The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 12

A Note on the Effect on Power of Score Tests via Dimension Reduction by Penalized Regression under the Null

Josue G. Martinez*

Raymond J. Carroll[†]

Samuel Muller[‡]

Joshua N. Sampson**

Nilanjan Chatterjee^{††}

*Texas A&M University, jgmartinez@stat.tamu.edu

[†]Texas A&M University, rcarroll@tamu.edu

[‡]University of Sydney, samuel.mueller@sydney.edu.au

**National Cancer Institute, joshua.sampson@nih.gov

^{††}National Cancer Institute, chattern@mail.nih.gov

A Note on the Effect on Power of Score Tests via Dimension Reduction by Penalized Regression under the Null*

Josue G. Martinez, Raymond J. Carroll, Samuel Muller, Joshua N. Sampson, and
Nilanjan Chatterjee

Abstract

We consider the problem of score testing for certain low dimensional parameters of interest in a model that could include finite but high dimensional secondary covariates and associated nuisance parameters. We investigate the possibility of the potential gain in power by reducing the dimensionality of the secondary variables via oracle estimators such as the Adaptive Lasso. As an application, we use a recently developed framework for score tests of association of a disease outcome with an exposure of interest in the presence of a possible interaction of the exposure with other co-factors of the model. We derive the local power of such tests and show that if the primary and secondary predictors are independent, then having an oracle estimator does not improve the local power of the score test. Conversely, if they are dependent, there is the potential for power gain. Simulations are used to validate the theoretical results and explore the extent of correlation needed between the primary and secondary covariates to observe an improvement of the power of the test by using the oracle estimator. Our conclusions are likely to hold more generally beyond the model of interactions considered here.

KEYWORDS: Adaptive Lasso, gene-environment interactions, Lasso, model selection, oracle estimation, score tests

*Martinez was supported by a Postdoctoral Training grant from the National Cancer Institute (CA90301). Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030). Chatterjee's research was supported by a Gene-Environment Initiative (GEI) grant from the National Heart Lung and Blood Institute (NHLBI) and by the Intramural research program of the National Cancer Institute. This paper benefited from the constructive comments of two referees and an associate editor.

1 Introduction

Dimension reduction in regression models via regularization has emerged as a topic that has undergone vigorous development, with oracle methods such as SCAD (Fan and Li, 2001) and the Adaptive Lasso (Zou, 2006) and other popular methods such as the Lasso (Tibshirani, 1996). We will say that an oracle method is one which is asymptotically consistent, i.e., selects the correct model with probability tending to 1.0 as the sample size $n \rightarrow \infty$, and is asymptotically efficient for estimating the non-zero variables.

We are interested in testing for the effect of a low-dimensional covariate Z when the model contains other higher dimensional covariates (X, S) which govern nuisance parameters. As in Fan and Li (2001) and Zou (2006), we work in the context that the number of covariates is large but smaller than the sample size, i.e., the dimensionality is fixed and does not increase with the sample size. For example, in logistic regression with a binary response Y , the model would be

$$\text{pr}(Y = 1|X, S, Z) = H(X^T \kappa + S^T \eta + Z^T \theta), \quad (1)$$

where $H(\cdot)$ is the logistic regression problem. The interest here is in testing the null hypothesis that $H_0 : \theta = 0$ and to produce an asymptotically valid significance level for that test. In model (1), (κ, η) is a nuisance parameter relating to (X, S) . One cannot use an oracle method to make valid probability statements about the null hypothesis, i.e., run SCAD or the Adaptive Lasso using all of (X, S, Z) and then somehow test for θ . This has been emphasized in a series of papers beginning with Leeb and Pötscher (2005), who note that the limit distribution of the oracle estimate of θ is not asymptotically normal under contiguous alternatives.

Our emphasis here is very different from the usual dimension reduction framework. The parameters (κ, η) are simply nuisance parameters and in this framework are not the main focus of interest.

