THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #241

Tournament Screening cum EBIC for Feature Selection with High Dimensional Feature Spaces

Zehua Chen and Jiahua Chen

Tournament Screening cum EBIC for Feature Selection with High Dimensional Feature Spaces

Zehua Chen[1] and Jiahua Chen[2]

[1] *National University of Singapore*

[2] *University of British Columbia*

Correspondence Author:

Zehua Chen

Department of Statistics & Applied Probability

National University of Singapore

3 Science Drive 2

Singapore 117543

Email: stachenz@nus.edu.sg

Abstract

The feature selection characterized by relatively small sample size and extremely high dimensional feature space is common in many areas of contemporary statistics. The high dimensionality of the feature space causes serious difficulties: (i) the sample correlations between features become high even if the features are stochastically independent; (ii) the computation becomes intractable. These difficulties make conventional approaches either inapplicable or inefficient. The reduction of dimensionality of the feature space followed by low dimensional approaches appears the only feasible way to tackle the problem. Along this line, we develop in this article a tournament screening cum EBIC approach for feature selection with high dimensional feature space. The procedure of tournament screening mimics that of a tournament. It is shown theoretically that the tournament screening has the sure screening property, a necessary property which should be satisfied by any valid screening procedure. It is demonstrated by numerical studies that the tournament screening cum EBIC approach enjoys desirable properties such as having higher positive selection rate and lower false discovery rate than other approaches.

Keywords: Extended Bayes information criterion, Feature selection, Penalized likelihood, Permutation aggregating, Reduction of dimensionality, Small-n-large-P, Sure screening.

1 Introduction

It becomes a common situation in many areas of contemporary statistics that a few features have to be identified from a huge feature space to explain the variation of a response variable with relatively a small number of observations. It gives rise to the so-called sparse small-n-large-P problem; that is, the number of candidate features (P) is much larger than the sample size (n), but the number of the features causal to the response variable is small. For example, in microarray tumor studies, DNA expression levels of thousands of genes are measured to identify, say, a few tens of genes which can be used for the classification or for the prognostics of tumors. Usually, the sample size is not more than a hundred. In genome-wide genetic association studies, to detect a handful of etiological variants of certain quantitative traits or common diseases, tens or hundreds of thousands of single nucleotide polymorphisms (SNPs) are genotyped for relatively a small number of samples. For some recent discussions, see, e.g., Hunter and Li (2005), Huang et al. (2007), Paul et al. (2007), Zhang and Huang (2007), Kosorok and Ma (2007), and Fan and Lv (2007).

The sparse small-n-large-P problem poses great challenges to feature selection. First, the causal features are buried among an extremely huge number of candidate features. Second, even if all the features are independent, the maximum sample correlation between the features (or maximum spurious correlation coined by some authors) can reach a very high level. Because of the high spurious correlation, non-causal features might appear highly correlated with the response variable, which makes the causal features hard to detect.

In genetic studies, some simple methods have been employed. For example, appropriate models are fitted to features (markers, SNPs, etc.) one at a time. Features with highest significant effects are selected with Bonferroni adjusted threshold value to control the family-wise type I error rate, or, an estimated false discovery rate (FDR) is used to determine significant single features, see Tusher et al. (2001), Tibshirani et al. (2002), Marchini et al. (2005), Benjamini and Hochberg (1995), and Storey and Tibshirani (2003). Another strategy is to pool together the strength of single feature statistics to increase the power for the detection of causal features. This includes the sum-statistics method developed by Hoh, Wille and Ott (2001), see also Hoh and Ott (2003), and the method using truncated product of p-values, see Zaykin et al. (2002) and Dudbridge and Koeleman (2003). These methods ignore multi-feature joint effects. They are particularly problematic with high spurious correlations when P is large.

A different, probably more appropriate, strategy is to incorporate feature selection into model selection. Apart from classical model selection methods such as all-subsets and stepwise forward or backward methods, more advanced approaches have been developed in the recent past. Tibshirani (1996) proposed the LASSO, a L_1 norm penalized likelihood approach. Fan and Li (2001) advocated the SCAD, a penalized likelihood approach with a modified penalty. Zou and Hastie (2005) proposed the Elastic Net, which adopts a combined L_1 and L_2 penalty. Efron et al. (2004) proposed the LARS, a sequential variable selection procedure which includes LASSO as a special case. The penalized likelihood methodology has been further advanced by more recent developments, see Hunter and Li (2005), Huang et al. (2007), Paul et al. (2007), and Zhang and Huang (2007). Bayesian approach with MCMC for model selection has also been developed, see Ishwaran and Rao (2003).

However, when P is extremely huge, as is common in genetic genome-wide studies, there are computational hurdles for the direct implementation of the above mentioned methods. The reduction of dimensionality of the feature space then becomes a natural way to step forward. Because of the sparsity of the causal features, we can hope that by an efficient screening procedure most of the non-causal features can be first screened out before any model selection method is used. Such a screening procedure must have the ability to retain the causal features while screening out non-causal features. This property is referred to as the sure screening property by Fan and Lv (2007).

Fan and Lv (2007) considered a sure independent screening (SIS) procedure which satisfies the sure screening property. The SIS is similar to the single-feature-statistics method mentioned in an earlier paragraph. But it is only used for pre-screening, not for the final feature selection. It reduces the dimensionality of the feature space (P) to a level p around the sample size (n) by retaining the first p relatively most significant features. Before it is formalized by Fan and Lv (2007), the SIS has in fact been used in an ad hoc way in genetic studies, see, e.g., the analysis of a leukemia data in Zou and Hastie (2005).

In this article, we develop a procedure called tournament screening (TS). The basic idea of the tournament screening is as follows. The features are sequentially screened in consecutive stages. At each stage, the features retained from the previous stage are divided into non-overlapping groups. The features in each group are subjected to a penalized likelihood mechanism and a given number of them are selected. The features selected are then pooled together and enter the next stage. This process is continued until the dimensionality of the feature space is reduced to a desirable level. The name of TS is coined because of its similarity to the competitions in a tournament. We show in this article that the sure screening property is also satisfied by TS. We also demonstrate by simulations that TS dominates SIS in the final feature selection in a sense to be made clear later.

