International Journal of Statistics and Probability; Vol. 4, No. 3; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Unbiased Estimation for Linear Regression When $n < v$

Saeed Aldahmnai[1] & Hongsheng Dai[1]

[1] Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK

Correspondence: Hongsheng Dai, Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK. E-mail: hdaia@essex.ac.uk

Received: April 28, 2015 Accepted: May 21, 2015 Online Published: July 1, 2015

doi:10.5539/ijsp.v4n3p61 URL: http://dx.doi.org/10.5539/ijsp.v4n3p61

The first author is supported by the PhD scholarship from United Arab Emirates University, Al-Ain, UAE

Abstract

In this paper a new method is proposed for solving the linear regression problem when the number of observations n is smaller than the number of predictors v. This method uses the idea of graphical models and provides unbiased parameter estimates under certain conditions, while existing methods such as ridge regression, least absolute shrinkage and selection operator (LASSO) and least angle regression (LARS) give biased estimates. Also the new method can provide a detailed graphical correlation structure for the predictors, therefore the real causal relationship between predictors and response could be identified. In contrast, existing methods often cannot identify the real important predictors which have possible causal effects on the response variable. Unlike the existing methods based on graphical models, the proposed method can identify the potential networks while doing regression even if the data do not follow a multivariate distribution. The new method is compared with some existing methods such as ridge regression, LASSO and LARS by using simulated and real data sets. Our experiments reveal that the new method outperforms all the other methods when $n < v$.

Keywords: graphical model, unbiased estimation, LARS, LASSO, ridge regression

1. Introduction

Consider a linear regression model with a univariate response, v covariates and n independent and identically distributed (i.i.d.) observations. Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{x}_j = (x_{1j}, ..., x_{nj})'$ $(j = 1, \cdots, v)$ be the observations for the response and the jth covariate, respectively. Denote the matrix of the covariate observations by $\mathbf{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_v)$. Assume the following linear regression model representing the relationship of the response and the covariates,

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

where the elements of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ are i.i.d. random variables with mean 0 and variance σ^2.

When $n < v$, many methods have been proposed for the above models, such as Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Least Angle Regression (LARS) (Efron et al., 2004) and ridge regression (Hoerl & Kennard, 1970). Greatest attention has been paid by these methods to the case with the number of non-zero coefficients to be much less than n or v, which is usually called the 'sparse' case. These methods, however, provide biased estimates. In addition, the selected model based on LASSO and LARS can take at most n covariates (Zou & Hastie, 2005; McCann & Welsch, 2007). This will be problematic in some areas where more or even all covariates have to be included in the model. Another problem with LASSO is overshrinking the final coefficients which might produce inaccurate estimates for the coefficients (James & Radchenko, 2009). Ridge regression can include all covariates in the model, but the biased estimate makes it difficult to justify the significance levels for each covariate. This can also lead to a non-sparse model which is difficult to interpret when the number of features is large (Yuan et al., 2007). Other related approaches include the recent research of Candes & Tao (2007), Meinshausen & Yu (2009), Bickel et al. (2009), Zhang (2010) and Lin et al. (2014). However, their estimates are still biased which might not be recommended in general (Washington et al., 2010; Zhang, 2010).

For regression analysis with $n > v$, standard Least Squares Estimate (LSE) actually uses the saturated model assumption, i.e. any pair of covariates are correlated conditional on the other covariates. When $n < v$, the full saturated least squares cannot be used. A good alternative is the unsaturated model, by adding constraints to the conditional dependency structure for the covariates. However, all existing methods did not consider the conditional correlation and thus they cannot identify the true model when $n < v$. This might be very important in genetic studies

where the identification of gene networks is needed. Traditional graphical methods only focus on networks based on multivariate distributions. However, if the distribution is not multivariate and we need to do regression, then traditional methods will not work.

This paper develops an unbiased estimation method via graphical models (GLSE), which can provide a much better solution than all other existing methods. The proposed theory has a strong connection with graphical models and takes into account the conditional correlation between covariates. Therefore it could possibly include the causal interpretation between response and covariates and identify the true regression model. Under certain trivial assumptions, the estimator produced by the new method is unbiased in addition to identifying the required networks.

Another advantage of GLSE is that it considers much more candidate models than existing methods. For the regression model with covariates $(X_1, \cdots, X_v)$, the estimate produced under the situation where all X_j are independent should be different from the regression estimate produced under the situation when some X_j are correlated. By taking into account the detailed conditional correlation structure, this new method actually searches from a much larger space of potential regression models than all existing methods. With v covariates, existing methods carry out model selection from 2^v different sets of models (at most). In contrast, if all covariates should be included in the model, the GLSE could search from $2^{v(v-1)/2}$ different models, since there will be so many different models according to the different conditional correlation structures among covariates X_j.

This paper is structured as follows. We provide necessary notations and definitions of graphical models in Section 2.1. Then in Section 2.2 we present the main methodology of GLSE and its properties. The model selection for the underlying graph structure is provided in Section 2.3. Simulation studies and data analysis are given in Sections 3. The paper ends with a conclusion in Section 4.

2. Method

2.1 Notations

2.1.1 Graphs

We follow the notations in Lauritzen (1996). One may also refer to Whittaker (2009) or Dawid and Lauritzen (1993) for more details on graphical models.

An *undirected graph* is denoted by $G = (V, \mathcal{E})$, where V is the set of vertices, say $V = \{1, 2, \cdots, v\}$, and $\mathcal{E}$ is the set of edges, a subset of $V \times V$. We usually use notation $\{i, j\}$ to denote the edge between vertex i and j. The elements in V correspond to the index set for all covariates in the regression model. Therefore each covariate corresponds to a vertex in the graph G.

If $A \subseteq V$, the subset A *induces* a subgraph $G_A = (A, \mathcal{E}_A)$, where $\mathcal{E}_A = E \cap A \times A$. A graph is *complete* if all vertices are joined by an edge and a subset is *complete* if it induces a complete subgraph from G. A complete subset that is maximal (with respect to $\subseteq$) is called a *clique*.

