool for variable selection, it's hard thresholding rule presents some problems. In some cases, LASSO is less than ideal. As a result, Fan and Li (2001) introduced a continuous penalty function they call the Smoothly Clipped Absolute Deviation Penalty. Central to this approach is a continuous penalty function λ_i. By allowing λ to vary with the size of the coefficient, Fan and Li (2001) both avoid downward bias and maintain the soft thresholding provided by LASSO.

The performance of a given variable selection method was recently evaluated by (Dezeure et al. (2015)). In the R package "hdi", Dezeure and colleagues introduce methods for evaluating the capacity of the de-sparsified LASSO and unbiased Ridge projection to accurately select relevant variables. The results from their simulations suggest that ridge and sample splitting procedures may provide the most type-I error control but are limited in their power to detect true sets of variables. The authors note that under well-posed designs, methods such as the de-sparsified LASSO and ridge projection maintain high power yet fail under more challenging designs.

Traditionally, variable selection has assumed that the source of error is largely exogenous and can thus be accounted for explicitly (Antonakis et al. (2010)). Exogeneity is surely important yet the role of endogeneity in error propogation has remained largely ignored in contemporary literature related to variable selection. Hao et al. (2016) explore the role of endogeneity in population genetic structure of humans globally. Under conditions where only genetic information is known (i.e. databases), how can we attempt to estimate the population structure? The authors introduce a logistic factor analysis that accounts for the discrete latent population structure of the human genome, efficiently capturing this variability.

Here, we explore the performance of ridge projection and the de-sparsified LASSO under conditions of endogeneity. For a given $X \sim N(0,1)$, we assume there is some latent variable structure Z upon which X depends. Although in reality, latent structure can present as both continuous and discrete functions, we only consider the case where $Z \sim Bin(n,p)$ and let $P(Z=z) = \pi_i(z_j)$. For fixed Gaussian design X, we have $X|Z \sim N(\pi_i(z_j), 1)$. Under these conditions, we evaluate the False Discovery Rate (FDR), true discovery rate, corrected p-values, and confidence intervals of both ridge projection and the de-sparsified LASSO under endogeneity.

2 Methods

2.1 Formulation

We consider a scalar Gaussian response Y and p predictors$\{X_i\}_{i=1}^p$ in the

$$Y|(X,Z) = \sum_{i=1}^{P} X_i\beta_i + \epsilon \tag{1}$$

where we have chosen $\epsilon_i \sim N(0,1)$

The true β vector is a sparse vector with s0 nonzero components. Each of the s0 nonzero components was generated from a Uniform(-2,2) distribution, with the rest of the components set to zero.

2.2 Latent Variable

The latent variable Z can be continuous or discrete for this project. The simulations were performed picking Z as a discrete Bernoulli random variable, $Z \sim Bern(0.2)$. For our simulations, we generated an $n \times p$ matrix of these independent Bernoulli values, which we will refer to as the Z-matrix. The choice in $\pi = 0.2$ is an arbitrary choice which we hold constant across each simulation.

2.3 Conditional Data Matrix

We need to consider the cases of the distribution of the conditional data matrix $X|Z$ as either heavy tailed, or light tailed. Our simulations consider the heavy tailed case, where $X|Z$ is an $n \times p$ multivariate normal distribution with the mean dependent on the latent variable Z, $X|Z \sim N(\mu(Z), \Sigma)$. The covariance matrix Σ is equal to a correlation matrix with $\rho = 0.9$ in the off diagonals.

Generation of this matrix was done using the 'rXb' function in the hdi R-package seen in Dezeure et al. (2015) to generate an initial X matrix, X^{init}. For each column entry X_{ij}^{init} of the initial

matrix, the mean was shifted depending on the corresponding value in the Z-matrix, Z_{ij}. In summary, $X|Z = X^{init} + Z_{matrix}$. We note a concern that the covariance matrix for $X|Z$ does not remain "equicorr" with $\rho = 0.9$, as the X^{init} was generated with the correct covariance, but the shift may have disrupted this.