Although feature selection can be incorporated into model selection, there is still a nuance between the two. Model selection involves not only features but also the estimation of the parameters associated with the features. It focuses more on the prediction accuracy of the models, even if the model contains spurious features in addition to the causal ones. On the other hand, the emphasis of feature selection is different where the aim is to detect the causal features. It is concerned more with the capacity of detecting causal features and the purity of the selection. The capacity of detecting causal features is measured by positive selection rate (PSR) which is the ratio of the number of selected causal features and the total number of causal features. The purity of the selection is measured by false discovery rate (FDR) which is the ratio of the number of non-causal features selected and the total number of selected features. A method dominates another if it achieves higher PSR and lower FDR than the other. In this article, we also propose a procedure for the final feature selection which combines the penalized likelihood methodology with an extended Bayes information criterion (EBIC) developed by Chen and Chen (2007). The TS followed by this procedure for feature selection is referred to as the TS cum EBIC approach.

The remainder of the article is arranged as follows. The TS and its sure screening property are described in §2. The procedure for final feature selection is presented in §3. Numerical studies are reported in §4. Some conclusion remarks are made in §5. Technical details are given in the Appendix.

2 TS and its sure screening property

Let $\boldsymbol{y}$ be an $n \times 1$ vector of response values and X an $n \times P$ matrix of feature values. Assume that, given X,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta}$ is a $P \times 1$ vector of parameters with sparsity property, i.e., only a small number of its components are non-zero, and $\boldsymbol{\epsilon}$ is a vector of n independent and identically distributed random variables with mean zero and variance σ^2. More generally, we can consider the model

$$\boldsymbol{y} \sim f(\boldsymbol{y}, X\boldsymbol{\beta}, \sigma^2), \tag{2}$$

where f is the density function of $\boldsymbol{y}$ given X. The general model includes generalized linear models.

Denote by $\mathcal{S}^1$ the set of integers from 1 to P. Let $\mathcal{S}_0$ be the subset of $\mathcal{S}^1$ corresponding to the non-zero components of $\boldsymbol{\beta}$; that is, $\beta_j \neq 0$ if and only if $j \in \mathcal{S}_0$. For any subset $\mathcal{S}$ of $\mathcal{S}^1$, let $\nu(\mathcal{S})$ denote the cardinality of $\mathcal{S}$. In particular, let $\nu_0 = \nu(\mathcal{S}_0)$. Denote by $X(\mathcal{S})$ the sub-matrix consisting of the columns of X with column indices in $\mathcal{S}$, and by $\boldsymbol{\beta}(\mathcal{S})$ the corresponding components of $\boldsymbol{\beta}$. The penalized negative log-likelihood function is defined as:

$$l_p(\boldsymbol{\beta}(\mathcal{S}), \sigma^2|\lambda) = -2\ln f(\boldsymbol{y}, X(\mathcal{S})\boldsymbol{\beta}(\mathcal{S}), \sigma^2) + n\sum_{j\in\mathcal{S}} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function regulated by a tuning parameter λ. The penalty function can be taken as $p_\lambda(|\beta|) = \lambda|\beta|$, the L_1 penalty used in LASSO by Tibshirani (1996), or as the penalty used in SCAD by Fan and Li (2001) defined through the following derivative:

$$p'_\lambda(|\beta|) = \lambda\left\{I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\right\}$$

for some choice of $a > 2$. The crucial property required of the penalty function is that it must be singular at zero. Because of the singularity, when the penalized negative likelihood is minimized with λ fixed at a certain value, a number of components of fitted $\boldsymbol{\beta}(\mathcal{S})$ will be forced to zero. By tuning the value of λ, practically any number of components of the fitted $\boldsymbol{\beta}(\mathcal{S})$ can be forced to zero.

We now describe the TS procedure in detail. Let n_g be a pre-specified group size. It is determined such that $n_g < n$ and that the numerical minimization of the penalized negative likelihood can be effectively carried out. Let K be a pre-determined number which serves as the selection size for each group in TS. In principle, K should be large enough in order to be able to retain all the causal features in the group and small enough to reduce the dimensionality efficiently. If one has a rough idea about how big ν_0 is, K can be chosen as $2\nu_0$ or $3\nu_0$. The TS procedure goes as follows.

Stage 1: Partition $\mathcal{S}^1$ at random into groups of nearly equal size n_g to yield

$$\mathcal{S}^1 = \mathcal{S}_{11} \cup \cdots \cup \mathcal{S}_{1J_1},$$

where J_1 is the largest integer such that $[n_g J_1] \leq P$. For each group $\mathcal{S}_{1j}$, minimize $l_p(\boldsymbol{\beta}(\mathcal{S}_{1j}), \sigma^2|\lambda)$ by tuning λ so that there are only K non-zero components of fitted $\boldsymbol{\beta}(\mathcal{S}_{1j})$. The features corresponding to these K components are selected from this group. Let $\mathcal{S}_{1j}^*$ denote the index set of these K features. At the end, pool $\mathcal{S}_{1j}^*, j = 1, \ldots, J_1$, together to form $\mathcal{S}^2$. The dimensionality of the feature space is reduced to KJ_1 at this stage.

Stage 2: Repeat the process in Stage 1 with $\mathcal{S}^1$ replaced by $\mathcal{S}^2$. The dimensionality of the feature space is further reduced to KJ_2 where J_2 is the largest integer such that $[n_g J_2] \leq KJ_1$.

Further Stages: The above process continues until the dimensionality of the feature space is reduced to K.

The TS procedure described above starts with a random partition of the feature space. The eventually reduced feature space is dependent of the initial partition in general. To reduce this dependence, a permutation aggregating procedure can be applied. The permutation aggregating procedure is as follows. The TS procedure is repeated a given number, say B, of times, each time with an independent partition of the feature space at the beginning and with a larger reduced dimensionality, say $2K$. At the end, count the frequencies of the features appearing in the obtained B reduced feature spaces, and retain the K most frequent features for the final feature selection. The idea of permutation aggregating is the same as the bootstrap aggregating (Bagging) proposed by Breiman (1996). The only difference is that bootstrapping in bagging is here replaced by random permutation.

The sure screening property of TS is stated in the following theorem under some conditions.