A triple (A, B, C) of disjoint subsets of the vertex set V of an undirected graph G is said to form a *decomposition* of G if $V = A \cup B \cup C$ and the following conditions hold (Lauritzen, 1996):

- B separates A from C;
- B is a complete subset of V;

An undirected graph G is *decomposable* if it holds one of the following:

- Graph G is complete.
- There is a proper decomposition (A, B, C) into decomposable subgraphs $G_{A\cup B}$ and $G_{B\cup C}$.

Consider a sequence of sets $C_1, ..., C_q$ and define that

$$\begin{aligned} H_i &= C_1 \cup ... \cup C_i, \\ S_i &= H_{i-1} \cap C_i. \end{aligned}$$

Then if the following conditions are satisfied, the given sequence $C_1, ..., C_q$ is said to be a *perfect sequence* (Lauritzen, 1996):

- There exist an $i < j$, for any $j > 1$, such that $S_j \subseteq C_i$;
- The sets S_j are complete for all values of j;

A decomposable graph G allows a perfect ordering of cliques (Golumbic, 2004).

2.1.2 Matrices and Vectors

Throughout this paper, we reserve the capital letters A, B, C and S to denote a subset of V and $|\cdot|$ to denote the number of elements for an edge or vertex set. The lower letters i, j, k are reserved for the notation of indices.

The following notations mainly follow the style in Lauritzen (1996). A vector $\boldsymbol{Z} \in \mathbb{R}^v$ can also be written as $(Z_j)_{j\in V}$. The jth element Z_j can also be written as $(\boldsymbol{Z})_j$. Denote $\boldsymbol{Z}_A$ as an a-dimensional subvector ($a = |A|$) of $\boldsymbol{Z}$, with $\boldsymbol{Z}_A = (Z_j)_{j\in A}$. Denote $[\boldsymbol{Z}_A]^\Gamma$ as a v-dimensional vector obtained by filling up 0s, with

$$\left([\boldsymbol{Z}_A]^\Gamma\right)_j = \begin{cases} Z_j & \text{if } j \in A \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We define similar notations for matrices. A $v \times v$ matrix $\mathbf{z}$ can also be written as $(z_{kj})_{k,j\in V}$. We may also write $z_{kj} = (\mathbf{z})_{kj}$. Denote $\mathbf{z}_{AB} = (z_{kj})_{k\in A, j\in B}$, a submatrix of $\mathbf{z}$. Denote $[\mathbf{z}_{AB}]^\Gamma$ as a $v \times v$-dimensional matrix obtained by filling up 0s, with

$$\left([\mathbf{z}_{AB}]^\Gamma\right)_{jk} = \begin{cases} z_{jk} & \text{if } j \in A,\ k \in B \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

With the above notation, we can use $\mathbf{x}_A$ to denote the covariate matrix only having covariates with indices in set A. The notation $[(\mathbf{x}_A'\mathbf{x}_A)^{-1}]^\Gamma$ denotes a $v \times v$-dimensional matrix obtained by filling up 0s, with

$$\left([(\mathbf{x}_A'\mathbf{x}_A)^{-1}]^\Gamma\right)_{jk} = \begin{cases} \left((\mathbf{x}_A'\mathbf{x}_A)^{-1}\right)_{jk} & \text{if } j,k \in A \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Clearly $[\mathbf{x}_B]^\Gamma[(\mathbf{x}_A'\mathbf{x}_A)^{-1}]^\Gamma = \mathbf{0}$ if $B \cap A = \phi$, the empty set.

2.2 The Idea of Unbiased Estimation via Graphical Model

In this section, we present the basic idea of the GLSE for the regression model (1). We assume that all response values and covariate values have been centered and the covariates variables are quantitative. First we introduce a condition.

Condition 1 *Suppose that the set V can be partitioned into disjoint sets A, B and C and the sample size* $n > \max\{|A| + |B|, |B| + |C|\}$. Note that this condition is weaker than standard LSE requirement $n > v = |A| + |B| + |C|$.

Denote the sample covariance matrix as $ssd = \mathbf{x}'\mathbf{x}$. Given that Condition 1 holds true, an estimator for $\boldsymbol{\beta}$ can be defined as

$$\hat{\boldsymbol{\beta}}^s = \left[\left[(ssd_{A\cup B})^{-1}\right]^\Gamma + \left[(ssd_{B\cup C})^{-1}\right]^\Gamma - \left[(ssd_B)^{-1}\right]^\Gamma\right]\mathbf{x}'\boldsymbol{y}. \tag{5}$$

Clearly under Condition 1, the matrix inversions in the above formula are available.

To discuss the unbiasedness property of the above estimator, we first introduce some definitions. For an estimator $\hat{\boldsymbol{\beta}}$, we introduce three different types of unbiasedness regarding the regression model given in (1).

1. Unbiased estimate (UE): If $\mathbb{E}_{\mathbf{x},\boldsymbol{y}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, the estimator $\hat{\boldsymbol{\beta}}$ is said to be unbiased.
2. Strong-conditional unbiased estimate (SUE): If $\mathbb{E}_{\boldsymbol{y}}(\hat{\boldsymbol{\beta}}|\mathbf{x}) = \boldsymbol{\beta}$ then the estimator $\hat{\boldsymbol{\beta}}$ is said to be strongly conditionally unbiased.
3. Weak-conditional unbiased estimate (WUE): If $\mathbb{E}_{\mathbf{x}_{V\setminus A},\boldsymbol{y}}(\hat{\boldsymbol{\beta}}|\mathbf{x}_A) = \boldsymbol{\beta}$, then the estimator $\hat{\boldsymbol{\beta}}$ is said to be weakly conditionally unbiased, where A is denoting a proper nonempty subset of V.

It is easy to show that SUE $\Rightarrow$ WUE $\Rightarrow$ UE, where $\Rightarrow$ mean "implies".

Write the covariate matrix as $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)$, then under the following condition, we can show that (5) is an unbiased estimate.