As a motivating example for this article, we consider an extension of the model (1) that Chatterjee et al. (2006) described for testing the association of a disease outcome with an exposure when that exposure may be interacting with other co-factors that influence the risk of the disease. See also Maity, et al. (2009) for a semiparametric extension. In Chatterjee et al., Z could be a specific genetic or environment exposure of interest, X could be an array of genetic and/or environmental covariates of higher dimension that may be potentially interacting with Z and S could be certain basic covariates, such as age and sex, that the model needs to be adjusted for. The interest focuses on whether Z is associated with the disease outcome Y . To improve power, Chatterjee et al proposed using a parsimonious, Tukey's one-degree-of-freedom style, interaction model between Z and X of the

form

$$\text{pr}(Y = 1|X, S, Z, \gamma) = H(X^T \kappa + S^T \eta + Z^T \theta + \gamma X^T \kappa Z^T \theta), \quad (2)$$

where the scalar γ is meant to capture the interaction. A full description of this method is given in the Appendix. The null hypothesis of no association between Y and Z corresponds to testing $H_0 : \theta = 0$. A complication, however, is that under $\theta = 0$, the parameter γ also disappears from the model and hence is not identifiable from the data. Nevertheless, Chatterjee et al. noticed that for each fixed value of γ , the model (2) can be used to construct a valid score-test for $H_0 : \theta = 0$. They proposed to use the maximum of such score-statistics over a range of the parameter γ as the final test statistic. They observed that the score-test has particular computational advantages, because under the null hypothesis the model (2) reduces to a standard logistic regression model involving only main effects of X and S .

Motivated by model (2) and Chatterjee, et al. (2006), we study in this paper the following algorithm. Let the loglikelihood function be denoted by $\mathcal{L}(Y, X^T \kappa, S^T \eta, Z^T \theta, \zeta)$, where for example in linear regression ζ would be the regression error variance. The usual score test of course would fit the model $\mathcal{L}(Y, X^T \kappa, S^T \eta, 0, \zeta)$. An oracle would do the following, which we call an *oracle score test*.

- Under the null hypothesis $H_0 : \theta = 0$, fit the parameters (κ, η) by using an oracle estimator that consistently chooses the correct null model.
- Perform a score test using the selected components of (X, S) .

This paper addresses the simple question: does having access to an oracle under the null model improve the local power of the testing, i.e., is the oracle score test more powerful than the ordinary score test that does not do any model selection under the null? We will show that the answer to this depends on how (X, S) are correlated with Z : when they are independent, there is no point in having access to an oracle. We will demonstrate however that if (X, S) has information carrying elements that are highly correlated with Z , then having access to an oracle can be valuable.

An outline of this paper is as follows. In Section 2, we develop the general framework and state the main result on the local power of the score test. In Section 3, we discuss the implications of this result for oracle methods. In Section 4 we give simulation evidence that suggests that having access to an oracle may not be much help in a variety of practical problems. In Section 5, we briefly describe a setting that having access to an oracle will be useful, although we point out some peculiar behavior of the Adaptive Lasso implemented with 10-fold crossvalidation. All technical details are given in an Appendix.

2 Local Power of the Score Test

Here we compute the local power of the score test against alternatives of the form $\theta = \lambda n^{-1/2}$.

Suppose that the loglikelihood function is $\mathcal{L}(Y, X^T \kappa, S^T \eta, Z^T \theta, \zeta)$ and that the null hypothesis is $H_0 : \theta = 0$. In the general form, the function $\mathcal{L}(Y, v_1, v_2, v_3, v_4)$ has first derivatives $\mathcal{L}_j(Y, v_1, v_2, v_3, v_4) = \partial \mathcal{L}(Y, v_1, v_2, v_3, v_4) / \partial v_j$ and second derivatives $\mathcal{L}_{jk}(Y, v_1, v_2, v_3, v_4) = \partial^2 \mathcal{L}(Y, v_1, v_2, v_3, v_4) / \partial v_j \partial v_k$. Let $\beta = (\kappa^T, \eta^T, \zeta^T)^T$. Let the Fisher information matrix for (θ, β) be partitioned to have diagonal elements $\mathcal{I}_{\theta\theta}$ and $\mathcal{I}_{\beta\beta}$ with top right hand off-diagonal element $\mathcal{I}_{\theta\beta}$. Under the null hypothesis, with $(\bullet) = (Y, X^T \kappa, S^T \eta, 0, \zeta)$, the Fisher information matrix is

