Sankhyā : The Indian Journal of Statistics
2021, Volume 83-B, Part 1, pp. 170-184

Model Selection With Mixed Variables on the Lasso Path

X. Jessie Jeng, Huimin Peng and Wenbin Lu
North Carolina State University, Raleigh, USA

Abstract

Among the most popular model selection procedures in high-dimensional regression, Lasso provides a solution path to rank the variables and determines a cut-off position on the path to select variables and estimate coefficients. In this paper, we consider variable selection from a new perspective motivated by the frequently occurred phenomenon that relevant variables are often mixed with noise variables on the solution path. We propose to characterize the positions of the first noise variable and the last relevant variable on the path. We then develop a new variable selection procedure to control over-selection of the noise variables ranking after the last relevant variable, and, at the same time, retain a high proportion of relevant variables ranking before the first noise variable. Our procedure utilizes the recently developed covariance test statistic and Q statistic in post-selection inference. In numerical examples, our method compares favorably with existing methods in selection accuracy and the ability to interpret its results.

AMS (2000) *subject classification.* Primary: 62J07; Secondary: 62F07.
Keywords and phrases. Jayanta K. Ghosh, Large p small n, Penalized regression, Variable selection

1 Introduction

The authors of the paper deeply mourn the loss of Dr. Jayanta K. Ghosh, whose dedication to research and mentoring have benefited generations of statisticians and who has set an eminent example of excellence as a scholar and role model. Dr. Ghosh has made substantial contributions to a wide range of research areas such as higher order asymptotics, Bayesian analysis, and high-dimensional inference. One of the authors, X.J. Jeng, was fortunate to have Dr. Ghosh as her PhD advisor in Purdue University. Dr. Ghosh's work in model selection, multiple testing, and their biomedical applications (e.g. Wilbur et al., 2002; Chakrabarti and Ghosh, 2007, 2011; Bogdan et al., 2008, 2011) has inspired the research in this paper.

Supported in part by NSF Grant DMS-1811360.

We consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}), \tag{1.1}$$

where $\mathbf{y}$ is a vector of response with length n, $\mathbf{X}$ is the $n \times p$ design matrix of standardized predictors, and $\boldsymbol{\beta}^*$ a sparse vector of coefficients. In high-dimensional regression, p can be greater than n. Among the most popular methods for model selection and estimation in the high-dimensional regression, Lasso (Tibshirani, 1996) solves the following optimization problem

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\beta \in \mathcal{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{1.2}$$

where $\lambda \geq 0$ is a tuning parameter. Lasso provides a solution path, which is the plot of the estimate $\hat{\beta}(\lambda)$ versus the tuning parameter λ.

Lasso solution path is piecewise linear with each knot corresponding to the entry of a variable into the selected set. Knots are denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$, where $m = \min(n-1, p)$ is the length of the solution path (Efron et al., 2004).

Recent developments in high-dimensional regression focus on hypothesis testing for variable selection. Impressive progress has been made in Zhang and Zhang (2014), Van De Geer et al. (2014), Lockhart et al. (2014), Barber and Candès (2015), Bogdan et al. (2015), Lee et al. (2016), G'sell et al. (2016), & Jeng and Chen (2019a), etc. Specifically, innovative test statistics based on Lasso solution path have been proposed. For example, Lockhart et al. (2014) construct the covariance test statistic as follows. Along the solution path, the variable indexed by j_k enters the selected model at knot λ_k. Define active set right before knot λ_k as $A_k = \{j_1, j_2, \cdots, j_{k-1}\}$. In addition, define true active set to be $A^* = \{j : \beta_j^* \neq 0\}$ and the size of true active set as $s = |A^*|$. Lockhart et al. (2014) considers to test the null hypothesis $H_{0k} : A^* \subset A_k$ conditional upon the active set A_k at knot λ_k and propose the covariance test statistic as

$$T_k = \left(\langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1}) \rangle - \langle \mathbf{y}, \mathbf{X}_{A_k}\tilde{\boldsymbol{\beta}}_{A_k}(\lambda_{k+1}) \rangle \right) / \sigma^2, \tag{1.3}$$

where $\hat{\boldsymbol{\beta}}(\lambda_{k+1}) = \arg\min_{\beta \in \mathcal{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_{k+1}\|\boldsymbol{\beta}\|_1$ and $\tilde{\boldsymbol{\beta}}_{A_k}(\lambda_{k+1}) = \arg\min_{\beta \in \mathcal{R}^{|A_k|}} \frac{1}{2}\|$ $\mathbf{y} - \mathbf{X}_{A_k}\boldsymbol{\beta}_{A_k}\|_2^2 + \lambda_{k+1}\|\boldsymbol{\beta}_{A_k}\|_1$.

Lockhart et al. (2014) derived that under orthogonal design, if all s relevant variables rank ahead of noise variables with probability tending to 1, then for any fixed d,

$$(T_{s+1}, T_{s+2}, \cdots, T_{s+d}) \xrightarrow{d} (\text{Exp}(1), \text{Exp}(1/2), \cdots, \text{Exp}(1/d)), \tag{1.4}$$

as $p \to \infty$, and that $T_1, T_2, \cdots, T_d$ are asymptotically independent.