Condition 2 *The sets A, B and C form a decomposition, with B as the separator. The sets* $\mathbf{x}_A$ *and* $\mathbf{x}_C$ *are conditionally independent on* $\mathbf{x}_B$*, such that*

(a)

$$\begin{aligned} \mathbf{x}_C &= \mathbf{x}_B \cdot \mathbf{r}_{BC} + \boldsymbol{\xi}_C, \ \ \mathbb{E}(\boldsymbol{\xi}_C) = \mathbf{0}, \\ \mathbf{x}_A &= \mathbf{x}_B \cdot \mathbf{r}_{BA} + \boldsymbol{\xi}_A, \ \ \mathbb{E}(\boldsymbol{\xi}_A) = \mathbf{0}, \end{aligned} \tag{6}$$

(b) *where $\mathbf{r}_{BC}$ and $\mathbf{r}_{AB}$ are constant matrices;*

$$\boldsymbol{\xi}_A \perp \boldsymbol{\xi}_C | \mathbf{x}_B. \tag{7}$$

Now it is ready to introduce the following theorem.

Theorem 1 *Under Condition 1 and Condition 2, the estimator in (5) is unbiased and even weak-conditional unbiased,*

$$\mathbb{E}_{\mathbf{x},y}(\hat{\boldsymbol{\beta}}^s) = \mathbb{E}_{\mathbf{x}_{V\setminus B},y}(\hat{\boldsymbol{\beta}}^s | \mathbf{x}_B) = \boldsymbol{\beta}.$$

Proof. We have

$$\begin{aligned} & \mathrm{E}(\hat{\boldsymbol{\beta}}^s | \mathbf{x}_B) \\ = \ & E\left(\left\{\left([(ssd_{A\cup B})^{-1}]^\Gamma + [(ssd_{B\cup C})^{-1}]^\Gamma - [(ssd_B)^{-1}]^\Gamma\right)\mathbf{x}'\right\}\mathbf{x}\boldsymbol{\beta} | \mathbf{x}_B\right) \\ = \ & \mathrm{E}\left[\begin{pmatrix} \mathbf{I}_A & \mathbf{0} & \\ \mathbf{0} & \mathbf{I}_B & (ssd_{A\cup B})^{-1}\mathbf{x}'_{A\cup B}\mathbf{x}_C \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (ssd_{B\cup C})^{-1}\mathbf{x}'_{B\cup C}\mathbf{x}_A & \mathbf{I}_B & \mathbf{0} \\ & \mathbf{0} & \mathbf{I}_C \end{pmatrix}\right. \\ & \left. - \begin{pmatrix} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_A & \mathbf{I}_B & (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_C \\ & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \middle| \mathbf{x}_B \right] \boldsymbol{\beta}. \end{aligned} \tag{8}$$

Condition 2 (a) gives

$$\mathrm{E}(\mathbf{x}_C | \mathbf{x}_B) = \mathbf{x}_B \boldsymbol{r}_{BC} = \mathbf{x}_{A\cup B}\begin{pmatrix} \mathbf{0} \\ \boldsymbol{r}_{BC} \end{pmatrix}$$

and we then have

$$\begin{aligned} & \mathrm{E}\left[(ssd_{A\cup B})^{-1}\mathbf{x}'_{A\cup B}\mathbf{x}_C - \begin{pmatrix} \mathbf{0} \\ (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_C \end{pmatrix} \middle| \mathbf{x}_B \right] \\ = \ & (ssd_{A\cup B})^{-1}\mathbf{x}'_{A\cup B}\mathbf{x}_{A\cup B}\begin{pmatrix} \mathbf{0} \\ \boldsymbol{r}_{BC} \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_B\boldsymbol{r}_{BC} \end{pmatrix} \\ = \ & \begin{pmatrix} \mathbf{0} \\ \boldsymbol{r}_{BC} \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ \boldsymbol{r}_{BC} \end{pmatrix} = \mathbf{0}. \end{aligned} \tag{9}$$

Similarly we can prove

$$\mathrm{E}\left[(ssd_{B\cup C})^{-1}\mathbf{x}'_{B\cup C}\mathbf{x}_A - \begin{pmatrix} (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_A \\ \mathbf{0} \end{pmatrix} \middle| \mathbf{x}_B \right] = \mathbf{0} \tag{10}$$

Substitution (9) and (10) back to (8), we have

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}^s | \mathbf{x}_B) = \boldsymbol{\beta}$$

and therefore $\mathrm{E}(\hat{\boldsymbol{\beta}}^s) = \boldsymbol{\beta}$.

Remark 1 The standard least squares estimate (available when $n > v$) is SUE, i.e. $\mathbb{E}_{\mathbf{x},y}(\hat{\boldsymbol{\beta}}^{lse}) = \mathbb{E}_y(\hat{\boldsymbol{\beta}}^{lse} | \mathbf{x}) = \boldsymbol{\beta}$. When $n < v$, all existing methods such as ridge regression, LASSO, or LARS are biased, i.e. $\mathbb{E}_{\mathbf{x},y}(\hat{\boldsymbol{\beta}}^{other}) \neq \boldsymbol{\beta}$.

Remark 2 Note that Condition 2 will be satisfied in many cases. For example, if $\boldsymbol{X}_i = (x_{i1}, \cdots, x_{iv})$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and the distribution of $\boldsymbol{X}_i$ factorizes according to a decomposable graph $g = (V, \mathcal{E}_g)$, with cliques $A \cup B$ and $B \cup C$ and separator B. Condition 2 implies the conditional independence of $\mathbf{x}_A$

and $\mathbf{x}_C$ given $\mathbf{x}_B$ and a linear relationship between $\mathbf{x}_A$ (and $\mathbf{x}_C$) and $\mathbf{x}_B$. Given a different underlying graph structure for the covariates, the proposed estimator will be different. In practice, we can search the best underlying graph structure and this will be provided in Section 2.3.