$$\begin{aligned} \mathcal{I} &= \begin{bmatrix} \mathcal{I}_{\theta\theta} & \mathcal{I}_{\theta\beta} \\ \mathcal{I}_{\beta\theta} & \mathcal{I}_{\beta\beta} \end{bmatrix} \\ &= -E \begin{bmatrix} ZZ^T \mathcal{L}_{33}(\bullet) & ZX^T \mathcal{L}_{13}(\bullet) & ZS^T \mathcal{L}_{23}(\bullet) & Z\mathcal{L}_{34}(\bullet) \\ XZ^T \mathcal{L}_{13}(\bullet) & XX^T \mathcal{L}_{11}(\bullet) & XS^T \mathcal{L}_{12}(\bullet) & X\mathcal{L}_{14}(\bullet) \\ SZ^T \mathcal{L}_{23}(\bullet) & SX^T \mathcal{L}_{12}(\bullet) & SS^T \mathcal{L}_{22}(\bullet) & S\mathcal{L}_{24}(\bullet) \\ Z^T \mathcal{L}_{34}(\bullet) & X^T \mathcal{L}_{14}(\bullet) & S^T \mathcal{L}_{24}(\bullet) & \mathcal{L}_{44}(\bullet) \end{bmatrix}. \end{aligned}$$

Further, define $\Omega = \mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta} \mathcal{I}_{\beta\beta}^{-1} \mathcal{I}_{\beta\theta}$. Let $\hat{\beta}(\theta_n)$ be the maximum likelihood estimate of β assuming that the null hypothesis is true but when θ_n is the true parameter value. Let $\hat{\Omega}$ be the usual estimate of Ω with the Fisher information matrix estimated under the null hypothesis. Define

$$\mathcal{J}_n = n^{-1/2} \sum_{i=1}^n Z_i \mathcal{L}_3(Y_i, X_i^T \hat{\kappa}, S_i^T \hat{\eta}, 0, \hat{\zeta}).$$

The score test statistic is $\mathcal{T}_n = \mathcal{J}_n^T \hat{\Omega}^{-1} \mathcal{J}_n$. Under the null hypothesis the score test statistic converges $\rightarrow \chi_{p_\theta}^2(0)$, the central chisquared distribution with p_θ degrees of freedom, where p_θ is the dimension of θ . The following is a well-known but useful result.

Result 1 *Under the local alternatives $\theta = \lambda n^{-1/2}$, $\mathcal{T}_n \rightarrow \chi_{p_\theta}^2(\lambda^T \Omega \lambda / 2)$, the non-central chisquared distribution with p_θ degrees of freedom and noncentrality parameter $\lambda^T \Omega \lambda / 2$.*

3 Implications for Oracles

We are going to show that if Z and (X, S) are independent, under two simple assumptions that are usually satisfied, using an oracle to reduce the dimension of

the fit yields no increase in local power when compared to using all components of (X, S) , and not just the ones that predict the response. This has an important implication. In perhaps the vast majority of cases, one will not expect information-carrying components (X, S) to be highly correlated with Z . Thus, in much of actual practice, oracle estimation when doing a score test will not improve power substantially.

We will make the following assumptions. Both Assumptions 1-2 hold for the score tests for models (1) and (2).

Assumption 1 Under the null hypothesis, the distribution of Y depends on (X, S, Z) only through $(X^T \kappa, S^T \eta)$.

Assumption 2 The score test statistic is invariant to location changes in Z , so that for example it has the same value whether Z or $Z - E(Z)$ is used.

Result 2 Suppose that Z and (X, S) are independent, and that Assumptions 1-2 hold. Then under the contiguous alternatives that $\theta = \theta_n = \lambda n^{-1/2}$, the local power of the score test depends on $W = (X, S)$ only through $(X^T \kappa, S^T \eta)$.