Later, G'sell et al. (2016) proposed to perform sequential test on $H_{0k} : A^* \subset A_k$ for k increasing from 0 to m and developed the Q statistics for a stopping rule. The Q statistics are defined as

$$q_k = \exp\left(-\sum_{j=k}^{m} T_j\right) \tag{1.5}$$

for $k = 1, \ldots, m$. It has been proved that in the case of perfect separation where all s relevant variables rank ahead of noise variables, if $T_{s+1}, \cdots, T_m$ are independently distributed as $(T_{s+1}, \cdots, T_m) \sim (\text{Exp}(1), \text{Exp}(1/2), \cdots \text{Exp}(1/(m-s)))$, then

$$q_k \overset{d}{=} (k-s)\text{th order statistic of } m-s \text{ independent standard uniform random variables} \tag{1.6}$$

for $s + 1 \leq k \leq m$. G'sell et al. (2016) developed a stopping rule (TailStop) implementing the Q statistics in the procedure of Benjamini and Hochberg (1995). Given the distribution of q_k in Eq. 1.6, TailStop controls false discovery rate at a pre-specified level.

In this paper, we consider more general scenarios where relevant variables and noise variables are not perfectly separated and some (or all) relevant variables intertwine with noise variables on the Lasso path. Such scenarios would occur when the effect sizes of relevant variables are not large enough. In fact, even with infinitely large effect size, perfect separation on solution path is still not guaranteed when the number of relevant variables is relatively large (Wainwright, 2009; Su et al., 2017; Jeng and Chen, 2019). Studies in theory and method are limited in such general scenarios because the inseparability among relevant and noise variables make it difficult to estimate Type I and/or Type II errors. In order to perform variable selection in the more general and realistic settings, we propose a new theoretical framework to characterize the region on the solution path where relevant and noise variables are not distinguishable.

Figure 1 illustrates the indistinguishable region on solution path. Denote m_0 as the position right before the first noise variable on the path such that all entries up to m_0 correspond to relevant variables and m_1 as the position of the last relevant variable where all entries afterwards correspond to noise variables. Given a solution path, both m_0 and m_1 are realized but unknown, and the region between m_0 and m_1 is referred to as the indistinguishable region.

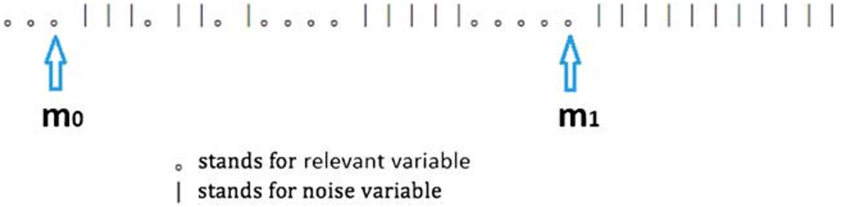

Figure 1: An example of m_0 and m_1 on Lasso solution path. m_0 is the entry right before the first noise variable. m_1 is the entry of the last relevant variable

Given the solution path, a sensible variable selection procedure would select all variables up to m_0 but no variables after m_1. Since both m_0 and m_1 are unknown stochastic quantities, the selection procedure should automatically adapt to the unknown m_0 and m_1.

We develop a new variable selection procedure utilizing the Q statistic in Eq. 1.5. We refer to the new procedure as Q-statistic Variable Selection (QVS). QVS searches through the solution path and determines a cut-off position that is likely between m_0 and m_1 under certain conditions that are more general than Eq. 1.6 on the distribution of the Q statistic.

2 Method and Theory

QVS is inspired by earlier works on estimating the proportion of non-null component in a mixture model of p-value (Meinshausen and Buhlmann, 2005, 2006). We extend the technique to high-dimensional regression considering the general scenarios with indistinguishable relevant and noise variables.

Given a Lasso path, QVS searches through the empirical process $k/m - q_k - c_m\sqrt{q_k(1-q_k)}, 1 \leq k \leq m$, where q_k is defined in Eq. 1.5, and determines the cut-off position as

$$\hat{k}_{qvs} = m \cdot \max_{1\leq k\leq m/2}\left\{\frac{k}{m} - q_k - c_m\sqrt{q_k(1-q_k)}\right\}, \tag{2.1}$$

where c_m is a bounding sequence to control over selection of noise variables after m_1 and constructed as follows. For $0 < t < 1$, let

$$U_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1(U_i \leq t),$$

where $U_1, \cdots, U_m$ are i.i.d. standard uniform random variables. Further, let

$$V_m = \sup_{t\in(0,1)} \frac{U_m(t) - t}{\sqrt{t(1-t)}}. \tag{2.2}$$

Then determine c_m as an upper bound of V_m so that $P(V_m > c_m) < \alpha_m \to 0$ as $m \to \infty$.