Remark 3 The linear assumption in Condition 2 will not limit the applicability of the method since one can always use variable transformation to achieve a linear relationship, if the covariates are quantitative variables. In addition, as any non-linear relation can be approximated via a polynomial, the nonlinear dependence on X can be viewed as linear dependence on $X, X^2, \cdots$.

2.2.1 The General Unbiased Estimates

In general, the associated graph g may be decomposed into many cliques or separators. Suppose that the graph g is decomposable and let C denote the set of cliques and $\mathcal{S}$ denote the set of separators. The general formula of the GLSE is

$$\hat{\boldsymbol{\beta}}^g = \left[\sum_{C\in C}[(ssd_C)^{-1}]^{\Gamma} - \sum_{S\in\mathcal{S}}[(ssd_S)^{-1}]^{\Gamma}\right]\mathbf{x}'\boldsymbol{y}. \tag{11}$$

Obviously, under the following condition, the GLSE exists.

Condition 3 *The sample size* $n > \max_{C\in C}\{|C|\}$.

Under the following condition, we can also show that $\hat{\boldsymbol{\beta}}^g$ is unbiased.

Condition 4 *Write the cliques and separators of g in the perfect ordering, as* $C_1, \cdots, C_q$ *and* $S_2, \cdots, S_{q-1}$. *They are such*

(a)

$$\begin{aligned}\mathbf{x}_{C_k\setminus S_k} &= \mathbf{x}_{S_k}\cdot\mathbf{r}_k + \boldsymbol{\xi}_k, \quad E(\boldsymbol{\xi}_k) = \mathbf{0}, \quad k = 2, \cdots, q,\\ \mathbf{x}_{C_1\setminus S_2} &= \mathbf{x}_{S_2}\cdot\mathbf{r}_1 + \boldsymbol{\xi}_1, \quad E(\boldsymbol{\xi}_1) = \mathbf{0},\end{aligned} \tag{12}$$

where $\mathbf{r}_k$ $(k = 1, \cdots, q)$ *are constant matrices;*

(b) For any $k = 2, \cdots, q$,

$$(\boldsymbol{\xi}_k, \cdots, \boldsymbol{\xi}_q) \perp (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{k-1})|\mathbf{x}_{S_k}. \tag{13}$$

Proposition 1 *Under Condition 3 and Condition 4, the estimator in (11) is unbiased,*

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^g) = \boldsymbol{\beta}.$$

The proof of Proposition 1 can be done recursively similar to Theorem 1.

2.2.2 Covariance Matrix Estimation for the GLSE

By denoting $\mathbf{K} = \sum_{C\in C}\left[(ssd_C)^{-1}\right]^{\Gamma} - \sum_{S\in\mathcal{S}}\left[(ssd_S)^{-1}\right]^{\Gamma}$, we can show that under Condition 4,

$$\text{var}(\hat{\boldsymbol{\beta}}^g) = \mathbb{E}((\hat{\boldsymbol{\beta}}^g - \boldsymbol{\beta})^2) = \mathbb{E}\left[(\mathbf{K}\mathbf{x}'\mathbf{x}\boldsymbol{\beta} - \boldsymbol{\beta})(\mathbf{K}\mathbf{x}'\mathbf{x}\boldsymbol{\beta} - \boldsymbol{\beta})' + \sigma^2(\mathbf{K}\mathbf{x}'\mathbf{x}\mathbf{K}')\right].$$

It is not easy to simplify the above formula and derive a simple estimator for the above variance of $\hat{\boldsymbol{\beta}}^g$. The bootstrap estimate is a good alternative. We can also use the following estimator for the variance,

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}^g) = \left(\mathbf{K}\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}^g - \hat{\boldsymbol{\beta}}^g\right)\left(\mathbf{K}\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}^g - \hat{\boldsymbol{\beta}}^g\right)' + \hat{\sigma}^2(\mathbf{K}\mathbf{x}'\mathbf{x}\mathbf{K}') \tag{14}$$

since any random variable is always an unbiased point estimate to its own mean. We will show that (14) does not work well for data with small sample size, but when the sample size is large, the estimator works well.

When we use bootstrap to estimate $\text{var}(\hat{\boldsymbol{\beta}}^g)$, care should be taken in that a bootstrapped sample can be used only if the number of distinct observations in the resampled data is greater than the number of variables within the maximum clique.

2.2.3 Discussion on the Unbiasedness and Condition 4

Condition 4 will be reasonable in many real applications. When there are many predictors, some of them are more likely to be conditionally independent given a subset of variables. For example, in finance studies the return of a portfolio may depend on a large number of assets. If we simply look at the correlation of these assets they may all be correlated marginally. However, some of them may be independent given a certain subset (for other example see Carvalho & West (2007)). Also in bioinformatics studies, the covariates may be a pool of genes where some genes may be related due to an intermediate gene. There has been a vast literature in the research of graphical models, which can learn the conditional dependency structure for many random variables even if the sample size is very small. For further study see Dobra et al. (2004); Markowetz & Spang (2007); Kramer et al. (2009); Talluri & Shete (2014) and the work cited there.

Condition 4 allows the possibility to study the detailed conditional independence/correlation structure under a regression framework and this will help us to solve the challenges when $n < v$. Traditional analyses did not consider the conditional dependency between the covariates. It is fine to ignore this and apply the saturated model (any pair of covariates is conditionally correlated given the remaining) when the sample size is large enough. But when $n < v$, certain conditions have to be added to achieve a reasonable result. Traditional methods uses other constraints or penalty terms and trade off the bias and variance. This will distort the real conditional correlation between covariates. That is why existing methods for $n < v$ may provide spurious correlation between covariates and predictors and cannot identify the real important causal predictors. For instance, Sun & Zhang (2010) showed that LASSO selects at least one variable at a 50% chance to predict y which might be totally non-significant variable(s) thus leading to a spurious correlation. Researchers in these field often select certain variables manually without any statistical justification.

2.3 Graphical Model Selection

In practice, the underlying graph structure g is unknown. We need to develop an algorithm to find which graph g is the *best* for the data. This is another advantage of GLSE since the underlying graph structure could lead to certain causal interpretations, which cannot be done via all existing methods.

