Hybrid Graphical Least Square Estimation and its application in Portfolio Selection

Saeed Aldahmani[1], Hongsheng Dai [2] and Qiaozhen Zhang[3]

[1]Department of Statistics, College of Business and Economics, United Arab Emirates University, UAE
[2]Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK
[3]Institute of Statistics, Nankai University, China

May 22, 2019

Abstract

This paper proposes a new regression method based on the idea of graphical models to deal with regression problems with the number of covariates v larger than the sample size N. Unlike the regularization methods such as ridge regression, LASSO and LARS, which always give biased estimates for all parameters, the proposed method can give unbiased estimates for important parameters (a certain subset of all parameters). The new method is applied to a portfolio selection problem under the linear regression framework and, compared to other existing methods, it can assist in improving the portfolio performance by increasing its expected return and decreasing its risk. Another advantage of the proposed method is that it constructs a non-sparse (saturated) portfolio, which is more diversified in terms of stocks and reduces the stock-specific risk. Overall, four simulation studies and a real data analysis from London Stock Exchange showed that our method outperforms other existing regression methods when $N < v$.

Keywords: Graphical Model; Graphical Least Squares; LASSO; Ridge Regression; Unbiased Estimation.

1 Introduction

1.1 Portfolio selection and its relation with linear regression

In finance, a portfolio is considered as a collection of two or more risk (or risk-free) assets such as shares, government bonds and derivative securities which are held directly by investors or managed by a financial institution (Ennew et al., 2014). Investors or financial institution managers seek to efficiently allocate and diversify their capital over a number of available assets by creating a portfolio that leads to maximizing expected returns on the investment and minimizing the associated risk. Such an efficient allocation of capital among different assets could be made by a portfolio optimization problem. This problem was solved by Markowitz (1952) using a model known as the Mean-Variance model or Markowitz theory. For a portfolio consisting of v assets, with expected return vector $\boldsymbol{\mu}$ (row vector) and covariance matrix $\boldsymbol{\Sigma}$, the Markowitz model selects an efficient (optimal) portfolio weight vector $\boldsymbol{w}$ (column vector), which minimizes the risk (volatility) of the portfolio for a predetermined targeted expected return. In other words, it is a trade-off between two factors, the risk and the return of the portfolio. Here $\boldsymbol{w}$ can be interpreted as the proportion of capital invested in each asset in the portfolio. The construction of such an efficient (optimal) portfolio for a given targeted expected return R is done through finding the optimal portfolio weights $\boldsymbol{w}$ by solving the following quadratic optimization problem (Ledoit and Wolf, 2003):

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \boldsymbol{w}^{'}\boldsymbol{\Sigma}\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{\mu}\boldsymbol{w} = R, \quad \boldsymbol{w}^{'}\mathbf{1}_v = 1 \tag{1}$$

where $\mathbf{1}_v$ denotes a $v \times 1$ vector of ones. The set of all efficient portfolios lies on a risk-return relationship curve called efficient frontier (see Figure 1) where every point on the curve offers the highest expected return for a given level of risk or the lowest risk for a given level of expected return (Lee and Lee, 2010).

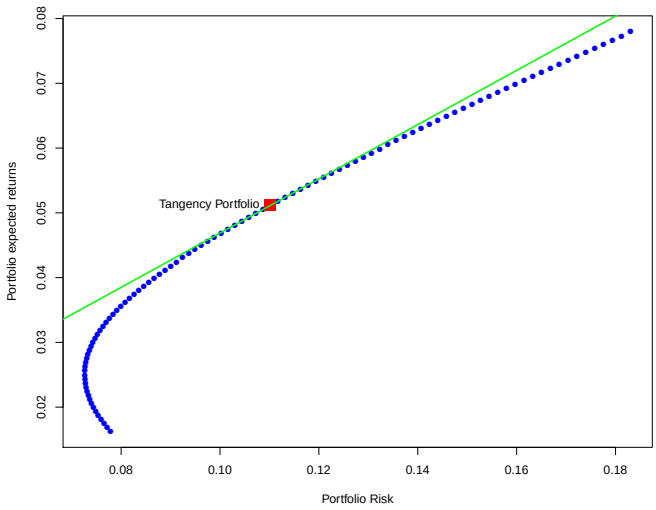


Figure 1: Graph of efficient portfolios curve and tangency portfolio.