We consider the setting where some relevant variables intertwine with noise variables on the Lasso path. In order to gain theoretical insights for the performance of QVS, we adopt a similar strategy as in Section 4.2.1 of G'sell et al. (2016), and simplify the problem by considering a sequence of arbitrary statistics $q_1, \ldots, q_m$ corresponding to the m ranked variables.

Define $U_{(1),m-s}, \ldots, U_{(m-s),m-s}$ as increasingly ordered statistics of $m-s$ independent standard uniform random variables, independent of $q_1 \ldots, q_m$. Assume that

$$q_{m_0+1} \le U_{(1),m-s} \tag{2.3}$$

with probability tending to 1 as $m \to \infty$. It is easy to see that in the special case of perfect separation with $m_0 = m_1 = s$, Eq. 2.3 is implied by condition (1.6) from G'sell et al. (2016). In the more general case with $m_0 \neq m_1$, we show that QVS provides an asymptotic upper bound for the unknown m_0 under (2.3).

Theorem 1. *Consider the stopping rule in* Eq. 2.1 *under condition* (2.3). *Assume that the number of relevant variables $s = o(m)$ and that $\sqrt{\log m}/m_0 \to 0$ in probability. Then*

$$P\left(\hat{k}_{qvs} \ge (1-\epsilon)m_0\right) \to 1 \tag{2.4}$$

as $m \to \infty$ for any small constant $\epsilon > 0$.

The proof of Theorem 1 is provided in Appendix A. The condition, $\sqrt{\log m}/\ m_0 \to 0$ in probability, holds when m_0 is large enough, which may be satisfied if the number of relevant variables and their effect sizes are large enough. The result in Theorem 1 implies that QVS can asymptotically retain a high proportion of relevant variables ranked up to m_0.

Next, we show the property of QVS to provide a lower bound for m_1 in Fig. 1. Recall the definition of $U_{(j),m-s}$ as the jth ordered statistic of $m-s$ independent standard uniform random variables. Assume that for any $t \in (0, 1)$,

$$\sum_{k=m_1+1}^{m} 1(q_k \le t) \le \sum_{k=m_1-s+1}^{m-s} 1(U_{(k),m-s} \le t), \tag{2.5}$$

with probability tending to 1 as $m \to \infty$, where $1(\cdot)$ denotes an indicator function. Note that in the special case with $m_0 = m_1 = s$, condition (2.5)) is implied by Eq. 1.6 because the latter assumes that q_k has the same distribution as that of $U_{(k-s),m-s}$ for $k = s+1, \ldots, m$. In the more general setting when $m_0 \neq m_1$, we have the following result.

Theorem 2. *Consider the stopping rule in* Eq. 2.1 *under condition* (2.5). *As $m \to \infty$,*

$$P(\hat{k}_{qvs} \leq m_1) \to 1.$$

The proof of Theorem 2 is presented in Appendix B. Theorem 2 implies that QVS provides a parsimonious variable selection such that noise variables ranked after m_1 are not likely to be over-selected. Combining Theorem 1 and 2, the following corollary is straightforward.

Corollary 1. *Consider the stopping rule in* Eq. 2.1 *under conditions* (2.3) *and* (2.5). *If $s = o(m)$ and $\sqrt{\log m}/m_0 \to 0$ in probability, then*

$$P\left((1-\epsilon)m_0 \leq \hat{k}_{qvs} \leq m_1\right) \to 1$$

as $m \to \infty$ for any small constant $\epsilon > 0$.

3 Simulation

In our simulation study, design matrix $\mathbf{X}_{n\times p}$ is a Gaussian random matrix with each row generated from $N_p(0, \mathbf{\Sigma})$. Response $\mathbf{y}$ is generated from $N_n(\mathbf{X}\boldsymbol{\beta}^*, \mathbf{I})$, where $\boldsymbol{\beta}^*$ is the vector of true coefficients. The locations of non-zero entries of $\boldsymbol{\beta}^*$ are randomly simulated.

For the QVS procedure, we simulate the bounding sequence c_m by the following steps. We generate $\mathbf{X}_{n\times p}$ and $\mathbf{y}_{n\times 1}$ under the null model and compute the Lasso solution path using the *lars* package in R. Covariance test statistics and Q statistics $\{q_i\}_{i=1}^m$ are calculated by Eqs. 1.3 and 1.5, respectively. Then, we compute V_m using $V_m = \max_{1\leq i\leq m/2}(i/m - q_i)/\sqrt{q_i(1-q_i)}$. We repeat the above steps for 1000 times and obtain $V_m^1, V_m^2, \cdots, V_m^{1000}$. The bounding sequence c_m is computed as the upper α_m percentile of $V_m^1, V_m^2, \cdots,$ V_m^{1000}. We set $\alpha_m = 1/\sqrt{\log m}$ as recommended in Jeng et al. (2019) to bound the exceedance probability of V_m at a degenerating level. For each combination of sample size n and dimension p, we only need to simulate the bounding sequence once.