There have been some covariance model selection methods developed under a Bayesian framework, such as Jones et al. (2005) and Dai (2008). We here proposed a new one for graphical model selection under a regression framework. It is natural to consider the following criteria for model selection: the best graph is given by minimizing a target function $\mathbb{T}(\boldsymbol{\beta}, g, \lambda)$.

$$(\hat{\boldsymbol{\beta}}, \hat{g}, \hat{\lambda}) = \arg \min_{\boldsymbol{\beta}, g \in \mathcal{G}, \lambda} \mathbb{T}(\boldsymbol{\beta}, g, \lambda) \tag{15}$$

$$\mathbb{T}(\boldsymbol{\beta}, g, \lambda) = \|\mathbf{y} - \mathbf{x}\boldsymbol{\beta}\|^2 + \lambda |\mathcal{E}_g| \tag{16}$$

where $\mathcal{G}$ is the set of all possible graphs and $|\mathcal{E}_g|$ is the number of edges in graph g. This criteria aims to search for the estimate and graph structure, corresponding to the minimum of the sum of square errors plus a penalty term.

The penalty term tells us how we should shrink the graph. In practice, we may fix the value of λ and only search for $\boldsymbol{\beta}$ and g. If one prefers a very sparse graph, then a large value of λ should be chosen; if one prefers a more saturated graph, a small value of λ should be chosen.

We recommend to consider more saturated graphs in practice. This is because more saturated graphs will allow more information on the correlation structure. To show this, a graphical comparison is done by generating a data set similar to the one simulated in Section 3.1. For convenience, the graph generated is allowed to add only up to 30 edges shown on the y-axis in Figure 1. The figure shows that a smaller value of λ will give smaller sum of square errors, i.e. more accurate prediction. Therefore, we may simply set $\lambda = 0$ and ignore the penalty term, except that in some situations a sparse graph is preferred. In the analysis of real data, $\lambda = 0$ is taken.

When searching in the whole graph space $\mathcal{G}$, we can use a stepwise selection procedure. The method considers adding/deleting edges one by one to/from the current graph. When an edge under consideration is not in the current graph, it will be added if the addition makes an improvement in terms of the predetermined criteria, otherwise it will not be added. When an edge under consideration is in the current graph, it will be deleted if the deletion makes an improvement in terms of the predetermined criteria, otherwise it will not be deleted. Similar graphical model selection algorithms for multivariate Gaussian models have been proposed by Jones et al. (2005)..

As an example, we used this algorithm to find the underlying graph structure for the data set generated in Section 3.1. The true graph structure is given in Figure 3. The two selected graphs for different sample sizes are given in Figures 2.

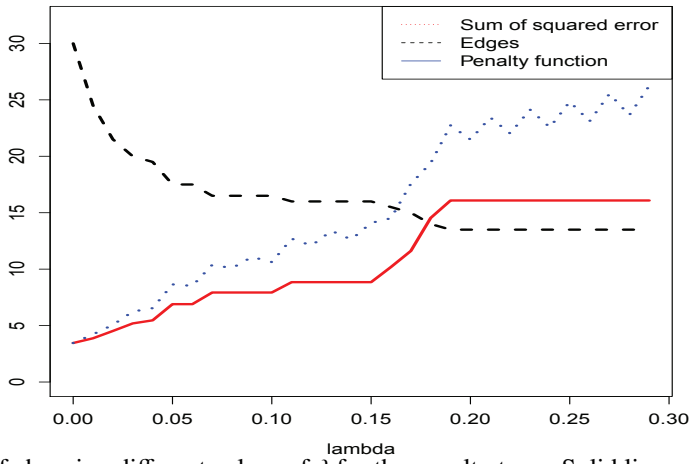

Figure 1. The effect of choosing different values of λ for the penalty term. Solid line: sum of squared errors; dotted line: penalty function; dashed line: the number of edges corresponding to the graph giving the minimum target function value.

Algorithm 1 Pseudocode of the GLSE graph selection

1: Start graph $g = (V, \mathcal{E})$, which can be an empty (or a given decomposable) graph such that $n > \max_{C \in \mathcal{C}} |C|$
2: Generate all possible graphs, g_i, such that there is only one edge difference between g_i and the current graph g. All such g_i are such that they have no cordless K-cycles ($K \geq 4$, to be decomposable) and such that $n > \max_{C \in \mathcal{C}} |C|$
3: Find the graph g_i^* such that g_i^* minimise the target function $\mathbb{T}(.)$ (given in (15))
4: Go to step 2 with the selected graph g_i^* and iterate until the best one is found.

In the original graph there are 13 edges having conditional correlations greater than absolute value 0.1. The graph chosen by the algorithm for $n = 15$, Figure 2 (a), has detected 9 of the these edges in the original graph. Also, the chosen graph detected 4 of the edges with conditional correlation less than or equal to absolute value 0.1, in the original graph. The graph for $n = 100$ detected 11 edges having conditional correlations greater than absolute value 0.1 and 7 edges with conditional correlation less than or equal to absolute value 0.1 of the original graph. It follows that by increasing the sample size, most of the original edges can be detected by using the algorithm.

The model selection step is the most challenging part for the GLSE method, since once a good graph structure is identified GLSE for regression parameters can be achieved straight forward. The computational cost is quite heavy to search the whole graph space, but step 2 of Algorithm 1 can be improved significantly via parallel computation.

3. Results

This section provides simulation studies and a real data analysis to assess the proposed method empirically.

3.1 Simulation

In this section, we provide detailed simulation study to show that our GLSE has much smaller bias than other existing methods such as LASSO, ridge regression and LARS. We compare the results of our proposal from the simulation model with those of LASSO, LARS and ridge regression where the penalty parameters are obtained by cross validation. Note that the GLSE $\hat{\boldsymbol{\beta}}^g$ depends on the graph g therefore we provide simulation results when the true graph is known and when it is unknown.