2.4 Parameters Estimate

For this project, we wish to compare the De-sparsified Lasso and Ridge Projection under varying sparsity and paradigms, through varying the p predictors, s0 nonzero predictors, and n observations. For each given combination of parameters, we generate 100 repetitions. Two variable selection techniques were used for our simulations:

1. De-sparsified Lasso
2. Ridge Projection

For each estimate under each set of parameters specified, we generate 6 statistics listed below.

1. **FDR:** This stands for False Discovery Rate, which is calculated by taking the ratio $\frac{\sum_{i=1}^{p\times 100} I(P_{corr,i}<0.05)|\beta_i=0}{\sum_{j=1}^{p\times 100} I(P_{corr,j}<0.05)}$ or the number of incorrectly rejected β_i values over the total number of rejected β_i values based on the corrected p-value seen in (Dezeure et al. (2015)).
2. **AVG LB, AVG UB**: For each of the $p \times 100$ β_i estimates, a single confidence interval is generated. We take the average lower bound and upper bound of the 95% CI.
3. **Prop. Captured**: For each of the $p \times 100$ confidence intervals generated, the 95% confidence interval generated either contains or does not contain the true value of β_i. The total proportion of true β_is captured by the confidence intervals out of the $p \times 100$ values is the Prop. Captured
4. **p-value ($\beta = 0$)**: Over all $(p - s0) \times 100$ values of $\beta_i = 0$ coefficients, we take the average corrected p-value.
5. **p-value ($\beta \neq 0$)**: Over all $(s0) \times 100$ values of $\beta_i \neq 0$ coefficients, we take the average corrected p-value.

3 Results

The role of endogeneity in variable selection has been largely underexplored in the literature. Under some precise conditions, we show that endogeneity has little impact on variable selection methods such as de-sparsified lasso and ridge projection. For both methods, when $p >> n$ and s_0 is very high, Ridge projection underestimates β when compared to LASSO (Fig. 3). However, LASSO appears to more closely estimate the true β parameter (Fig. 4). This trend persists

for various settings of n and p; reducing sparsity (i.e. increasing s_0) increases the error of the estimate.

Figure 1 further highlights these differences between LASSO and Ridge under endogeneity. Instead however, comparing $\hat{\beta}$ estimates against the true β values, we see that the differences in performance are largely due to the ratio of n/p. When the sample size is small and latent variables create a source of endogenous variance, it can become difficult for either method to efficiently estimate true coefficient values.

Consider s_0 to be fixed. As shown in Dezeure et al. (2015), desparsified LASSO maintains better performance than Ridge in point value estimation (Figs. 3 4). In essence, this highlights the improved power of LASSO over Ridge. In contrast, consider the setting where n/p = c. For fixed s_0, but especially very small s_0, ridge performs exceptionally well when comparing point estimation (Figs. 3 4). As s_0 increases or $n >> p$, LASSO and ridge perform similarly. Under endogeneity, the confidence intervals of both methods is consistent at capturing $> 95\%$ of the true β values. However, when n/p = c and $s_0 = 30$ (shown in Table 5), the confidence intervals capture only 92% of the true parameters. This appears to be an exception in the simulation and not the rule. In general, the type 1 error rate of these confidence intervals remains low under endogeneity for both methods.

In the conclusion, we consider potential reasons behind these results yet inspection of the point estimation may provide some explanation. Ridge is a hard-thresholding approach that downward-biases even large estimates of β. Thus when n is very small and sparsity is low, ridge will consistently underestimate the true parameter values (Figs. 3 4).

Although LASSO may be a more efficient estimate under very sparse conditions, it does not succeed in controlling the false discovery rate. Tables 1-6 compare the false discovery rates of each method with different values of n and p. Ridge projection has a low false discovery rate under nearly all conditions due to it's downward bias (Fig. 6). Selection of non-relevant parameters is efficiently controlled though this comes at the cost of identifying true parameters, a task it performs poorly at when compared to LASSO (Fig. 3).

LASSO appears to represent a middle ground approach (Fig. 5). When $p >> n$, false discovery rate is amplified and many non-relevant parameters may be selected. In addition, true parameters may be missed. However, as n increases relative to p, the low performance vanishes and LASSO begins to reliably estimate true β coefficients.