3.1. Positions of $\hat{k}_{qvs}$, m_0, and m_1 Recall the definitions of the m_0 and m_1 on the Lasso path and Fig. 1. Table 1 reports the realized values of $\hat{k}_{qvs}$, m_0, m_1. It can be seen that the distance between m_0 and m_1 increases as the number of relevant variables s increases. In all the cases, $\hat{k}_{qvs}$ is greater than m_0, which agrees with the theoretical property of QVS in Theorem 1. On the other hand, $\hat{k}_{qvs}$ is less than m_1 with high frequency when $n = 200$. When $n = 300$, $\hat{k}_{qvs}$ is mostly less than m_1 with relatively large s but greater than m_1 with smaller s. We suspect that in the latter case, condition (2.5) in Theorem 2 may not hold.

Table 1: Mean values of the QVS cut-off positions ($\hat{k}_{qvs}$), the positions of m_0, and the positions of m_1 from 1000 replications

n	s	$\hat{k}_{qvs}$	m_0	m_1	$F(\hat{k}_{qvs} \geq m_0)$	$F(\hat{k}_{qvs} \leq m_1)$
200	30	79(6.2)	4(2.8)	142(32.9)	1.00	0.99
	40	92(3.4)	3(2.4)	171(21.4)	1.00	1.00
	50	96(1.9)	3(2.2)	180(17.5)	1.00	1.00
300	30	68(8.1)	10(4.8)	58(17.7)	1.00	0.18
	40	99(9.9)	8(4.5)	120(41.2)	1.00	0.68
	50	127(7.7)	6(3.9)	206(49.8)	1.00	0.98

Standard errors are in parenthesis. $F(\hat{k}_{qvs} \geq m_0)$ and $F(\hat{k}_{qvs} \leq m_1)$ represent the frequencies of $\hat{k}_{qvs} \geq m_0$ and $\hat{k}_{qvs} \leq m_1$, respectively. In these examples, $p = 2000$, $Cov(X) = I$, and non-zero coefficients are equal to 0.5

Recall that condition (2.3) is imposed in Theorem 1 for QVS to retain relevant variables before m_0, and condition (2.5) is imposed in Theorem 2 to avoid over-selecting noise variables after m_1. Consider the settings in Table 1. Figure 2 shows the empirical distributions of q_{m_0+1} and $U_{(1),m-s}$,

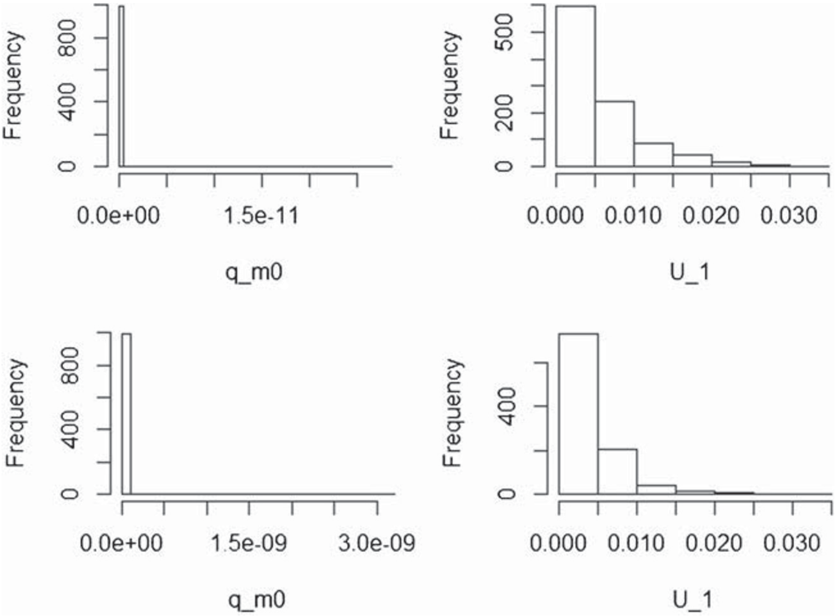

Figure 2: Histograms of q_{m_0+1} and $U_{(1),m-s}$. The top row has $n = 200$, $s = 30$, $p = 2000$, $Cov(X) = I$, and non-zero coefficients equal to 0.5. The bottom row increases the sample size to $n = 300$

respectively, from 1000 replications. The top row has $n = 200$ and $s = 30$, and the bottom row increases the sample size to $n = 300$. The results seem to support condition (2.3) that $q_{m_0+1} \leq U_{(1),m-s}$ with high probability. Condition (2.5) is more difficult to check in simulation because it is supposed to hold for arbitrary $t \in (0, 1)$ with high probability. We would defer the verification of this condition to future research.

3.2. Comparisons with other methods We compare the performance of QVS with other variable selection methods, such as Lasso with BIC ("BIC"), Lasso with 10-fold cross-validation ("LCV"), Bonferroni procedure applied to the Q statistics ("Q-BON"), and Benjamini-Hochberg procedure applied to the Q statistics ("Q-FDR"). Specifically, Q-BON and Q-FDR select the top-ranked variables on the solution path with sizes equal to $\arg\max_k \{k : \; q_k \leq 0.05/m\}$ and $\arg\max_k \{k : \; q_k \leq 0.05k/m\}$, respectively. The nominal levels for both Q-BON and Q-FDR are set at 0.05. We note that Q-FDR is equivalent to the TailStop method introduced in G'sell et al. (2016).