Investors, however, usually prefer to use the Sharpe ratio (the ratio of return to the standard deviation of the return) to evaluate a portfolio performance (Lai et al., 2011). The combination of the v risk assets which gives the maximum Sharpe ratio among portfolios on the efficient frontier is called tangency portfolio $\boldsymbol{w}^{\star}$ (Gilli et al., 2011) and graphically represented as a point where a line through the origin (or any other point in the vertical axis if the portfolio includes risk-free assets) is tangent to the curve of efficient frontier (see Figure 1).

Markowitz theory requires that the population mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of the assets' returns be known. However, in practice, these two parameters are unknown and should be estimated using historical data set. Britten-Jones (1999) shows that the tangency portfolio (the optimal solution) based on the historical data is given as:

$$\boldsymbol{w}^{\star} = \frac{\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\beta}}' \mathbf{1}_v} \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares (OLS) estimate of the coefficient parameter $\boldsymbol{\beta}$ for the

linear regression model

$$\boldsymbol{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{y}$ is a column vector of 1s with length N (sample size), $\mathbf{x}$ is the observed $N \times v$ dimensional asset return matrix and ϵ is an N dimensional column vector of the residuals.

1.2 Regression with $N < v$ and its challenges

In the regression model for portfolio selection, the total number of observations N (usually about several hundreds, about 10 to 20 years monthly data points) is usually much less than the total number of assets v (usually thousands), i.e. $\hat{\boldsymbol{\beta}}$ is not available based on standard regression methods, which is the main challenge for such a modern portfolio optimization problem.

Therefore, in order to overcome the issue for model (3) when $N < v$, many recent studies (DeMiguel et al., 2009; Still and Kondor, 2010; Carrasco and Noumon, 2011; Fastrich et al., 2015; Long et al., 2018; Norouzirad et al., 2018) focus on regularization methods such as ridge regression (Hoerl and Kennard, 1970), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 2011), Least Angle Regression (LARS) (Efron et al., 2004), Adaptive LASSO (Zou, 2006) and the Dantzig selector (Candes and Tao, 2007). However, all of their estimates are biased giving emphasis to the famous notion of "bias-variance tradeoff", which might over-shrink the coefficients (Radchenko et al., 2011) and perhaps produce inaccurate portfolio weights. Some of these methods, LASSO, LARS and Sure independence screening for example, also suffer from the problem of not selecting more than N covariates (Zou and Hastie, 2005; Hastie et al., 2009; Fan and Lv, 2008) and giving a sparse portfolio (Fastrich et al., 2015), which is less preferable for risk management and diversification (Li, 2014).

1.3 The new methodology and paper structure

This study considers using an unbiased estimation method via graphical models for solving the linear regression problem when $N < v$. This estimation is named as Graphical Least Squares estimation (GLSE) (Aldahmani and Dai, 2015). A potential weakness of the GLSE is that, for a very large number of covariates/assets, the involved computation cost will be very heavy. To overcome this computational issue, we combine the idea of ridge regression and GLSE and further propose a methodology named as the hybrid GLSE (HGLSE). This method imposes a ridge type penalty on some covariates while ensuring that the other covariate coefficient estimates are still unbiased. Therefore, this HGLSE incorporates the advantages of both GLSE (unbiased) and ridge regression (computationally simple). The HGLSE can give unbiased coefficient estimates for the most important assets (assets with high return and low risk) and maintain them in the portfolio with large weights, while penalizing only the weights of the other less important assets. Such an advantage will lead to increasing the Sharpe ratio and the expected rate of returns and decreasing the risk of the portfolio for both in-and-out-of-sample periods. It can also generate diversified portfolios (like the ridge method) across a large number of stocks, as it produces a non-sparse portfolio (including all stocks in the market). The diversified portfolio risk is very important (Malkiel and Xu, 1997), because if one or more sectors of the economy decline, the other sectors could help in reducing the significant loss due to market fluctuations.