Theorem: *Assume model (1) holds. In addition, assume that the entries of X are independent bounded random variables with mean 0 and variance 1, and they are independent of $\boldsymbol{\epsilon}$. Suppose that $P = O(n^{\kappa})$ for some $\kappa > 0$ as $n \to \infty$. Let the penalty function p_λ be taken as either L_1 penalty or SCAD penalty. Let $\mathcal{S}$ be any subset of $\mathcal{S}^1$ such that $K < \nu(\mathcal{S}) < n$ and λ^* be the smallest value of λ such that the minimum of*

$$l_p(\boldsymbol{\beta}(\mathcal{S})|\lambda^*) = \|\boldsymbol{y} - X(\mathcal{S})\boldsymbol{\beta}(\mathcal{S})\|^2 + n\sum_{j\in\mathcal{S}} p_{\lambda^*}(|\beta_j|)$$

attains at some $\hat{\boldsymbol{\beta}}(\mathcal{S})$ with K non-zero components. Let $\mathcal{S}^*$ be the index set of the nonzero components of $\hat{\boldsymbol{\beta}}(\mathcal{S})$. Then, as $n \to \infty$,

$$P(\mathcal{S}_0 \cap \mathcal{S} \subset \mathcal{S}^*) \to 1.$$

Our condition on the entries of X is motivated by Fan and Lv (2007). They assumed the same condition for the sure screening property of SIS. This condition as well as the assumption of regression model (1) can be relaxed. We do not intend the relaxation in this article because the regression model still covers a broad class of practical problems. Furthermore, the relaxation of the conditions will involve too much technicalities which we want to avoid at the current stage. The proof of the theorem is given in the Appendix.

3 TS cum EBIC for final feature selection

The TS reduces the dimensionality $P(>> n)$ of the original feature space to $K(< \ n\ )$. Then any conventional model selection methods can be employed in the final selection. We will concentrate on the penalized likelihood methodology in the final selection because of its superiority over classical model selection methods which has been shown in the literature. The penalized likelihood methodology is usually coupled with cross validation, and feature selection and parameter estimation are done simultaneously. Implicitly, it aims to select a model (including both features and the estimates of the parameters) to minimize prediction error. However, we found that the penalized likelihood methodology coupled with cross validation tends to select too many spurious features overwhelming the causal features. For feature selection, the accountability of individual features is more important than the overall predicting ability of the model. A procedure which emphasizes on the accountability of individual features is more appropriate. In this section, we describe a procedure that combines the penalized likelihood methodology with EBIC. The penalized likelihood methodology is used to order models and the EBIC is then used to make the final selection.

Let $\mathcal{S}^*$ be the index set of the K features retained from the TS procedure. First, consider the penalized likelihood of the model containing all the K features:

$$l_p(\boldsymbol{\beta}(\mathcal{S}^*)|\lambda) = -2\ln f(\boldsymbol{y}, X(\mathcal{S}^*)\boldsymbol{\beta}(\mathcal{S}^*)) + n\sum_{j\in\mathcal{S}^*} p_\lambda(|\beta_j|).$$

The parameter λ is tuned to a value λ_1 such that it is the smallest to make at least one component of $\hat{\boldsymbol{\beta}}(\mathcal{S}^*)$ obtained by minimizing $l_p(\boldsymbol{\beta}(\mathcal{S}^*)|\lambda_1)$ to be zero. There might be more than one zero components. But, for the ease of presentation, we assume there is only one such component. Let $j_K \in \mathcal{S}^*$ be the index corresponding to this zero component. Next, update $\mathcal{S}^*$ to $\mathcal{S}^*/j_K$. This is equivalent to eliminating feature $X(\{j_K\})$ from further consideration. With the updated $\mathcal{S}^*$, the above minimization and tuning procedure is repeated, and another feature, say $X(\{j_{K-1}\})$, is eliminated. Continuing this way, eventually, we obtain an ordered sequence of the indices in $\mathcal{S}^*$: $j_1, j_2, \ldots, j_K$.

From the ordered sequence above, we form a nested sequence of subsets of indices: $\mathcal{S}_k = \{j_1, \ldots, j_k\}, k = 1, \ldots, K$. For each $\mathcal{S}_k$, the un-penalized likelihood $\ln f(\boldsymbol{y}, X(\mathcal{S}_k)\boldsymbol{\beta}(\mathcal{S}_k))$ is maximized and then the EBIC is computed. The features in the model with the smallest EBIC value are eventually selected.

In the remainder of this section, we give some discussion on the EBIC. For model selection criteria, the traditional AIC (Akaike, 1973), BIC (Schwarz, 1978) and the like are too liberal in the sense that too many spurious features will be selected when the dimension of the feature space is extremely high. The final selection with the reduced feature space is not a feature selection problem with low dimensional feature space. The high spurious correlation of the original high dimensional feature space will pass to the reduced feature space. The high dimensionality should still be taken into account in the final selection. For example, a model based on $X(\mathcal{S}_k)$ can well be considered as a model selected among all models containing k features from the original feature space. The EBIC is recently developed by Chen and Chen (2007) particularly for feature selection with high dimensional feature spaces. The EBIC of a model with feature matrix $X(\mathcal{S})$ is defined as

$$\mathrm{EBIC}_{\gamma}(\mathcal{S}) = -2 \ln f(\boldsymbol{y}, X(\mathcal{S})\hat{\boldsymbol{\beta}}(\mathcal{S})) + \nu(\mathcal{S}) \ln n + 2\gamma \ln[\tau(\mathcal{S})], \tag{3}$$

where $\hat{\boldsymbol{\beta}}(\mathcal{S})$ is the maximum likelihood estimate of $\boldsymbol{\beta}(\mathcal{S})$, $\tau(\mathcal{S})$ is the total number of models which can be formed by $\nu(\mathcal{S})$ features from the original feature space of dimension P, and γ is a constant between 0 and 1 which is to be determined by the user.