We demonstrate the performances of these methods by presenting the true positive proportion (TPP), false discovery proportion (FDP), and g-measure of these methods. TPP is the ratio of true positives to the number of relevant variables entered the solution path. FDP is the ratio of false positives to the number of selected variables. TPP and FDP measure the power and type I error of a selection method, respectively. We also compute the g-measure, which is the geometric mean of specificity and sensitivity, i.e. g-measure $= \sqrt{\text{specificity} \times \text{sensitivity}}$, where specificity is the ratio of true negatives to the number of noise variables in the solution path and sensitivity is equivalent to TPP. G-measure can be used to evaluate the overall performance of a variable selection method. Higher value of g-measure indicates better performance (Powers, 2011).

Figure 3 summarizes the mean values of TPP, FDP, and g-measure for different methods under various model settings with $p = 2000$, $n = 200$ and $Cov(\mathbf{X}) = \mathbf{\Sigma} = (0.5^{|i-j|})_{i=1,\cdots,p;\; j=1,\cdots,p}$. The number of non-zero coefficients $s = 10, 20, 40$, and the non-zero effect vary from 0.3 to 2. It can be seen that the Lasso-based BIC and LCV tend to select fewer variables, which results in lower TPP and FDP. On the other hand, the Q Statistic-based methods, Q-BON, Q-FDR, and QVS, all have higher TPP and FDP. However, in these examples, Q-BON does not control family-wise error at the nominal level of 0.05, and Q-FDR does not control FDR at the nominal of 0.05. The reason is because relevant and noise variables are not perfectly separated in these examples. As illustrated in Table 1, m_0 is much smaller than m_1, and the

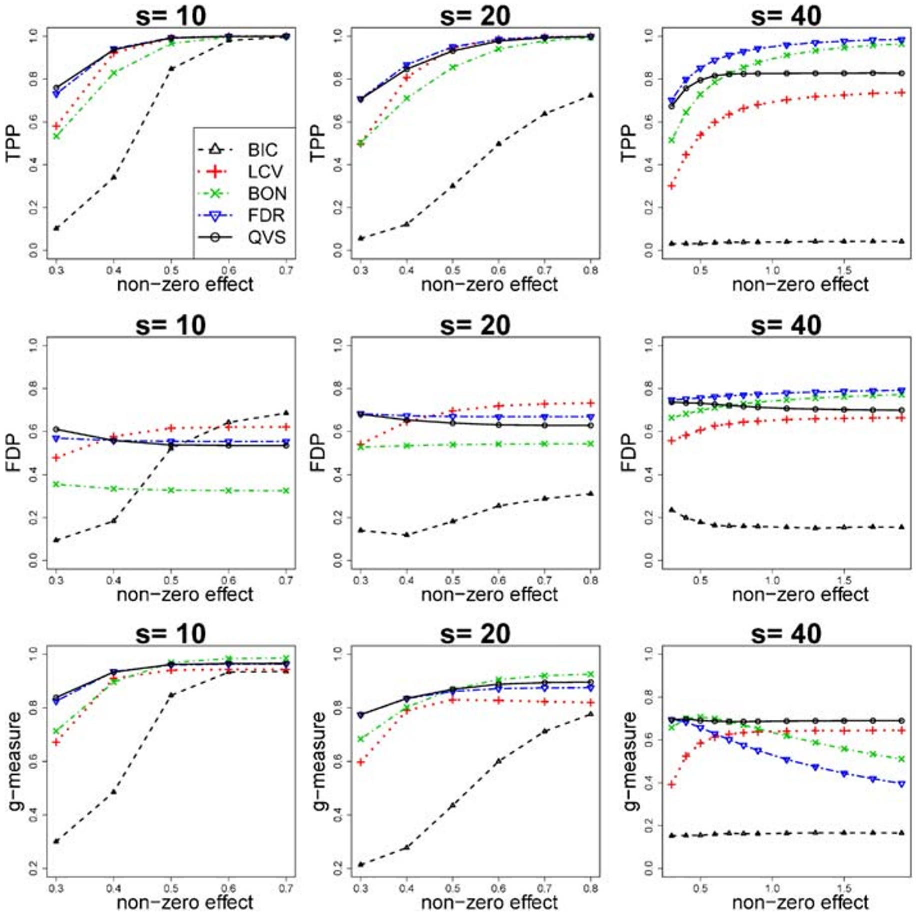

Figure 3: Comparisons of QVS with other methods when $p = 2000$, $\mathbf{\Sigma} = (0.5^{|i-j|})_{i=1,\cdots,p;\ j=1,\cdots,p}$, and $n = 200$

results of Q-BON and Q-FDR cannot be interpreted presumably. In terms of g-measure, QVS generally outperforms other methods. We suspect that the relatively better performance of QVS is related to its control of over-selecting noise variables and under-selecting relevant variables to certain degrees in the challenging scenarios when m_0 and m_1 are far apart.