We simulate our data from model (1) with $v = 20$ and $(X_1, \cdots, X_v)$ following a multivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Sigma}$. The concentration matrix $\boldsymbol{\Sigma}^{-1}$ is associated with a graph, shown in Figure 3, where the conditional correlation is indicated by numbers above each line (for the case when true graph is known). Then the response $\boldsymbol{y}$ is generated from model (1), where the random errors ϵ are normally distributed with mean 0 and standard error $\sigma = 0.5$. We choose $n = 15 < v = 20$. The true parameter values of $\boldsymbol{\beta}$ are given in Table 1.

(*a*) n=15 (*b*) n=100

Figure 2. The selected graphs.

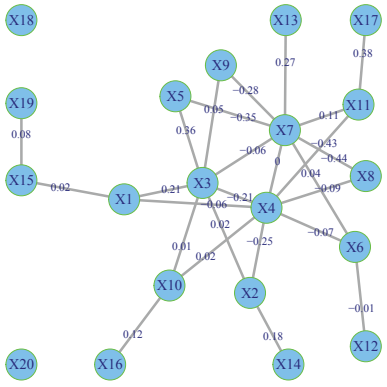

Figure 3. Graph structure for the covariates used in the simulation.

Table 1. Results from simulation (with known and unknown graphs), $n = 15, v = 20$; SD: the Monte Carlo standard error for the 500 replicates; SE is the mean of the 500 standard error estimates.

		Ridge		LASSO		LARS		GLSE (with known graph)				GLSE (with unknown graph)			
	β	$\hat{\beta}_{ridge}$	SD	$\hat{\beta}_{lasso}$	SD	$\hat{\beta}_{lar}$	SD	$\hat{\beta}$	SE (14)	SD	SE (Bootstrap)	$\hat{\beta}$	SE (14)	SD	SE (Bootstrap)
X_1	0.5	0.256	0.594	0.129	0.472	0.050	0.300	0.456	3.571	1.605	1.612	0.279	0.326	1.081	1.293
X_2	1	0.369	0.704	0.214	0.630	0.060	0.289	1.053	4.492	1.899	2.417	0.600	0.393	1.312	1.180
X_3	0.6	0.092	0.578	0.063	0.470	0.026	0.253	0.472	4.208	1.839	1.594	0.218	0.390	1.489	1.280
X_4	-1	-0.361	0.368	-0.391	0.498	-0.252	0.423	-1.009	2.734	1.178	0.919	-0.488	0.238	1.325	1.230
X_5	0.7	0.393	0.337	0.423	0.510	0.323	0.471	0.698	2.380	0.924	0.816	0.538	0.220	0.884	0.779
X_6	-1	-0.567	0.763	-0.317	0.658	-0.145	0.410	-0.975	4.541	1.863	2.042	-0.658	0.420	1.303	1.233
X_7	1.1	0.399	0.332	0.526	0.582	0.439	0.562	1.098	2.911	1.130	1.174	0.565	0.196	0.599	0.959
X_8	-0.6	-0.173	0.355	-0.219	0.448	-0.121	0.303	-0.602	2.358	0.965	1.011	-0.257	0.194	0.624	1.160
X_9	1.5	0.885	0.627	0.720	0.844	0.445	0.718	1.585	3.965	1.479	1.361	1.159	0.348	1.081	1.226
X_{10}	0.5	0.258	0.778	0.097	0.474	0.019	0.215	0.408	4.739	2.076	1.136	0.306	0.425	1.324	1.345
X_{11}	0.8	0.243	0.287	0.274	0.426	0.214	0.364	0.808	2.233	0.888	1.016	0.363	0.194	0.802	0.950
X_{12}	-0.8	-0.466	0.818	-0.218	0.588	-0.062	0.278	-0.807	4.623	1.912	1.625	-0.593	0.494	1.647	1.296
X_{13}	1.8	0.822	0.475	1.051	0.813	0.849	0.771	1.809	2.873	0.796	0.818	1.102	0.212	0.734	1.032
X_{14}	0.9	0.461	0.617	0.280	0.545	0.154	0.397	0.970	3.679	1.567	1.315	0.631	0.399	1.617	1.299
X_{15}	-0.6	-0.295	0.571	-0.179	0.501	-0.102	0.321	-0.588	3.191	1.303	2.647	-0.452	0.316	0.969	1.307
X_{16}	1.3	0.654	0.851	0.271	0.624	0.110	0.377	1.395	4.926	2.000	1.635	0.954	0.518	2.037	1.357
X_{17}	0.1	-0.040	0.364	-0.008	0.381	-0.009	0.286	0.120	2.567	1.109	0.947	0.019	0.230	0.700	1.356
X_{18}	1.7	0.882	0.934	0.467	0.825	0.218	0.611	1.599	5.135	2.094	1.985	1.241	0.509	1.754	1.258
X_{19}	-2	-1.027	0.678	-0.890	0.930	-0.592	0.818	-1.935	3.914	1.441	1.084	-1.414	0.384	1.045	1.337
X_{20}	0.8	0.426	0.814	0.167	0.559	0.079	0.305	0.662	4.429	1.868	1.873	0.465	0.484	1.457	1.296

Table 1 gives coefficient estimates and Monte Carlo estimates of standard errors for the ridge, LASSO, LARS and the proposed GLSE method. The table also shows the standard error estimates for the GLSE (with known and unknown graph structures) method. When the true graph is known, GLSE outperforms all other three methods in the coefficients estimates. Although the Monte Carlo estimates of standard errors for ridge, LASSO and LARS are smaller than that of the GLSE, the new method being an unbiased estimation should be preferred. We can also see from the results that the mean of the 500 bootstrap standard error estimate is very close to the Monte Carlo standard deviation for the 500 replicated estimates. This implies that bootstrap estimation for the standard error performs well. However, we also noted that the standard error estimate based on (14) does not work well for small sample sizes. This problem can be fixed by increasing n.

In the case when the graph is estimated, the results given in the last 4 columns of Table 1 leads to the same conclusion. Here also, the GLSE outperforms the others in coefficient estimation.