Using Result 2 our main result follows:

Theorem 1 Make the same assumptions as in Result 2, including that Z is independent of (X, S) . Partition β into information and non-information carrying components, so that $\beta = (\beta_1, \beta_2 = 0)$. Partition the components of $W = (X, S)$ similarly as (W_1, W_2) . Then, in the score test, knowing that $\beta_2 = 0$ and using only W_1 does not affect the local power of the score test. It follows that oracle estimators of β will have no more local power than the naive test that does not attempt dimension reduction.

Corollary 1 Using Results 1-2 and Result 3 in Appendix A.1, if Z is not independent of (X, S) , then in general oracle estimators will have greater power.

It is tempting to believe that just because (X, S) are independent of Z , then estimates of $\beta = (\kappa^T, \eta^T, \zeta^T)^T$ are independent of estimates of θ . This is not the case under alternatives in general, including in our example.

4 Simulations

In this section, we report two simulations of the Chatterjee method of model (2). The method itself is described in Appendix A.3. In all cases, 1,000 data sets were randomly generated. In all cases, we first generated X_* to have dimension $p_\kappa = 100$, S to have dimension $p_\eta = 10$ and Z_* was scalar. In all cases, marginally X_* and S_* were normally distributed with mean zero, variance 1 and common correlation 0.50, while Z_* was normally distributed with mean zero and variance one. We then convert X and Z to SNP-type data, which take on the values 0, 1 and 2. More precisely, the individual components of X and Z had value 1.0 with probability 0.27, and they had value 2.0 with probability 0.20. The data were then generated by model (2) with $\gamma = 2$.

The sample size was $n = 5,000$. All components of κ were zero except for the first three, which had values (1.0, 1.0, 0.5). All components of η were zero except the first, which had value 0.75. The values of θ varied from the null, $\theta = 0$, to $\theta = 0.07$ in steps of 0.01. The Lasso using the penalized package in R, (Goeman, J. J., 2009a; Goeman, J. J., 2009b), and the Adaptive Lasso using the glmnet package in R, (Friedman J. et al. 2009), were used to estimate β under the null hypothesis. We also refitted β using only the four interesting predictors, as well as using all 110 components of (X, S) .

We point out that while our emphasis is on statistical power, we also evaluate the Type I error of the tests. As the results below show, all the methods have reasonable Type I error rate control.

4.1 The Case that (X, S, Z) are Mutually Independent

In the first simulation, (X_*, S_*, Z_*) were mutually independent. According to our Theorem 1, we expect no gain in power for being an oracle. This is confirmed by Table 1, where the largest single difference in power was 0.026, although there is consistently a very small power increase with an oracle.

4.2 The Case that (X, S, Z) are Mutually Dependent

In the second simulation, we set all components of (X_*, S_*, Z_*) to have correlation 0.50. In this case, the theory predicts that the oracle method and the method that uses all 110 predictors will have different power, see Corollary 1. In Table 2 we produce the results. There is some evidence of very modest increased power for the oracle method, with the biggest difference in power being when $\theta = 0.05$, this is an increase of 0.051. The adaptive Lasso is very close to the oracle. Disappointingly,

θ	All 110		Adaptive	
	Oracle	Predictors	Lasso	Lasso
0.00	0.040	0.039	0.038	0.040
0.01	0.103	0.098	0.094	0.096
0.02	0.248	0.240	0.249	0.240
0.03	0.509	0.491	0.495	0.493
0.04	0.754	0.744	0.745	0.752
0.05	0.908	0.882	0.893	0.908
0.06	0.981	0.973	0.971	0.965
0.07	0.997	0.991	0.997	0.989

Table 1: Simulation of power for a 5%-level test in the case that (X, S, Z) are mutually independent. Here the theory says that all methods should have the same local power.

in the second simulation, the Lasso method actually loses considerable power compared to using all 110 predictors, or compared to the adaptive Lasso. It is not clear why this is happening.