4 Real Application

We obtain a dataset for expression quantitative trait loci (eQTL) analysis related to Down Syndrome. Down Syndrome is one of the most common gene-associated diseases. Our dataset includes the expression levels of gene

Table 2: Covariance test statistics and Q statistics along Lasso solution path for samples from Asian, Yoruba, and European populations, respectively

Knots	Asian		Yoruba		European	
	Covtest	Q statistic	Covtest	Q statistic	Covtest	Q statistic
1	6.09e-01	0.00	4.20e-01	0.00	9.81e-03	0.85
2	3.05e 00	0.01	6.46e 00	0.00	2.94e-02	0.85
3	1.98e-01	0.26	4.41e-02	0.45	1.64e-02	0.88
4	5.66e-02	0.32	1.23e-01	0.47	1.78e-03	0.89
5	1.24e-01	0.34	1.06e-04	0.54	1.05e-02	0.90
6	3.94e-03	0.39	2.77e-02	0.54	6.99e-03	0.91
7	2.05e-02	0.39	1.60e-03	0.55	1.06e-03	0.91
8	2.24e-02	0.40	4.31e-02	0.55	6.49e-03	0.91
9	1.09e-02	0.41	9.26e-03	0.58	1.03e-04	0.92
10	4.61e-02	0.41	1.29e-02	0.58	7.83e-03	0.92

CCT8, which contains a critical region of Down syndrome, and genome-wide single-nucleotide polymorphism (SNP) data from three different populations (Bradic et al., 2011): Asia (Japan and China) with sample size $n = 90$, Yoruba with $n = 60$, and Europe with $n = 60$. We perform eQTL mapping to identify SNPs that are potentially associated with the expression level of gene CCT8 for each population. Due to the limited sample size, we randomly select subsets of SNPs with $p = 6000, 2000, 4000$ for the three populations, respectively.

For the sample of each population, we first compute the covariance test statistics by Eq. 1.3 and Q statistics by (1.5) based on Lasso solution path. Table 2 presents these statistics for the top 10 knots on the path.

We apply QVS as well as all the other methods analyzed in Section 3.2 to the datasets. Table 3 presents the number of selected variables along the solution path for different methods. It can be seen that QVS generally selects more variables than the other methods for these datasets. Particularly, when there exists a gap in the list of Q-statistics, such as for Asian and Yoruba samples, QVS tends to stop right after the gap. This is because such gap is

Table 3: The number of selected variables long the Lasso solution path for different methods

Population	n	p	BIC	LCV	Q-BON	Q-FDR	QVS
Asian	90	6000	0	1	0	0	3
Yoruba	60	2000	1	2	2	2	3
European	60	4000	0	0	0	0	0

Table 4: Locations on the Lasso solution paths of the reference variables identified in Bradic et al. (2011)

Population	n	p	location of reference variables	m_0	m_1	QVS
Asian	90	1955	1, 2, 6, 46	2	46	3
Yoruba	60	1978	1, 4, 17, 34, 50, 58	1	58	3
European	60	2146	2, 4, 18, 30	0	30	0

likely to occur in the indistinguishable region between m_0 and m_1. Stopping right after the gap would include relevant variables ranked before m_0 and, at the same time, not over-select the noise variables ranked after m_1.

We further verify the result of QVS by comparing with the findings in literature. Bradic et al. (2011) studied the same samples for eQTL mapping but only focused on cis-eQTLs. Therefore, the numbers of SNPs included in their analysis are much smaller with $p = 1955, 1978, 2146$ for the three populations, respectively. More SNP variables are identified in Bradic et al. (2011) for each population due to larger ratio of sample size to dimension. Table 4 reports the locations on the solution path of the variables identified in Bradic et al. (2011). Note that the Lasso solution path is computed using our datasets with lower ratio of sample size to dimension. We utilize this result as a reference to evaluation the results of QVS.

For the solution path of Asian population, according to (Bradic et al., 2011), the first noise variable enters after two relevant variables and the last relevant variable enters at the position 46. Therefore, $m_0 = 2$ and $m_1 = 46$. QVS selects the top 3 variables on the path, which is in-between m_0 and m_1. This result supports the theoretical property of QVS as a sensible variable selection procedure. Similar results are observed in the other two cases.

5 Conclusion and Discussion

We develop a new variable selection procedure whose result is interpretable in the scenarios where relevant variables may be mixed indistinguishably with noise variables on the Lasso solution path. Our theoretical findings are very different from the existing results which consider variable selection properties in the ideal setting where all relevant variables rank ahead of noise variables on the solution path. The new analytic framework is unconventional but highly relevant to Big Data applications.