The EBIC of a model is essentially the negative of 2 times the log posterior probability of the model in a Bayesian framework. The original BIC is a special case of EBIC with $\gamma = 0$ which corresponds to a constant prior that assigns equal probability mass on every individual model. When the dimension P of the feature space is huge, the absurdity of the BIC becomes obvious. Let the model space be partitioned into subclasses according to the number of features a model contains. Under the constant prior, the probability assigned to a subclass is proportional to its size, and the subclasses of models with larger number of features have much larger probability than subclasses of models with small number of features. As a consequence, the BIC is in favour of models with more features rather than those with less features. The EBIC rectifies BIC by assigning the prior probability to a subclass inversely proportional to the γth power of its size. Thus EBIC is in favour of models with fewer features. Under some mild conditions, Chen and Chen (2007) shown that if $P = O(n^{\kappa})$ and $\gamma > 1 - 1/(2\kappa)$ then EBIC is consistent in the sense that asymptotically the EBIC will select the model with feature matrix $X(\mathcal{S}_0)$ with probability 1. The result also indicates that the original BIC might not be consistent when $\kappa \geq 1/2$. The EBIC with $\gamma = 1$ is universally consistent but it is also the most stringent criterion.

4 Numerical Studies

In this section, we report three sets of numerical studies. In the first set, we compare TS with SIS on their consequence for the final feature selection. In the second set, the TS cum EBIC approach with different penalties are compared with each other, and they are also compared with other feature selection procedures. In the third set, the TS cum EBIC approach is applied to a real problem.

Numerical Study 1

In the comparison of TS and SIS, we are not concerned with the sure screening property of the two procedures since the property is theoretically justified for both TS and SIS and some simulation results on SIS have been reported by Fan and Lv (2007). Our goal is the final feature selection, therefore, we focus on their consequences for the final selection. Fan and Lv (2007) considered the combination of SIS with a number of model selection methods such as SCAD, Dantzig selector (Candes and Tao, 2007) and so on. While focusing on the prediction accuracy of the selected model, they made comparison among those different combinations and found that the combination of SIS with SCAD outperforms all the other combinations in terms of both parsimony and estimation error. Therefore, in our study, we only compare TS followed by SCAD (TS-SCAD) with SIS followed by SCAD (SIS-SCAD). But, instead of using cross-validation to choose the tuning parameters, which aims at minimizing prediction error but fails in selecting causal features, we apply the procedure described in §3 and choose the tuning parameters by EBIC$_1$. We refer these procedures as TS cum EBIC and SIS cum EBIC. The failure of cross validation for causal feature selection will be exposed in Numerical Study 2. For each simulated data set, the original feature space is reduced to the same low dimensionality by both TS and SIS. The positive selection rate (PSR) and the false discovery rate (FDR) in the final selection are then compared. Two settings for generating the data sets are utilized.

In the first setting, the sample size is taken as $n = 200$ and the dimension of the feature space is taken as $P = 1000$. The features are generated in groups of 100 each. For each group, the features are generated as the components of a one-hundred-variate normal vector with mean zero, variance 1 and correlation $\text{corr}(i, j) = |\rho|^{|i-j|}$ for given ρ's. The generated features are shuffled and 15 of them are randomly chosen as the causal features. The response variable is then generated as

$$y_i = \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + \epsilon_i,$$

where $\beta_0 = 0, \beta_j = 3, j = 1, \ldots, 15$, and ϵ_i are i.i.d. normal variables with mean zero and standard deviation 9.5. The standard deviation is chosen such that each individual feature has a 4% marginal contribution to the variation of the response variable. The dimensionality of the feature space is reduced to 30, i.e., two times the number of causal features. The PSR and FDR of the two procedures are averaged over 100 replicates of simulations.

The second setting is adapted from Fan and Lv (2007). The sample size is taken as $n = 400$, the dimensionality of the feature space is taken as $P = 2500$. The causal features are taken as $X_1, \ldots, X_{14}$ which are generated as dependent normal variables with mean zero, variance 1 and correlation $\text{corr}(i, j) = 0.4^{|i-j|}$. The remaining $P - 14$ features are generated as follows. Let $Z_{15}, \ldots, Z_P$ be generated as i.i.d. standard normal variables. Then, for $i = 15, \ldots, 28$, X_i is generated as $X_i = Z_i + rX_{i-14}$, for the remaining i, $X_i = Z_i + (1 - r)X_1$, where $r = 1 - 5 \ln n/P$. The response variable

Table 1: The average positive selection rate (PSR) and false discovery rate (FDR) of the TS cum EBIC with SCAD (TS-SCAD) and SIS cum EBIC with SCAD (SIS-SCAD) over 100 replicates of simulations as described in Simulation 1.

		PSR		FDR	
(n, P)	ρ	TS-SCAD	SIS-SCAD	TS-SCAD	SIS-SCAD
(200, 1000)	0.00	0.130	0.136	0.035	0.033
	0.25	0.122	0.120	0.036	0.037
	0.75	0.151	0.137	0.217	0.218
(400, 2500)	0.40	0.917	0.862	0.014	0.011

is then generated as

$$y_i = \sum_{j=1}^{14} \beta_j X_{ij} + \epsilon_i,$$

where ϵ_i are i.i.d. normal variables with mean 0 and standard deviation 2 and β_j's are generated as $\beta_j = (-1)^{u_j}(a + |z_j|)$ with z_j being i.i.d. standard normal variables, u_j being i.i.d. Bernoulli variables with probability of success 0.4, and $a = 4 \ln n/\sqrt{n}$. Again, the simulation is repeated for 100 replicates.

The results under both settings are given in Table 1. The results demonstrate that when the features are independent or slightly correlated, TS-SCAD and SIS-SCAD are comparable in terms of both PSR and FDR (the cases with $\rho = 0$ and 0.25 in the first setting), but when the correlation among the features are high, TS-SCAD has much higher PSR and lower or comparable FDR (the case with $\rho = 0.75$ in the first setting and the case in the second setting). The PSR in the first setting is not high because the effect of each causal feature is not strong (each accounts for only 4% of the total variation) and the sample size is relatively small. With stronger effects and larger sample size, both procedures have high PSR and low desirable FDR, as demonstrated in setting 2.