The lack of a great increase in power for the oracle estimator in this second simulation is predictable from our theory. Write $W = (X^T, S^T)^T$ and remember that $\beta = (\kappa^T, \eta^T)^T$. Fix γ . Let $H^{(1)}(x) = H(x)\{1 - H(x)\}$. Detailed calculations based on Theorem 1 show that the elements of the information matrix are given as

$$\begin{aligned}\mathcal{I}_{\theta\beta}(0, \beta_0) &= E \left\{ (1 + \gamma X^T \kappa) H^{(1)}(W^T \beta) Z W^T \right\}; \\ \mathcal{I}_{\beta\beta}(0, \beta_0) &= E \left\{ H^{(1)}(W^T \beta) W W^T \right\}; \\ \mathcal{I}_{\theta\theta}(0, \beta_0) &= E \left\{ (1 + \gamma X^T \kappa)^2 H^{(1)}(W^T \beta) Z Z^T \right\},\end{aligned}$$

from which Ω and the noncentrality parameter can be calculated from Result 1.

In our simulation study, with $\gamma = 2$ and $\theta = 0.05$, the noncentrality parameter when using all 110 predictors is 0.42, while it is 0.44 for the oracle. Such a modest increase in noncentrality is reflected in the very modest power gains for the oracle in the second simulation. If instead we make the common correlation in the second simulation equal to 0.90, the noncentrality parameter is 0.20 and 0.22 for the 110 variables and the oracle, respectively.

θ	All 110		Adaptive	
	Oracle	Predictors	Lasso	Lasso
0.00	0.060	0.061	0.052	0.067
0.01	0.099	0.101	0.072	0.097
0.02	0.216	0.218	0.144	0.206
0.03	0.389	0.373	0.293	0.387
0.04	0.614	0.581	0.482	0.608
0.05	0.819	0.768	0.680	0.802
0.06	0.933	0.891	0.822	0.938
0.07	0.976	0.963	0.936	0.967

Table 2: Simulation of power for a 5%-level test in the case that (X, S, Z) are mutually correlated. Here the theory says that oracle methods should have higher local power than the method that uses all 110 predictors.

5 High Correlation and Information-Carrying Covariates

Our main purpose has been to show that there will be little to be gained by having access to an oracle if the secondary variables are not highly correlated with the variables of interest. However, as a secondary matter, it is not always the case that having access to an oracle will not help in terms of power. Using our theory, we developed cases where X had 15 components, only 2 of which were information-carrying. In addition, Z had three components, one of which was also a predictor. One of the information-carrying components of X was uncorrelated with Z , while the other had correlation > 0.85 for each component of Z .

We investigated two cases, one in which the information-carrying components of X had relative risks ≈ 4.0 , and the other in which the relative risks were a more modest ≈ 0.5 . Here there was a decided increase in power if an oracle was available. Disappointingly, the Adaptive Lasso with 10-fold crossvalidation was only modestly more powerful than the approach that used all covariates, and much less powerful than the oracle. The Lasso was essentially equivalent to using all the covariates. It is possible, and indeed we hope, that newer oracle methods such as the Adaptive Elastic Net (Zou and Zhang, 2009) that deal with collinearity better than the Adaptive Lasso will lead to improved performance in these cases, although currently there is no software available in R for this method. The elastic net (Zou and Hastie, 2005) is its non-oracle version, and in these simulations was essentially equivalent to the Adaptive Lasso.

6 Discussion

We believe that using regularization methods such as the Adaptive Lasso is a natural idea in score testing for a primary variable when there are many secondary variables. Our simulations and empirical work demonstrate that if the secondary variables are not highly correlated with the variables of interest, or if these secondary variables are not informative about the response, then little will be gained by having an access to an oracle. Our theoretical numerical exercises reveal the possibility of power increases with strong correlation between highly informative secondary variables and the primary variables of interest. This situation may not commonly occur. As a point of future research, in these settings that power gains are possible, the Adaptive Lasso with 10-fold crossvalidation had fairly disappointing behavior.

It would be interesting to find empirical examples of score testing where oracle methods actually do achieve significantly greater power in practice.