The proposed QVS procedure utilizes the Q statistic (G'sell et al., 2016) that is built upon the limiting distribution of the covariance test statistic developed in Lockhart et al. (2014) under orthogonal design. In a more general setting where design matrix is in general position as described in Lockhart et al. (2014), the theoretical analysis on covariance test statistic is much more complicated and its limiting distribution has not be fully derived. Lockhart et al. (2014) provides a control on the tail probability of the covariance test statistic. It will be interesting to characterize the indistinguishable region on the Lasso solution path and interpret the result of the proposed QVS method in the more general setting. We note that the simulation and real data analyses in the paper have design matrices that are not orthogonal. Compared with other popular methods, QVS shows advantages in selection accuracy and the ability to interpret its results.

Appendix

A Proof of Theorem 1

By the construction of $\hat{k}_{qvs}$ in Eq. 2.1,

$$\begin{aligned}\frac{\hat{k}_{qvs}}{m_0} - 1 &= \max_{1\le k\le m/2}\left\{\frac{k}{m_0} - 1 - \frac{m}{m_0}q_k - \frac{m}{m_0}c_m\sqrt{q_k(1-q_k)}\right\} \\ &\ge \frac{m_0+1}{m_0} - 1 - \frac{m}{m_0}q_{m_0+1} - \frac{m}{m_0}c_m\sqrt{q_{m_0+1}} \\ &> -\frac{m}{m_0}q_{m_0+1} - \frac{m}{m_0}c_m\sqrt{q_{m_0+1}}, \end{aligned} \tag{A.1}$$

where the second step above is by taking $k = m_0 + 1$.

By condition (2.3), $P(q_{m_0+1} \le U_{(1),m-s}) \to 1$, where $U_{(1),m-s} \sim Beta(1, m-s)$. Therefore, given $s = o(m)$ and $\sqrt{\log m}/m_0 \to 0$ in probability,

$$\frac{m}{m_0}q_{m_0+1} = O_p(\frac{1}{m_0}) = o_p(1). \tag{A.2}$$

On the other hand, it has been shown in Meinshausen and Rice (2006) that $c_m = O(\sqrt{\log m}/\sqrt{m})$. Then, by $s = o(m)$ and $\sqrt{\log m}/m_0 \to 0$ in probability, we have

$$\frac{m}{m_0}c_m\sqrt{q_{m_0+1}} = O_p(\frac{\sqrt{\log m}}{m_0}) = o_p(1). \tag{A.3}$$

Combining (A.1) - (A.3) gives (2.4).

B Proof of Theorem 2

Define $F_m(t) = \frac{1}{m}\sum_{k=1}^{m} 1(q_k \le t)$. Re-write $\hat{k}_{qvs}$ in Eq. 2.1 as

$$\hat{k}_{qvs} = m \max_{0<t<1}\{F_m(t) - t - c_m\sqrt{t(1-t)}\}. \tag{A.4}$$

For notation convenience, define $\pi_1 = m_1/m$. Then

$$F_m(t) \;=\; \frac{m_1}{m}\frac{1}{m_1}\sum_{j=1}^{m_1} 1(q_j \le t) + \frac{1}{m}\sum_{j=m_1+1}^{m} 1(q_j \le t) \le \pi_1 + \frac{1}{m}\sum_{j=m_1+1}^{m} 1(q_j \le t).$$

Therefore,

$$\begin{aligned}
& P(\hat{k}_{qvs} > m_1) \\
\le\;& P(\sup_{0<t<1}\{F_m(t) - t - c_m\sqrt{t(1-t)}\} > \pi_1) \\
\le\;& P(\sup_{0<t<1}\{\pi_1 + \frac{1}{m}\sum_{j=m_1+1}^{m} 1(q_j \le t) - t - c_m\sqrt{t(1-t)}\} > \pi_1) \\
=\;& P(\sup_{0<t<1}\{\frac{1}{m}\sum_{j=m_1+1}^{m} 1(q_j \le t) - t - c_m\sqrt{t(1-t)}\} > 0) \\
\le\;& P(\sup_{0<t<1}\{\frac{1}{m}\sum_{j=m_1-s+1}^{m-s} 1(U_{(j),m-s} \le t) - t - c_m\sqrt{t(1-t)}\} > 0) + o(1),
\end{aligned} \tag{A.5}$$

where the last step above is by condition (2.5). Note that $\sum_{j=m_1-s+1}^{m-s} 1$ $(U_{(j),m-s} \le t)$ is stochastically dominated by $\sum_{j=1}^{m-s} 1(U_{(j),m-s} \le t)$, which has the save distribution as that of $\sum_{j=1}^{m-s} 1(U'_j \le t)$, where $\{U'_j\}_{j=1}^{m-s}$ is a sequences of independent standard uniform random variables, independent of $\{U_j\}_{j=1}^{m}$. Further, $\sum_{j=1}^{m-s} 1(U'_j \le t)$ is stochastically dominated by $\sum_{j=1}^{m} 1(U'_j \le t)$, which has the same distribution as that of $mU_m(t)$. Sum up the above, we have

$$\begin{aligned}
& P(\sup_{0<t<1}\{\frac{1}{m}\sum_{j=m_1-s+1}^{m-s} 1(U_{(j),m-s} \le t) - t - c_m\sqrt{t(1-t)}\} > 0) \\
\le\;& P(\sup_{0<t<1}\{U_m(t) - t - c_m\sqrt{t(1-t)}\} > 0).
\end{aligned} \tag{A.6}$$