The computational cost for the proposed algorithm is not heavy with modern parallel computing technology. Both serial and parallel computing time is considered when the true graph is unknown. It is noted that on a machine with 8GB of memory and 2.7GHz processor, the time taken is approximately 10 hours. When the parallel processing was used, with 60 cores, the computational time reduced to approximately 9 minutes. For the case when the true graph structure is known, parallel computing is not needed in that the time taken serially (using one processor) is just a few seconds.

3.2 Real Data Analysis

We consider the data used in Scheetz et al. (2006) which consist of gene expressions from eye tissues of 12-week-old rats. The data set has 120 observations on more than 31,000 genes. The data set is available with in the R library "flare" (Li et al., 2014) where the dimension of the data set is reduced to 200 genes, which could have high possibilities to relate to TRIM32. The gene TRIM32 carries a significant biological information, for example, this gene might lead to Bardet-Biedl syndrome, which is a heterogeneous genetic disease in several organs including eye retina (Huang et al., 2008).

In summary, the data analysis is a regression problem with $n = 120$ and $v = 200$ where identifying the correlation structure in the data set is desirable in that most of the variable in the data are correlated. To show this, 10 variable are selected randomly from the data and their scatter plot matrix depicting the marginal correlations is given in Figure 4.

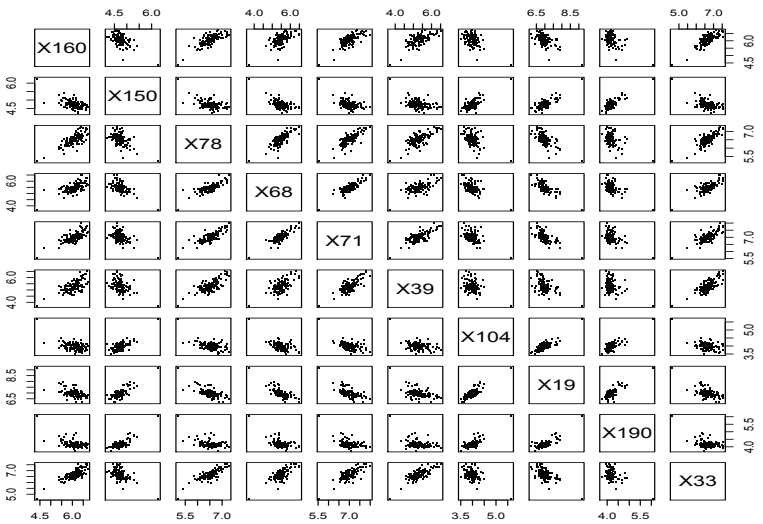

Figure 4. Scatter plot (marginal correlations) for randomly chosen 10 variables of the data. Most of the variables are correlated with each other.

We applied ridge, LASSO, LARS and our proposed GLSE method on this data set. For ridge, LASSO and LARS, cross validation is used for obtaining the penalty parameter. The LASSO method found 24 non-zero coefficients while LARS found 25 coefficients to be non-zero. Ridge regression selected all the variables and results are not provided here as it cannot give any information about significance levels. The proposed method kept all 200 variables in the model, among which 7 coefficients are significantly non-zero under a 10% significance level and

12 are significant under a 20% significance level. Note that we used bootstrap to find the confidence interval for coefficient estimates and decide the significance level for the GLSE. The results are shown in Table 2. One advantage of GLSE is that it can provides the significance levels, while none of other existing methods can do this properly. The reason for this is that reporting a standard error of a penalized estimate can give a mistaken impression of high precision. It completely ignores the inaccuracy introduced due to the bias. Therefor, confidence statements, such as those made based on bootstrap confidence intervals, might lead to incorrect decisions (Goeman et al., 2012; Van & Putter, 2011).

Table 2. Variable selected and coefficient estimates by LARS, lasso and GLSE. Significant at 5%: "**", significant at 10%: "*" and significant at 20%: " ".

LARS		Lasso		GLSE		
Variable selected	$\hat{\beta}_{lars}$	Variable selected	$\hat{\beta}_{lasso}$	Variable selected	$\hat{\beta}_{GLSE}$	
X_2	-0.009	X_2	-0.019	X_{48}	0.116	
X_{11}	0.017	X_4	0.042	X_{54}	0.077	*
X_{13}	0.001	X_8	-0.031	X_{62}	-0.064	*
X_{42}	0.040	X_{11}	-0.011	X_{64}	0.087	**
X_{54}	0.041	X_{13}	0.021	X_{65}	0.068	*
X_{58}	0.008	X_{21}	-0.045	X_{87}	-0.169	**
X_{60}	0.013	X_{29}	-0.006	X_{102}	-0.076	
X_{62}	-0.062	X_{33}	0.006	X_{146}	0.189	*
X_{76}	-0.012	X_{36}	0.039	X_{149}	-0.106	
X_{87}	-0.116	X_{42}	0.100	X_{155}	0.078	
X_{90}	-0.016	X_{54}	0.043	X_{170}	0.067	
X_{106}	0.011	X_{55}	-0.007	X_{180}	0.118	**
X_{109}	-0.008	X_{58}	0.024			
X_{110}	-0.004	X_{60}	0.026			
X_{136}	-0.003	X_{62}	-0.079			
X_{146}	0.022	X_{65}	0.008			
X_{148}	0.008	X_{70}	-0.103			
X_{153}	0.079	X_{72}	-0.019			
X_{155}	0.025	X_{87}	-0.140			
X_{158}	-0.005	X_{106}	0.038			
X_{180}	0.021	X_{146}	0.031			
X_{185}	-0.027	X_{153}	0.152			
X_{187}	-0.015	X_{155}	0.062			
X_{188}	-0.018	X_{160}	0.047			
X_{200}	-0.043					

The estimated graph showing the conditional correlations structure of genes is given in Figure 5. It has been shown in the original study of Scheetz et al. (2006) the some of the genes are highly correlated that in turn regulate the genes responsible for the eye disease. This means that the genes have a dependency structure that need to be identified. Our graphical model has identified this really important structure whereas none of the existing methods can.