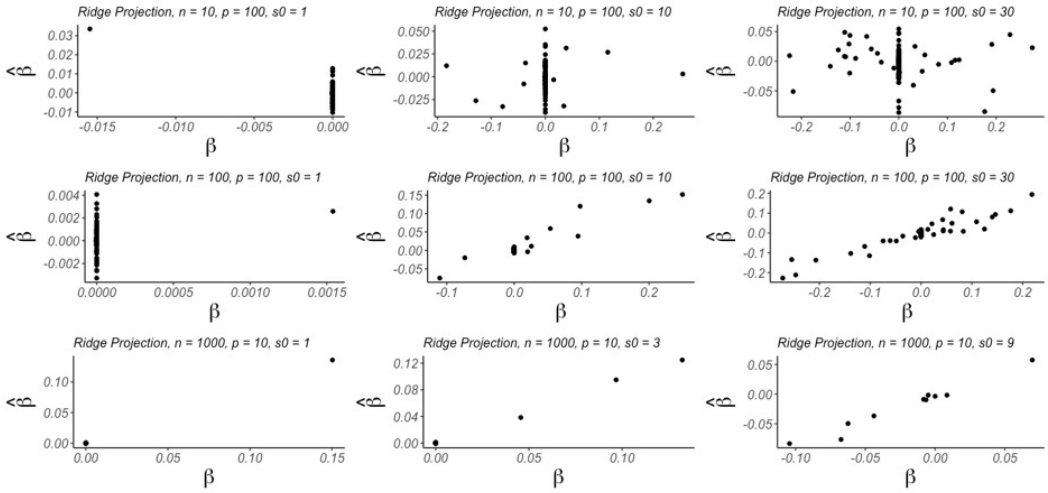

Figure 1: Beta-beta plots for ridge projection under various sparsity conditions. Comparison of the estimated beta values against the true beta values. Under very sparse conditions, $\hat{\beta}$ tends to agree well with the true β values. As sparsity expands but $p >> n$, $\hat{\beta}$ estimates no longer estimate the true parameter values well. When p/n = c and $n >> p$, ridge projection performs well regardless of the sparsity condition.

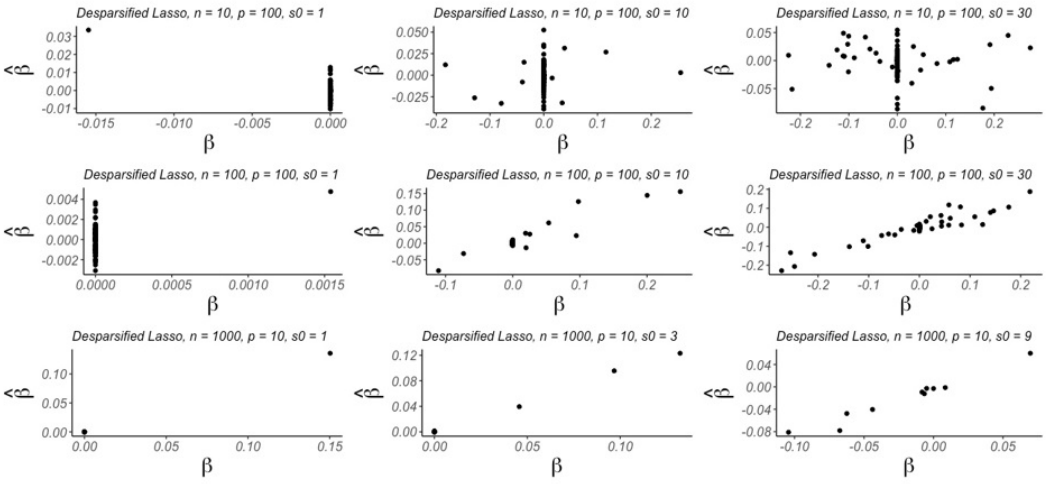

Figure 2: Beta-beta plots for the de-sparsified LASSO under various sparsity conditions. Comparison of the estimated beta values against the true beta values. Under very sparse conditions, $\hat{\beta}$ tends to agree well with the true β values. As sparsity expands but $p >> n$, $\hat{\beta}$ estimates no longer estimate the true parameter values well. When p/n = c and $n >> p$, de-sparsified LASSO performs well regardless of the sparsity condition.