By the definition of the bounding sequence c_m,

$$P\left(\sup_{0<t<1}\frac{U_m(t)-t}{\sqrt{t(1-t)}} > c_m\right) = \alpha_m.$$

Further,

$$P\left(\sup_{0<t<1}\frac{U_m(t)-t-c_m\sqrt{t(1-t)}}{\sqrt{t(1-t)}} > 0\right) \le P\left(\sup_{0<t<1}\frac{U_m(t)-t}{\sqrt{t(1-t)}} > c_m\right) = \alpha_m.$$

And for every $t \in (0,1)$,

$$\frac{U_m(t) - t - c_m\sqrt{t(1-t)}}{\sqrt{t(1-t)}} > U_m(t) - t - c_m\sqrt{t(1-t)}$$

almost surely. The above implies that

$$P\left(\sup_{0<t<1}\{U_m(t) - t - c_m\sqrt{t(1-t)}\} > 0\right) \leq \alpha_m = o(1). \tag{A.7}$$

Combining (A.5) with (A.6) and (A.7) gives $P(\hat{k}_{qvs} > m_1) \to 0$ as $m \to \infty$.

References

BARBER, RF and CANDÈS, EJ (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43**, 5, 2055–2085.

BENJAMINI, Y and HOCHBERG, Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**, 1, 289–300.

BOGDAN, M, GHOSH, J and ŻAK-SZATKOWSKA, M (2008). Selecting explanatory variables with the modified version of the bayesian information criterion. *Quality and Reliability Engineering International* **24**, 6, 627–641.

BOGDAN, M, CHAKRABARTI, A, FROMMLET, F and GHOSH, J (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* **39**, 3, 1551–1579.

BOGDAN, M, VAN DEN BERG, E, SABATTI, C, SU, W and CANDÉS, E (2015). SLOPE - adaptive variable selection via convex optimization. *The Annals of Applied Statistics* **9**, 3, 1103–1140.

BRADIC, J, FAN, J and WANG, W (2011). Penalized composite quasi-likelihood for unltahigh-dimensional variable selection. *Journal of the Royal Statistical Society: Series B* **73**, 3, 325–349.

CHAKRABARTI, A and GHOSH, J (2007). Some aspects of bayesian model selection for prediction. *Bayesian Statistics* **8**, 51–90.

CHAKRABARTI, A and GHOSH, J (2011). Aic, bic, and recent advances in model selection. *Handbook of the Philosophy of Science* **7**, 583–605.

EFRON, B, HASTIE, T, JOHNSTONE, I and TIBSHIRANI, R (2004). Least angle regression. *The Annals of Statistics* **32**, 2, 407–499.

G'SELL, M, WAGER, S, CHOULDECHOVA, A and TIBSHIRANI, R (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B* **78**, 2, 423–444.

JENG, XJ and CHEN, X (2019a). Predictor ranking and false discovery proportion control in high-dimensional regression. *Journal of Multivariate Analysis* **171**, 163–175.

JENG, XJ and CHEN, X (2019). Variable selection via adaptive false negative control in linear regression. *Electronic Journal of Statistics* **13**, 2, 5306–5333.

JENG, XJ, ZHANG, T and TZENG, J-Y (2019). Efficient signal inclusion with genomic applications. *Journal of the American Statistical Association* **114**, 1787–1799.

LEE, J, SUN, D, SUN, Y and TAYLOR, J (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44**, 3, 907–927.

LOCKHART, R, TAYLOR, J, TIBSHIRANI, R and TIBSHIRANI, R (2014). A significance test for the lasso. *The Annals of Statistics* **42**, 2, 413–468.

MEINSHAUSEN, N and BUHLMANN, P (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika* **92**, 4, 893–907.

MEINSHAUSEN, N and RICE, J (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics* **34**, 1, 373–393.

POWERS, D (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies* **2**, 37–63.

SU, W, BOGDAN, M and CANDES, E (2017). False discoveries occur early on the lasso path. *The Annals of Statistics* **45**, 5, 2133–2150.

TIBSHIRANI, R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

VAN DE GEER, S, BUHLMANN, P, RITOV, Y and DEZEURE, R (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 3, 1166–1202.

WAINWRIGHT, M (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* **55**, 5, 2183–2202.

WILBUR, JD, GHOSH, J, NAKATSU, C, BROUDER, S and DOERGE, R (2002). Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community dna fingerprints. *Biometrics* **58**, 2, 378–386.

ZHANG, C and ZHANG, SS (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* **76**, 1, 217–242.

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

X. Jessie Jeng, Huimin Peng and
Wenbin Lu
Department of Statistics, North
Carolina State University, Raleigh,
NC, USA
E-mail: xjjeng@ncsu.edu