Appendix: Sketch of Technical Arguments

A.1 General Considerations

Consider a general problem where the loglikelihood function is $\mathcal{L}(\mathcal{Y}, \theta, \beta)$. The null hypothesis is $H_0 : \theta = 0$, and we consider contiguous alternatives of the form $\theta = \theta_n = \lambda n^{-1/2}$. Let $\hat{\beta}(\theta_n)$ be the estimator computed as if the null hypothesis were true with data generated under the contiguous alternative.

Subscripts will denote derivatives, e.g., $\mathcal{L}_\theta(\mathcal{Y}, \theta, \beta) = \partial \mathcal{L}(\mathcal{Y}, \theta, \beta) / \partial \theta$ and $\mathcal{L}_{\theta\beta}(\mathcal{Y}, \theta, \beta) = \partial^2 \mathcal{L}(\mathcal{Y}, \theta, \beta) / \partial \theta \partial \beta^T$. Define $\mathcal{J}_{\theta\theta}(\theta, \beta) = -E\{\mathcal{L}_{\theta\theta}(\mathcal{Y}, \theta, \beta)\}$ and similarly for $\mathcal{J}_{\theta\beta}(\theta, \beta)$, $\mathcal{J}_{\beta\theta}(\theta, \beta)$, and $\mathcal{J}_{\beta\beta}(\theta, \beta)$. These components of the Fisher information are estimated as

$$\hat{\mathcal{J}}_{\theta\theta}(\theta, \beta) = -n^{-1} \sum_{i=1}^n \mathcal{L}_{\theta\theta}(\mathcal{Y}_i, \theta, \beta).$$

The numerator of the score statistic is

$$\mathcal{S}_n = n^{-1/2} \sum_{i=1}^n \mathcal{L}_\theta\{\mathcal{Y}_i, 0, \hat{\beta}(\theta_n)\}. \quad (\text{A.1})$$

Define $\Omega = \mathcal{J}_{\theta\theta}(0, \beta_0) - \mathcal{J}_{\theta\beta}(0, \beta_0) \mathcal{J}_{\beta\beta}^{-1}(0, \beta_0) \mathcal{J}_{\beta\theta}(0, \beta_0)$ and let $\hat{\Omega}$ be its estimate. The score test statistic is $\mathcal{T}_n = \mathcal{S}_n^T \hat{\Omega}^{-1} \mathcal{S}_n$.

Result 3 Let p_θ be the dimension of θ . Under contiguous alternatives $\theta = \theta_n = \lambda n^{-1/2}$,

$$\begin{aligned}\mathcal{S}_n &\rightarrow \text{Normal}(-\Omega\lambda, \Omega); \\ \mathcal{T}_n &\rightarrow \chi_{p_\theta}^2(\lambda^\top \Omega \lambda / 2),\end{aligned}$$

where $\chi_p^2(a)$ is the noncentral chisquared distribution.

Here is a sketch of the argument for Result 3. Under the null hypothesis, it is well known that the numerator of the score statistic has the expansion

$$\mathcal{S}_n = n^{-1/2} \sum_{i=1}^n \left\{ \mathcal{L}_\theta(\mathcal{Y}_i, 0, \beta_0) - \mathcal{J}_{\theta\beta}(0, \beta_0) \mathcal{J}_{\beta\beta}^{-1}(0, \beta_0) \mathcal{L}_\beta(\mathcal{Y}_i, 0, \beta_0) \right\} + o_p(1),$$

where β_0 is the true value of β .

To apply LeCam's third Lemma to obtain the distribution of the numerator of the score statistic under the alternatives $\theta = \theta_n$, we note that by a Taylor series expansion, under the null hypothesis,

$$\begin{aligned}\mathcal{Q}_n &= \sum_{i=1}^n \{ \mathcal{L}(\mathcal{Y}_i, \theta_n, \beta_0) - \mathcal{L}(\mathcal{Y}_i, 0, \beta_0) \} \\ &= \sum_{i=1}^n \mathcal{L}_\theta^\top(\mathcal{Y}_i, 0, \beta_0)(\theta_n - 0) + \\ &\quad (1/2)(\theta_n - 0)^\top \sum_{i=1}^n \mathcal{L}_{\theta\theta}^\top(\mathcal{Y}_i, 0, \beta_0)(\theta_n - 0) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^n \lambda^\top \mathcal{L}_\theta(\mathcal{Y}_i, 0, \beta_0) - (1/2) \lambda^\top \mathcal{J}_{\theta\theta}(0, \beta_0) \lambda + o_p(1). \quad (\text{A.2})\end{aligned}$$