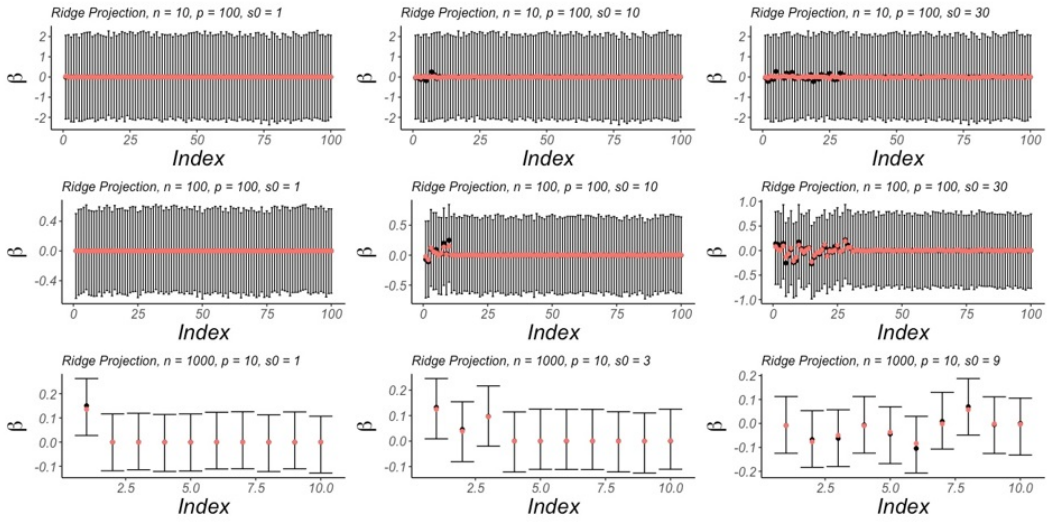

Figure 3: Ridge projection of parameter under $p >> n$. Ridge projection consistently underestimates the true beta parameter when $p >> n$. True β values (black points) are plotted with the confidence interval of the estimated beta parameter (red points). The x-axis refers to the index of the ith parameter in each model. When p/n=c and $n >> p$, ridge performs well at capturing the true parameter value.

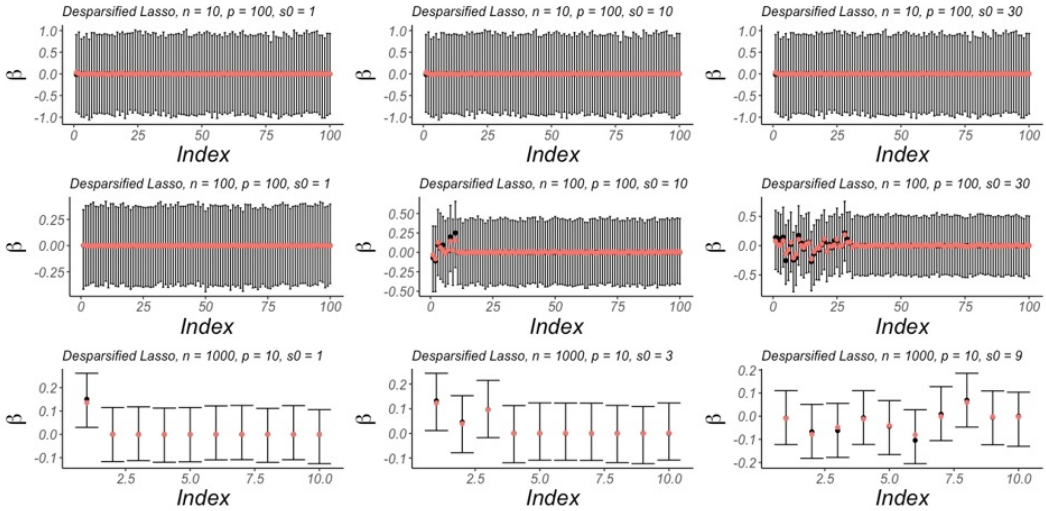

Figure 4: De-sparsified LASSO estimates of parameters. De-sparsified LASSO provides a robust estimate of the true parameter value when $p >> n$, p/n=c and $n >> p$. True β values (black points) are plotted with the confidence interval of the estimated β parameter (red points). The x-axis refers to the index of the ith parameter in each model.