4. Discussion

We have proposed a novel unbiased estimation method via graphical models (GLSE) for linear regression, specifically, when $n < p$. It has been shown that the proposed method is unbiased under some regularity constraints imposed on the predictor variables. These constraints are fulfilled quite often in many real life applications. Simulation has been done to verify empirically the validity and unbiasedness of the proposed method along with application on a microarray gene expression data set. The results of the method on simulated and real data sets are compared with those of ridge, LARS and LASSO for linear regression.

The main difference between our proposed GLSE method and the traditional graphical models is that unlike the latter, the GLSE does not consider the normality assumption for the covariates. This increases the scope of the application of the proposed method in that variables in a data set are not always normally distributed.

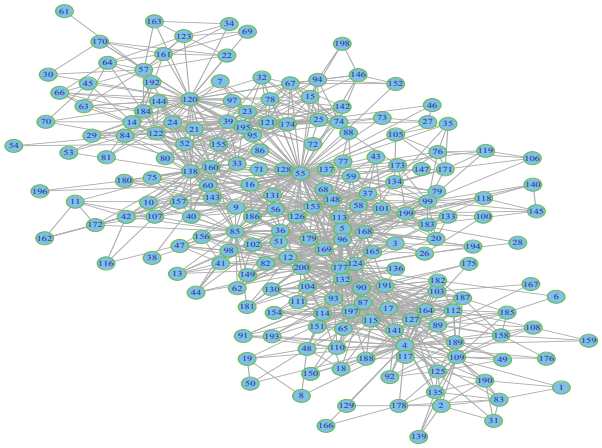

Figure 5. Estimated gene network (conditional correlation structure) for the covariates in the real data.

Our proposed method is based on a greedy search algorithm which searches a huge space of the potential graphs. This algorithm can be significantly improved if parallel computing is taken (step 2 of Algorithm 1). Similar discussion can be found in Jones et al. (2005).

To improve the efficiency of Algorithm 1, we may also focus on the most saturated graphs. This is because from our simulation studies, we found that when increasing the number of edges, the sum of square errors will decrease (Figure 1). Although this is quite reasonable, it is non-trivial to prove that when the number of edges in the graph is increased, the GLSE will provide a more accurate predicted value for the response. We also leave this to future work.

References

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 1705–1732. http://dx.doi.org/10.1214/08-AOS620.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351. http://dx.doi.org/10.1214/009053606000001523

Carvalho, C. M., & West, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis, 2*, 69–97. http://dx.doi.org/10.1214/07-BA204

Dai H. (2008). Perfect sampling methods for random forests. *Advances in Applied Probability, 40*, 897-917. http://dx.doi.org/10.1239/aap/1222868191

Dawid A. P., & Lauritzen S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics, 21*, 1272–1317. http://dx.doi.org/10.1214/aos/1176349260

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis, 90*, 196–212. http://dx.doi.org/10.1016/j.jmva.2004.02.009

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*, 407-499. http://dx.doi.org/10.1214/009053604000000067

Goeman, J., Meijer, R., & Chaturvedi, N. (2012). L1 and L2 penalized regression models [R package]. *cran.r-project.or*.

Golumbic, M. C. (2004). *Algorithmic graph theory and perfect graphs*. Elsevier.

Hoerl, A.E., & Kennard R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 5567. http://dx.doi.org/10.1080/00401706.1970.10488634

Huang, J., Ma, S., & Zhang, C. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Test, 18*, 270–275.

James, G. M., & Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika Trust,*

96, 323–337. http://dx.doi.org/10.1093/biomet/asp013

Jones B., Carvalho C., Dobra A., Hans C., Carter C., & West M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science, 4*, 388-400.

Krämer, N., Schäfer, J., & Boulesteix, A. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics, 10*, 384. http://dx.doi.org/10.1186/1471-2105-10-384

Lauritzen, S. (1996). *Graphical Models*. Oxford Univeristy Press.

Lin, W., Shi, P., Feng, R., & Li, H. (2014). Variable selection in regression with compositional covariates. *Statistical Science*, 388-400.

Markowetz, F., & Spang, R. (2007). Inferring cellular networks – a review. *BMC Bioinformatics, 8*, S5. http://dx.doi.org/10.1186/1471-2105-8-S6-S5

Meinshausen, N., & Yu, B. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 246–270. http://dx.doi.org/10.1214/07-AOS582

McCann, L., & Welsch, R. E. (2007) Robust variable selection using least angle regression and elemental set sampling. *Biometrika*, 1–13. http://dx.doi.org/10.1016/j.csda.2007.01.012

Scheetz, T. E., Kim, K. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L. Sheffield, V. C., & Stone, E. M. (2006) Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences, 39*, 14429–14434. http://dx.doi.org/10.1073/pnas.0602562103

Sun, T., & Zhang, C. (2010) Comments on: l 1-penalization for mixture regression models. *Biometrika, 19*, 270–275. http://dx.doi.org/10.1007/s11749-010-0201-7

Talluri, R., & Shete, S. (2014) Gaussian graphical models for phenotypes using pedigree data and exploratory analysis using networks with genetic and nongenetic factors based on Genetic Analysis Workshop 18 data.. *BMC Proceedings, 8*, S99. http://dx.doi.org/10.1186/1753-6561-8-S1-S99

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, B, 33*, 267–288.

Van, H. H., & Putter, H. (2011). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, 183.

Washington, S. P., Karlaftis, M. G., & Mannering, F. L. (2010). *Statistical and econometric methods for transportation data analysis*. CRC Press, 15.

Whittaker J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

Li, X., Zhao, T., Wang, L., Yuan, X., & Liu, H. (2014). flare: Family of Lasso Regression, R package version 1.5.0 [R package] *http://CRAN.R-project.org/package=flare*.

Yuan, M., Ekici, A., Lu, Z., & Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*, 329–346. http://dx.doi.org/10.1111/j.1467-9868.2007.00591.x

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 894–942. http://dx.doi.org/10.1214/09-AOS729

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x