Numerical Study 2

In this set of numerical studies, we compare the TS cum EBIC coupled with different penalties and other approaches such as LASSO with cross-validation and multiple tests with Bonferroni adjusted threshold values. In the TS cum EBIC approach, there is an issue on the choice of penalty function. Although, in principle, the LASSO penalty, the SCAD penalty and the elastic net penalty which is a linear combination of L_1 and L_2 penalties can all be used, their performance in feature selection could differ. Since elastic net penalty involves two tuning parameters, it does not produce a linear order of the features when the parameters are tuned, which entails a tremendous computational complexity. Hence, we only consider the LASSO and SCAD penalty. The LASSO with cross-validation is chosen as a representative for approaches focusing on model prediction accuracy. It is chosen because of its easy computation. The multiple tests with Bonferroni adjusted threshold value mentioned in the introduction, which has been used in the genetics literature, is one of the univariate soft thresholding approaches.

The elegant LARS algorithm developed by Efron et al. (2004) makes it possible to compute the whole solution path for LASSO sequentially when $P < n$. When $P \geq n$, the algorithm cannot continue after the first n features have been tentatively selected. In addition, the computational load is already very heavy even long before n features are selected when P is very large. With moderate dimensionality, we found that the TS procedure produces essentially the same reduced feature space as that obtained by the LARS algorithm stopped while the same number of features have been selected.

The features and responses are generated in the same way as in Setting 1 of Numerical Study 1 except that 20 simulated features are randomly chosen as causal

Table 2: The average positive selection rate (PSR) and false discovery rate (FDR) of the TS cum EBIC with SCAD (TS-S), TS cum EBIC with LASSO (TS-L), LASSO with cross-validation (L-CV) and univariate soft thresholding (UST) over 100 replicates of simulations as described in Simulation 2.

		PSR				FDR			
(n, P)	$\rho(\sigma)$	TS-S	TS-L	L-CV	UST	TS-S	TS-L	L-CV	UST
(200, 1000)	.20(6)	.528	.444	.997	.165	.151	.197	.802	.039
	.40(6)	.487	.385	.996	.173	.152	.209	.795	.065
	.75(6)	.533	.263	.988	.230	.238	.284	.793	.489
(400, 2500)	.40(6)	.993	.999	1.00	.509	.075	.099	.826	.046
(800, 6100)	.40(6)	1.00	1.00	—	.936	.004	.052	—	.056
(800, 6100)	.40(15)	.719	.771	—	.467	.035	.072	—	.027

features and that the standard deviation of ϵ_i is set at 6. The reduced dimensionality is set as 40, again, two times the number of causal features. But for the procedure of LASSO with cross validation, the dimensionality is reduced to 120, since by some preliminary simulation we found that the cross validation will choose almost all the features in the reduced feature space if the dimensionality is reduced to too low. The following settings are considered: (i) $(n, P) = (200, 1000)$, $\rho = 0.2, 0.4$ and 0.75, $\sigma = 6$; (ii) $\rho = 0.4$, (n,P) = (400, 2500), (800, 6100), $\sigma = 6$; (iii) $\rho = 0.4$, (n,P) = (800, 6100), $\sigma = 15$. The simulation results are reported in Table 2. In the heading of the table, TS-S, TS-L, L-CV and UST stand respectively for TS cum EBIC with SCAD, TS cum EBIC with LASSO, LASSO with cross-validation and univariate soft thresholding. For the UST, an overall 0.05 level critical value is used.

The findings of Simulation 2 are summarized as follows. (a) Generally, TS-S has higher or comparable PSR and lower FDR than TS-L and hence is obviously better than TS-L. (b) The performance of TS-S is not affected too much by the increment of the correlation among the features while the increment of correlation has an apparent adverse effect on the performance of TS-L. (c) The LASSO with cross-validation has intolerably high FDR and fails for feature selection. (d) The UST can tightly control the FDR only when features are not highly correlated, its power to detect causal features is too low. (e) There are two major factors which affect PSR and FDR: sample size and relative size of feature effects. The PSR increases and the FDR decreases as sample size increases. The larger the relative size of the feature effects, the larger the PSR and the smaller the FDR, as demonstrated by the results under setting (ii) and (iii).

Numerical Study 3

In this study, we first apply the TS cum EBIC approach to a real genetic data set, then use the data set to conduct some simulations, which provides guidelines on how the results of the analysis can be used for further genetic studies. The SCAD penalty is used in the approach. The data consists of the trait values together with the genotypes at 2155 SNPs over 23 chromosomes of 233 individuals belonging to 16 pedigrees obtained from an experiment. In the experiment, B lymphocytes from the blood samples of these individuals are transformed into immortalized lymphoblastoid cell lines (LCLs) by Epstein-Barr Virus (EBV) which express in the LCLs. The trait value is a measure of the mRNA expression level of the EBNA-3A gene, one of the EBV genes, in LCLs. As the result of a preliminary analysis, only 1414 SNPs are retained in the analysis because the other SNPs either are uninformative or have a large proportion of missing genotypes.

In this application, because of the pedigree structure, a variance-component-model considered by Amos (1994) is used instead of the linear regression model. The TS procedure is applied by first randomly dividing the 1414 SNPs into 14 groups of equal size 101, and selecting 20 SNPs from each group. The dimensionality of the SNPs is finally reduced to 30. Then the final selection procedure with EBIC is applied to the 30 SNPs. The EBIC_1, $\text{EBIC}_{1/2}$ and the original BIC are used in the final selection. The EBIC_1 selects 1 SNP, the $\text{EBIC}_{1/2}$ selects 4 SNPs and the original BIC selects 7 SNPs. More details about the analysis are given in Chen, Chen and Liu (2006). To what extent, can the selected SNPs by the different EBIC be trusted as the genuine ones associated with the trait? How can these results be used to guide further genetic investigation? To answer these questions, we conduct simulation studies based on the data structure to provide some guidelines.

In the simulation studies, we keep the SNP genotypes intact, but each time we randomly select a given number of SNPs and treat them as the ones which are responsible for the variation of the trait value, then the trait values are generated from these SNPs using a linear model with an error term having standard deviation 1. We consider two settings, at each setting, 10 SNPs are randomly selected to generate the trait values, but the values of the coefficients associated with the selected SNPs are different. In the first setting,

$$\boldsymbol{\beta}(\mathcal{S}_0) = (-1.56, -1.09, 1.22, -.06, -.08, -.012, .067, -.047, -.07, .05)^t.$$

In the second setting,

$$\boldsymbol{\beta}(\mathcal{S}_0) = (-.31, .23, .42, -.32, -.33, -.26, .41, .29, -.35, -.69)^t.$$

In the first setting, the first three SNPs have effect size larger than the standard deviation 1 and are considered as major SNPs. The other ones have effect size much smaller than 1 and are not expected to be detected. In the second setting, all the ten SNPs have about the same effect size which are not big and are considered as minor SNPs. These two settings are designed to see how the tournament screening cum

Table 3: Average positive selection rate (PSR) and false discovery rate (FDR) of the tournament screening cum EBIC based on the real data structure over 200 replicates (I — setting 1, II — setting 2).