This means that under the null hypothesis, $(\mathcal{S}_n, \mathcal{Q}_n)$ are jointly normally distributed with means 0 and $-(1/2)\lambda^\top \mathcal{J}_{\theta\theta}(0, \beta_0)\lambda$, respectively. Their variances are $\Omega = \mathcal{J}_{\theta\theta}(0, \beta_0) - \mathcal{J}_{\theta\beta}(0, \beta_0) \mathcal{J}_{\beta\beta}^{-1}(0, \beta_0) \mathcal{J}_{\beta\theta}(0, \beta_0)$ and $\lambda^\top \mathcal{J}_{\theta\theta}(0, \beta_0)\lambda$, respectively, and their covariance is $-\Omega\lambda$. Applying LeCam's third Lemma shows that when $\theta = \theta_n$, $\mathcal{S}_n \rightarrow \text{Normal}(-\Omega\lambda, \Omega)$. It then follows that $\mathcal{T}_n \rightarrow \chi_{p_\theta}^2(\lambda^\top \Omega \lambda / 2)$, as claimed.

Result 1 is a simple consequence of these results.

A.2 Proof of Theorem 1

We merely need to show that $\mathcal{J}_{\theta\beta} = 0$ and that $\mathcal{J}_{\theta\theta}$ depends on the distribution of (X, S) through $(X^\top \kappa, S^\top \eta)$.

Recalling Assumption 2, we can assume that $E(Z) = 0$. Because of Assumption 1, under the null hypothesis, $E(\mathcal{L}_{jk}(Y, X^T \kappa, S^T \eta, 0, \zeta) | X, S, Z) = \mathcal{G}_{jk}(X^T \kappa, S^T \eta, \zeta)$ say. Letting $(\bullet) = (X^T \kappa, S^T \eta, \zeta)$, we can now compute the Fisher information matrix under the null hypothesis as

$$\begin{aligned} \mathcal{I} &= \begin{bmatrix} \mathcal{I}_{\theta\theta} & \mathcal{I}_{\theta\beta} \\ \mathcal{I}_{\beta\theta} & \mathcal{I}_{\beta\beta} \end{bmatrix} \\ &= -E \begin{bmatrix} ZZ^T \mathcal{G}_{33}(\bullet) & ZX^T \mathcal{G}_{13}(\bullet) & ZS^T \mathcal{G}_{23}(\bullet) & Z\mathcal{G}_{34}(\bullet) \\ XZ^T \mathcal{G}_{13}(\bullet) & XX^T \mathcal{G}_{11}(\bullet) & XS^T \mathcal{G}_{12}(\bullet) & X\mathcal{G}_{14}(\bullet) \\ SZ^T \mathcal{G}_{23}(\bullet) & SX^T \mathcal{G}_{12}(\bullet) & SS^T \mathcal{G}_{22}(\bullet) & S\mathcal{G}_{24}(\bullet) \\ Z^T \mathcal{G}_{34}(\bullet) & X^T \mathcal{G}_{14}(\bullet) & S^T \mathcal{G}_{24}(\bullet) & \mathcal{G}_{44}(\bullet) \end{bmatrix}. \end{aligned}$$

Since Z is independent of (X, S) , and since $E(Z) = 0$, it is obvious that $\mathcal{I}_{\theta\beta} = 0$, and hence that $\Omega = \text{cov}(Z)E\{\mathcal{G}_{33}(X^T \kappa, S^T \eta, \zeta)\}$, completing the proof.