Table 1: De-sparsified (n = 1000, p = 10)

	FDR	AVG. LB	AVG.UB	Prop. Captured	p-value($\beta = 0$)	p-value($\beta \neq 0$)
s0 = 1	0.292331	0.02989887	0.2614957	0.957	0.06958573	0.5119485
s0 = 3	0.09375	0.01153819	0.2441021	0.963	0.0364141	0.2572743
s0 = 9	0.009259259	-0.1228556	0.1105431	0.951	0.02817667	0.0845905

Table 2: Ridge Proj. (n = 1000, p = 10)

	FDR	AVG. LB	AVG.UB	Prop. Captured	p-value($\beta = 0$)	p-value($\beta \neq 0$)
s0 = 1	0.02173913	0.02744472	0.2627581	0.956	0.07065849	0.9530567
s0 = 3	0.01052632	0.009029593	0.2453131	0.963	0.03681513	0.9197304
s0 = 9	0.005875441	-0.12452	0.1122729	0.957	0.02841396	0.592315

Table 3: De-sparsified (n = 10, p = 100)

	FDR	AVG. LB	AVG.UB	Prop. Captured	p-value ($\beta = 0$)	p-value ($\beta \neq 0$)
s0 = 1	0.9545455	-0.885819	0.9082666	0.9515	0.9283848	0.9618583
s0 = 10	0.7930175	-1.846066*	1.814109*	0.9499	0.9816953	0.9833593
s0 = 30	0.6474654	-2.656795*	2.700131*	0.9305	0.9880757	0.9874145

Table 4: Ridge Proj. (n = 10, p = 100)

	FDR	AVG. LB	AVG.UB	Prop. Captured	p-value ($\beta = 0$)	p-value ($\beta \neq 0$)
s0 = 1	1	-2.083213	2.060574	0.9987	0.9767084	0.996108
s0 = 10	0	-3.40*	3.445845*	0.9973	0.9988018	0.995506
s0 = 30	0	-5.085863*	4.805008*	0.9952	0.9984765	0.9948788

Table 5: De-sparsified (n = 100, p = 100)

	FDR	AVG. LB	AVG.UB	Prop. Captured	p-value ($\beta = 0$)	p-value ($\beta \neq 0$)
s0= 1	0.8346939	-0.4157424	0.3431213	0.9588	0.2414269	0.6207135
s0 = 10	0.3032258	-0.5075523	0.3404699	0.958	0.3455361	0.4044717
s0 = 30	0.1431095	-0.4056422	0.6037323	0.9236	0.4716329	0.4886375

Table 6: Ridge Proj. (n = 100, p = 100)

	FDR	AVG. LB	AVG.UB	Prop. Captured	p-value ($\beta = 0$)	p-value ($\beta \neq 0$)
s0= 1	0.05882353	-0.6373235	0.5017032	0.9618	0.516078	0.9955071
s0=10	0.003154574	-0.7136132	0.5685099	0.9643	0.5642348	0.9956483
s0=30	0.004643963	-0.6896516	0.7924367	0.9606	0.6653522	0.9903209

*** represents there were confidence intervals removed due to $\pm\infty$. For rows which s0 = 10 and s0 = 30, 3 and 2 CIS removed respectively**

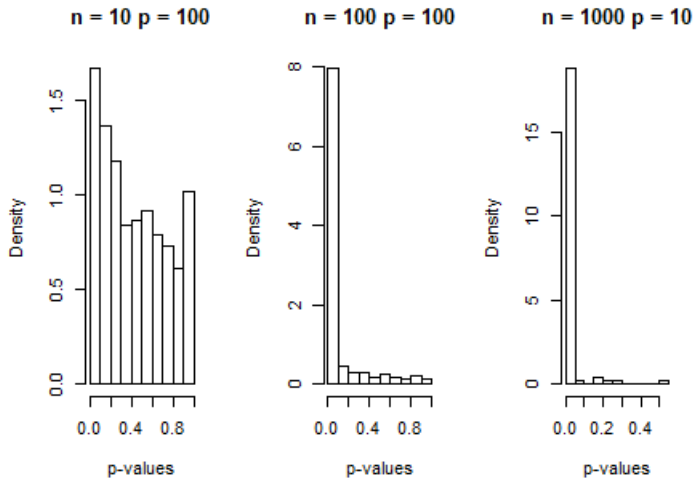

Figure 5: Desparsified p-value (un-adjusted)