EBIC	PSR		FDR	
γ	I	II	I	II
0	.993	.638	.541	.384
$\frac{1}{2}$	.993	.580	.059	.139
1	.993	.515	.007	.084

EBIC approach perform when the SNP effects are at different levels. The simulation based on each of the two settings is repeated 200 times and the results are given in Table 3.

The simulation results show that the EBIC_1 has a tight control over the false discovery rate whether or not the associated SNPs have major or minor effects, the $\text{EBIC}_{1/2}$ still has a reasonable control over the false discovery rate, but the original BIC is too liberal. These suggest that, in the real data analysis, the SNP selected by EBIC_1 is highly likely to be a genuine one associated with the trait. Furthermore, other SNPs selected by $\text{EBIC}_{1/2}$ also stand a very good chance to be the genuine ones. Thus, further genetic verification studies should be first focused on the one selected by EBIC_1. If fund allows, those selected by $\text{EBIC}_{1/2}$ should also be further investigated.

5 Conclusion remarks

The TS cum EBIC approach developed in this article is specifically devised for feature selection while the purpose is to detect causal features. In this situation, the PSR and FDR, which are similar concepts of power and probability of type I error in hypothesis testing, are more important than the prediction accuracy of the selected model. The TS cum EBIC approach enjoys desirable properties in terms of PSR and FDR compared with other approaches. This approach is especially useful for genetic genome-wide association studies where tens or hundreds of thousands SNPs are investigated to detect etiological genetic variants and efficient statistical methods are still in wanting. As demonstrated, the approach is particularly efficient for detecting major genetic variants, it enjoys very high PSR and very low FDR. Even for minor variants, the PSR and FDR of the approach are very appealing. We believe that the TS cum EBIC approach will provide the geneticists with an important tool in their scientific exploring.

In genetic studies, it is not uncommon that the interaction between two genes, termed as epistasis effect, is prominent but the marginal effects of the genes are negligible. Without consideration of interactions, those genes will never stand a chance to be detected. With P already large, the inclusion of the consideration of interactions becomes more challenging. The TS cum EBIC approach can be easily adapted for the consideration of interactions.

Though TS is only used together with the EBIC procedure in this article, it can also be used as a general screening procedure for other purposes because of its sure screening property.

Appendix

Proof of the Sure Screening Property

We first derive some preliminary results about the order of the sample correlation between the response variable and the feature variables. Denote by $x(i,j)$ the (i,j)th entry of X and $y(i)$ the ith component of $\boldsymbol{y}$. Let C be the bound of the entries of X.

Recall that the total index set of the feature space is denoted by $\mathcal{S}^1$ and the index set of the features with non-zero coefficients in the regression model (1) is denoted by $\mathcal{S}_0$. Under the assumptions of the Theorem, we have

Lemma: *For $j \in \mathcal{S}_0$,*

$$\frac{1}{n}\sum_{i=1}^{n} x(i,j)y(i) = \beta^2(j) + o_p(n^{-1/2}\ln n). \tag{4}$$

Uniformly over $j \notin \mathcal{S}_0$,

$$\frac{1}{n}\sum_{i=1}^{n} x(i,j)y(i) = o_p(n^{-1/2}\ln n). \tag{5}$$

This lemma states in words that the correlation of the response variable with a causal feature variable is of order $O_p(1)$ and those with a non-causal feature variable is of order $o_p(n^{-1/2}\ln n)$ which tends to zero as $n \to \infty$. It is this property that distinguishes the causal features from the non-causal features in the tournament procedure. The lemma follows from an inequality of Hoeffding (Serfling, 1980, pp75) for bounded random variables. By Hoeffding's inequality, we have

$$P\left\{|\sum_{i=1}^{n} x(i,j)x(i,k)| \geq n\delta\right\} \leq 2\exp\left\{-\frac{n\delta^2}{2C^4}\right\}$$

for any n and positive constant δ. Furthermore, we have

$$P\left\{\max_{j,k\in\mathcal{S}^1, j\neq k} |\sum_{i=1}^{n} x(i,j)x(i,k)| \geq n\delta\right\} \leq 2n^{2\kappa}\exp\left\{-\frac{n\delta^2}{2C^4}\right\},$$

since $\nu(\mathcal{S}^1) = P = O(n^\kappa)$ and hence the number of pairs (j,k) is of order $n^{2\kappa}$. Let $\delta = n^{-1/2}\log n$, this inequality implies that

$$\max_{j,k\in\mathcal{S}^1, j\neq k} |\sum_{i=1}^{n} x(i,j)x(i,k)| = o_p(n^{1/2}\ln n).$$

Because of the sparsity of $\boldsymbol{\beta}$, we have

$$\boldsymbol{y} = X(\mathcal{S}_0)\boldsymbol{\beta}(\mathcal{S}_0) + \boldsymbol{\epsilon}.$$

Then it is easy to see that, for $j \in \mathcal{S}_0$,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} x(i,j)y(i) &= n^{-1}\beta^2(j)\sum_{i=1}^{n} x^2(i,j) + \sum_{k\in\mathcal{S}_0, k\neq j}[n^{-1}\sum_{i=1}^{n} x(i,j)x(i,k)] \\ &\quad +n^{-1}\sum_{i=1}^{n} x(i,j)\epsilon(i) \\ &= \beta^2(j) + o_p(n^{-1/2}\ln n). \end{aligned}$$

and that, uniformly for $j \notin \mathcal{S}_0$,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} x(i,j)y(i) &= \sum_{k\in\mathcal{S}_0} \beta(k)\{n^{-1}\sum_{i=1}^{n} x(i,j)x(i,k)\} + n^{-1}\sum_{i=1}^{n} x(i,j)\epsilon(i) \\ &= o_p(n^{-1/2}\ln n). \end{aligned}$$

The lemma is thus proved.