A.3 The Chatterjee Method of Score Testing in Genomics

Here are the details of the Chatterjee procedure for model (2). Let $\beta = (\kappa^T, \eta^T)^T$ and $W = (X^T, S^T)^T$. The normalized score for estimating θ_0 when evaluated at the null hypothesis $\theta_0 = 0$ in the logistic context is

$$\mathcal{S}_n(\beta, \gamma) = n^{-1/2} \sum_{i=1}^n (1 + \gamma X_i^T \kappa) Z_i \{Y_i - H(W_i^T \beta)\}. \quad (\text{A.3})$$

The idea is that for each fixed γ , estimate $\beta_0 = (\kappa_0, \eta_0)$ by maximum likelihood at the null model, calling that estimate $\hat{\beta}$. Let $H^{(1)}(x) = H(x)\{1 - H(x)\}$. Define $\mathcal{A}(\beta_0) = E\{WW^T H^{(1)}(W^T \beta_0)\}$, $\mathcal{B}(\beta_0, \gamma) = E\{Z(1 + \gamma X^T \kappa_0)H^{(1)}(W^T \beta_0)W^T\}$, $\mathcal{G}(X, Z, S, \beta_0, \gamma) = Z(1 + \gamma X^T \kappa_0) - \mathcal{B}(\beta_0, \gamma)\mathcal{A}^{-1}(\beta_0)W$ and $\mathcal{K}(\beta_0, \gamma) = E\{\mathcal{G}(X, Z, S, \beta_0, \gamma)\mathcal{G}^T(X, Z, S, \beta_0, \gamma)H^{(1)}(W^T \beta_0)\}$. Further define $U(\gamma) = \mathcal{S}_n^T(\beta, \gamma)\mathcal{K}^{-1}(\beta, \gamma)\mathcal{S}_n(\beta, \gamma)$. All these terms are estimated by replacing true parameters by their estimates under the null model and expectations by averages over the data.

Chatterjee, et al. propose as a test statistic to reject the null hypothesis for large values of

$$\widehat{\mathcal{U}}_n = \max_{a \leq \gamma \leq b} \widehat{U}(\gamma), \quad (\text{A.4})$$

where they show that for each γ , under the null hypothesis $\widehat{U}(\gamma) \rightarrow \chi_{p_\theta}^2$, where p_θ is the dimension of θ . They also show how to compute p-values using only simulation,

as follows. Let $b = 1, \dots, B$, and for any b , let $(\mathcal{Z}_{b1}, \dots, \mathcal{Z}_{bn})$ be randomly generated standard normal random variables. Define

$$\mathcal{T}_b(\gamma) = n^{-1/2} \sum_{i=1}^n \mathcal{G}_i(X_i, Z_i, S_i, \hat{\beta}, \gamma) \left\{ Y_i - H(W_i^T \hat{\beta}) \right\} \mathcal{Z}_{bi} \quad (\text{A.5})$$

$$\mathcal{U}_b = \max_{a \leq \gamma \leq b} \mathcal{T}_b^T(\gamma) \hat{\mathcal{K}}^{-1}(\hat{\beta}, \gamma) \mathcal{T}_b(\gamma). \quad (\text{A.6})$$

Then, asymptotically, under the null hypothesis, $\mathcal{U}_b$ has the same limit distribution as does $\hat{\mathcal{U}}_n$ in (A.4). Hence, the p-value is just

$$\hat{p} = B^{-1} \sum_{b=1}^B I(\hat{\mathcal{U}}_n \leq \mathcal{U}_b). \quad (\text{A.7})$$

References

- Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006). Powerful multi-locus tests for genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics*, 79, 1002-1016.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.
- Friedman J., Hastie T. and Tibshirani R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.1-3. <http://www-stat.stanford.edu/hastie/Papers/glmnet.pdf>.
- Goeman, J. J. (2009a) Penalized R package, version 0.9-24.
- Goeman, J. J. (2009b). L1 penalized estimation in the Cox proportional hazards model, *Biometrical Journal*, 51, in press.
- Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, 21, 21-59.
- Maity, A., Carroll, R. J., Mammen, E. and Chatterjee, N. (2009). Powerful multi-locus tests for genetic association with semiparametric gene-environment interactions. *Journal of the Royal Statistical Society, Series B*, 71, 75-96.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

- Zou, H. (2006). The Adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320.
- Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37, 1733-1751.