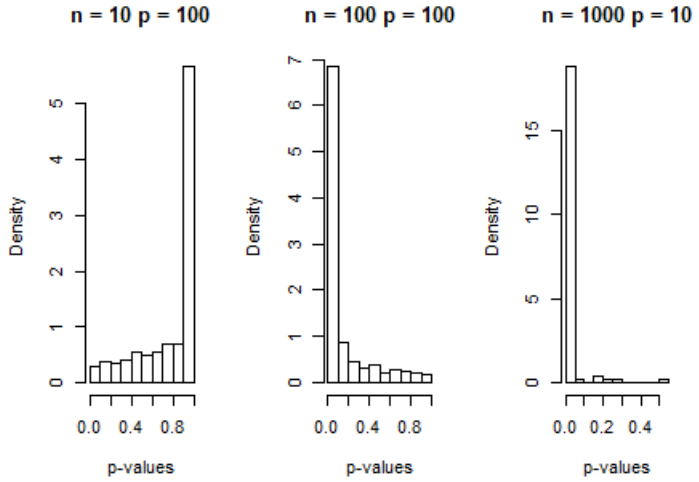

Figure 6: Ridge p-value (un-adjusted)

4 Discussion

Endogeneity is an understudied subject in variable selection methods. Latent dependence is a phenomenon that occurs in various places in real-world datasets. Further understanding the effect of endogeneity on model specification and variable selection methods is thus an important frontier in statistical research. How can we correctly estimate parameter coefficients when latent structures cannot be observed and how does this impact the performance of our estimation procedure?

There are many approaches to variable selection, among the most popular being ridge projection and the de-sparsified LASSO. Here, we explored how endogeneity effects the performance of both ridge and LASSO variable selection methods. The hard-thresholding provided by the ridge projection method implies it is a weak regularization method when $p >> n$, except under very sparse conditions (Figure 3). In contrast, LASSO provides a robust approach when compared to ridge (Figure 4). Comparing both methods, we can clearly see that as n approaches infinity, estimation error approaches zero. However, in most cases, n is fixed. Thus, a key implication of our result suggests that under endogeneity, LASSO is a consistently better method of variable selection for small n.

When n is very large, ridge may be a preferred approach for controlling type-I error. The histograms of our p-values further highlight the instability of LASSO for type-I error control when $p >> n$. Often, correct estimation of the full set of relevant parameters is often preferred over under selection. The preference of power over type-I error control may infer some cost in terms of overfitting, due to including excess variables in the model.

Endogeneity contributed little to no variance in the selection procedures, and we found results consistent with those of other authors (Dezeure et al. 2015). Specifically, our simulations suggest that under the equicorrelation covariance matrix with a bernoulli latent variable that changed the mean of a normally distributed random variable, the relationship between LASSO and Ridge is conserved under various sparsity conditions. Future work should explore how alternative latent variable structure, covariance matrices, and alternate values of n and p influence these results.

5 Appendix

References

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen, "High-dimensional inference: confidence intervals, p-values and r-software hdi," *Statistical science*, pp. 533–558, 2015.

R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, 2014.

J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

J. Antonakis, S. Bendahan, P. Jacquart, and R. Lalive, "On making causal claims: A review and recommendations," *The leadership quarterly*, vol. 21, no. 6, pp. 1086–1120, 2010.

R.-H. Hao, Y. Guo, S.-S. Dong, G.-Z. Weng, H. Yan, D.-L. Zhu, X.-F. Chen, J.-B. Chen, and T.-L. Yang, "Associations of plasma fgf2 levels and polymorphisms in the fgf2 gene with obesity phenotypes in han chinese population," *Scientific Reports*, vol. 6, no. 1, pp. 1–7, 2016.