Now, let $\mathcal{S}$ be any subset of $\mathcal{S}^1$ such that $K < \nu(\mathcal{S}) < n$ and λ is tuned to a value λ^* such that the minimum of

$$l_p(\boldsymbol{\beta}(\mathcal{S})|\lambda^*) = \|y - X(\mathcal{S})\boldsymbol{\beta}(\mathcal{S})\|^2 + n\sum_{j\in\mathcal{S}} p_{\lambda^*}(|\beta_j|)$$

attains at some $\hat{\boldsymbol{\beta}}(\mathcal{S})$ with exactly K non-zero components. Let $\mathcal{S}^*$ be the index set of the nonzero components of $\hat{\boldsymbol{\beta}}(\mathcal{S})$. So $\mathcal{S}^* \subset \mathcal{S}$ and $\nu(\mathcal{S}^*) = K$. Let $\mathcal{S}_0^+ = \mathcal{S}_0 \cap \mathcal{S}$ and $\mathcal{S}_0^- = \mathcal{S}_0/\mathcal{S}$. That is, $\mathcal{S}_0^+$ is the set of the causal features contained in $\mathcal{S}$, and $\mathcal{S}_0^-$ is the set of the causal features not contained in $\mathcal{S}$. We are going to show that $\mathcal{S}_0^+ \subset \mathcal{S}^*$. Obviously,

$$l_p(\hat{\boldsymbol{\beta}}(\mathcal{S}^*)|\lambda^*) = \inf_{\boldsymbol{\beta}(\mathcal{S})} l_p(\boldsymbol{\beta}(\mathcal{S})|\lambda^*) < l_p(\boldsymbol{\beta}(\mathcal{S}_0^+)|\lambda^*). \tag{6}$$

We have

$$\begin{aligned}
l_p(\boldsymbol{\beta}(\mathcal{S}_0^+)|\lambda^*) &= \sum_{i=1}^n [\sum_{j\in\mathcal{S}_0^-} \beta(j)x(i,j)+\epsilon(i)]^2 + n\sum_{j\in\mathcal{S}_0^+} p_{\lambda^*}(|\beta_j|) \\
&= \sum_{j\in\mathcal{S}_0^-} \beta^2(j)\sum_{i=1}^n x^2(i,j) + \sum_{j,k\in\mathcal{S}_0^-, j\neq k} \beta(j)\beta(k)\sum_{i=1}^n x(i,j)x(i,k) \\
&\quad + \sum_{j\in\mathcal{S}_0^-} \beta(j)\sum_{i=1}^n x(i,j)\epsilon(i) + \sum_{i=1}^n \epsilon^2(i) + n\sum_{j\in\mathcal{S}_0^+} p_{\lambda^*}(|\beta_j|) \\
&= n\sum_{j\in\mathcal{S}_0^-} \beta^2(j) + \sum_{i=1}^n \epsilon^2(i) + n\sum_{j\in\mathcal{S}_0^+} p_{\lambda^*}(|\beta_j|) + o_p(n^{1/2}\ln n). \qquad (7)
\end{aligned}$$

Now, suppose $\mathcal{S}_0^+ \not\subset \mathcal{S}^*$; that is, there is at least one causal feature in $\mathcal{S}_0^+$ which is not in $\mathcal{S}^*$. Let $\mathcal{S}^{*-} = \mathcal{S}_0^- \cup (\mathcal{S}_0^+/\mathcal{S}^*)$. Clearly, $\mathcal{S}_0^- \subset \mathcal{S}^{*-}$. Let

$$H^* = X(\mathcal{S}^*)\{X'(\mathcal{S}^*)X(\mathcal{S}^*)\}^{-1}X'(\mathcal{S}^*)$$

be the projection matrix of $X(\mathcal{S}^*)$. By (4) and (5), it is seen that

$$n\{X'(\mathcal{S}^*)X(\mathcal{S}^*)\}^{-1} = I_{K\times K} + o_p(n^{-1/2}\ln n).$$

That is, it deviates from an $K \times K$ identity matrix component-wise by an order of $n^{-1/2}\ln n$. Thus, we have

$$H^* = n^{-1}X(\mathcal{S}^*)\{I + o_p(n^{-1/2}\ln n)\}X'(\mathcal{S}^*).$$

For any $j, k \notin \mathcal{S}^*$, we have

$$\begin{aligned}
X'(j)H^*X(k) &= n^{-1}X'(j)X(\mathcal{S}^*)\{I + o_p(n^{-1/2}\ln n)\}X'(\mathcal{S}^*)X(k) \\
&= n^{-1}o_p(n^{1/2}\ln n)\{I + o_p(n^{-1/2}\ln n)\}o_p(n^{1/2}\ln n) \\
&= o_p([\ln n]^2).
\end{aligned}$$

This order is uniform over all possible pairs of j, k. Then, we have

$$\begin{aligned} l_p(\hat{\boldsymbol{\beta}}(\mathcal{S}^*)|\lambda^*) &\geq \inf_{\boldsymbol{\beta}(\mathcal{S}^*)} \|X(\mathcal{S}^{*-})\boldsymbol{\beta}(\mathcal{S}^{*-}) + \boldsymbol{\epsilon} - X(\mathcal{S}^*)\boldsymbol{\beta}(\mathcal{S}^*)\|^2 \\ &= \{\boldsymbol{\beta}'(\mathcal{S}^{*-})X'(\mathcal{S}^{*-}) + \boldsymbol{\epsilon}'\}(I - H^*)\{X(\mathcal{S}^{*-})\boldsymbol{\beta}(\mathcal{S}^{*-}) + \boldsymbol{\epsilon}\} \\ &= \boldsymbol{\beta}'(\mathcal{S}^{*-})X'(\mathcal{S}^{*-})(I - H^*)X(\mathcal{S}^{*-})\boldsymbol{\beta}(\mathcal{S}^{*-}) \\ &\quad +2\boldsymbol{\beta}'(\mathcal{S}^{*-})X'(\mathcal{S}^{*-})(I - H^*)\boldsymbol{\epsilon} + \epsilon'(I - H^*)\boldsymbol{\epsilon}. \end{aligned} \tag{8}$$

It can be seen that

$$\begin{aligned} & \boldsymbol{\beta}'(\mathcal{S}_1^-)X'(\mathcal{S}_1^-)(I - H^*)X(\mathcal{S}_1^-)\boldsymbol{\beta}(\mathcal{S}_1^-) \\ = \;& \boldsymbol{\beta}'(\mathcal{S}_1^-)X'(\mathcal{S}_1^-)X(\mathcal{S}_1^-)\boldsymbol{\beta}(\mathcal{S}_1^-) - \boldsymbol{\beta}'(\mathcal{S}_1^-)X'(\mathcal{S}_1^-)H^*X(\mathcal{S}_1^-)\boldsymbol{\beta}(\mathcal{S}_1^-) \\ = \;& n\boldsymbol{\beta}'(\mathcal{S}_1^-)\boldsymbol{\beta}(\mathcal{S}_1^-) + o_p(n^{1/2}\ln n) + o_p([\ln n]^2) \\ = \;& n\boldsymbol{\beta}'(\mathcal{S}_1^-)\boldsymbol{\beta}(\mathcal{S}_1^-) + o_p(n^{1/2}\ln n). \end{aligned}$$

The variance of $\boldsymbol{\beta}'(\mathcal{S}^{*-})X'(\mathcal{S}^{*-})(I - H^*)\boldsymbol{\epsilon}$ is clearly $O(n)$ and hence its order is $O_p(n^{1/2})$. It is obvious that

$$\boldsymbol{\epsilon}'(I - H^*)\boldsymbol{\epsilon} = \sum_{i=1}^n \epsilon^2(i) + O_p(1).$$

Thus,

$$l_p(\hat{\boldsymbol{\beta}}(\mathcal{S}^*)|\lambda^*) \geq n\boldsymbol{\beta}'(\mathcal{S}^{*-})\boldsymbol{\beta}(\mathcal{S}^{*-}) + \sum_{i=1}^n \epsilon^2(i) + o_p(n^{1/2}\ln n).$$

Eventually, we have

$$\begin{aligned} & l_p(\hat{\boldsymbol{\beta}}(\mathcal{S}^*)|\lambda^*) - l_p(\boldsymbol{\beta}(\mathcal{S}_0^+)|\lambda^*) \\ \geq \;& n\{\sum_{j\in\mathcal{S}^{*-}} \beta^2(j) - \sum_{j\in\mathcal{S}_0^-} \beta^2(j)\} - n\sum_{j\in\mathcal{S}_0^+} p_{\lambda^*}(|\beta_j|) + o_p(n^{1/2}\ln n). \end{aligned}$$

Since, by assumption, $\mathcal{S}_0^- \subset \mathcal{S}^{*-}$, the first term on the right hand side of the above inequality is positive and is of order $O(n)$. Thus the total on the right hand side is larger than 0 in probability provided that

$$\sum_{j \in \mathcal{S}_0^+} p_{\lambda^*}(|\beta_j|) = o_p(1). \tag{9}$$

Then this will contradict to (6). This contradiction shows that $\mathcal{S}_0^+ \subset \mathcal{S}^*$.

We now show that (9) is true. For convenience, take the penalty as the L_1 penalty. Since $\nu(\mathcal{S}^*) = K > \nu_0$, there is at least one $j \in \mathcal{S}^*$ which is not in $\mathcal{S}_0$ such that $\hat{\beta}(j) \neq 0$. This implies that

$$2\sum_{i=1}^{n} x(i,j)\{y(i) - \sum_{k \in \mathcal{S}^*} x(i,k)\hat{\beta}(k)\} + n\lambda^* = 0, \tag{10}$$

which follows from elementary calculus. Since $j \notin \mathcal{S}_0$, the order of

$$\sum_{i=1}^{n} x(i,j)\{y(i) - \sum_{k \in \mathcal{S}^*} x(i,k)\hat{\beta}(k)\}$$

is $o_p(n^{1/2}\ln n)$. Thus for (10) to hold, λ^* must be of order $o_p(n^{-1/2}\ln n)$. Hence $\sum_{j \in \mathcal{S}_0^+} p_{\lambda^*}(|\beta_j|) = o_p(n^{-1/2}\ln n) = o_p(1)$. A similar argument can be made for the SCAD penalty but we do not present details here. The theorem is thus finally proved.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski. Budapest: Akademiai Kiado, page 267.

Amos, C. I. (1994). Robust variance-component approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**, 535-543.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate — A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57**, 289-300.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **26**(2), 123-140.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Statist. Soc. B* **64**, 641-656.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* to appear.

Chen, J. and Chen, Z. (2007). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, under second review.

Chen, Z., Chen, J. and Liu, J. (2006). A tournament approach to the detection of multiple associations in genome-wide studies with pedigree data. Working Paper 2006-09, www.stats.uwaterloo.ca. Department of Statistics & Actuarial Sciences, University of Waterloo.

Dudbridge, F. and Koeleman, B. P. (2003). Rank truncated product of P-values, with application to genome-wide association scans. *Genet. Epidemiol.* **25**, 360-366.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. of Statist.*, **32**, 407-499.

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

Fan, J. and Lv, J. (2007). Sure independence screening for ultra-high dimensional feature space. *Ann. Statist.* to appear.

Hoh, J., Wille, A. and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* **11**, 2115-2119.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* **4**, 701-709.

Huang, J., Horowitz, J. and Ma, S. (2007). Asymptotic properties of bridge estimation in sparse high-dimensional regression models. *Ann. Statist.* to appear.

Hunter, D. and Li, R. (2005). Variable selection via MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *Ann. Statist.* **35**, 1456-1486.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Stat. Assoc.* **98**, 438-455.

Marchini, J. Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413-417.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2007). "Pre-conditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.* to appear.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440-9445.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *J. Roy. Statist. Soc. B* **58**, 267-288.

Tibshirani, R, Hastie, T., Narasimhan, B. and Chu, C. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA* , **99**, 6567-6572.

Tusher, V., Tibshirani, R. and Chu, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natn. Acad. Sci. USA* , **98**, 5116-5121.

Zhang, C. -H. and Huang, J. (2007). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* to appear.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P.H. and Weir, B. S. (2002). Truncated product method for combining p-values, *Genet. Epidemiol.* **22**, 170-185.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 301-320.