To empirically investigate the proposed method, the HGLSE is implemented on real historical London Stock Exchange data. Four different sizes of portfolios are constructed and their in-and-out-of-sample performances are tracked. It is shown that the HGLSE outperforms the ridge method in that the HGLSE constructs portfolios with much higher Sharpe ratio. Moreover, all portfolios constructed by the HGLSE method achieve a lower risk than the ones produced by the ridge method for both in-and-out-sample periods. Clearly, this is in the interest of the investor, who usually seeks a low-risk investment.

The rest of the paper is organized as follows. Section 2 provides necessary notations, defini-

tions of graphical models and the basic idea of the graphical least squares estimation. Section 3 provides the general form of HGLSE and the bias and variance of the proposed estimator. Section 4 provides the iteration algorithm of graph structure selection and the computational complexity of HGLSE. Section 5 presents four simulation scenarios for assessing the proposed method, while the analysis on real data examples is given in Section 6. A brief discussion is provided in Section 7.

2 Preliminaries

2.1 Notation

As the main methodology used in this paper is based on graphical models, we here follow the notations in Lauritzen (1996).

An *undirected graph* G is formed of two sets, a set V and a set $\mathcal{E}$. The set $V = \{1, 2, \cdots, v\}$ denotes the vertices representing the covariate variables in the regression model (assets in the portfolio) and $\mathcal{E}$ is the set of edges (a subset of $V \times V$) connecting the vertices. An edge between vertex i and j is usually denoted as $\{i, j\}$. We only consider undirected edges in this papers, since identifying causal relations (direct edges) between vertices is not related to this work. A path from vertex i to j is a sequence of vertices $i_1, \cdots, i_n$ such that $\{i_a, i_{a+1}\} \in \mathcal{E}$, for all a.

We say the graph G is complete, if all the pairs of vertices in V are joined by an edge. A subset A of V, together with edges in G whose endpoints are both in A, induces a subgraph, denoted by G_A. A subset A is complete if G_A is a complete subgraph. A complete subset that is maximal (with respect to $\subseteq$) is called a *clique*. Disjoint subsets (A, B, C) of V in an undirected graph G form a *decomposition* of G if $V = A \cup B \cup C$ provided that: B separates A from C (all paths from vertices in A to vertices in C intersect B) and B is a complete subset of V (definition of weak decomposition in Lauritzen (1996)).

If an undirected graph G is complete and/or there is a proper decomposition (A, B, C) into decomposable subgraphs $G_{A\cup B}$ and $G_{B\cup C}$ then G is called a *decomposable* graph.

For a sequence of sets $C_1, ..., C_q \subset V$, define

$$H_i = C_1 \cup ... \cup C_i; \qquad \text{and} \qquad S_i = H_{i-1} \cap C_i.$$

The sequence of sets $C_1, ..., C_q$ is called a *perfect sequence* if the following conditions hold (Lauritzen, 1996):

1. For all $j > 1$, there is an $i < j$, such that $S_j \subseteq C_i$;
2. The sets S_j are complete for all values of j;

Perfect ordering of cliques is obeyed in a decomposable graph G (Golumbic, 2004).

For any index set A, we usually use the corresponding lower letter a or $|A|$ to denote the number of elements in A.

In later sections, we also need the following matrix notations. A $v \times v$ matrix $\mathbf{z}$ can be written as $(z_{kj})_{k,j\in V}$. For $A \subset V, B \subset V$, 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

$$\left([\mathbf{z}_{AB}]^{\Gamma}\right)_{jk} = \begin{cases} z_{jk} & \text{if } j \in A,\ k \in B \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We can define $\boldsymbol{Z}_A$ and $[\boldsymbol{Z}_A]^{\Gamma}$ similarly for a vector $\boldsymbol{Z} = (z_1, \cdots, z_v)$.

Now we consider the observed covariate (assets return) matrix $\mathbf{x}$. Let $\mathbf{x}_A$ be the covariate matrix only having variables with indices in set A and $ssd_A = \mathbf{x}'_A\mathbf{x}_A$. Then $[(ssd_A)^{-1}]^{\Gamma}$ represents a $v \times v$-dimensional matrix obtained by filling up 0s, with

$$\left([(ssd_A)^{-1}]^{\Gamma}\right)_{jk} = \begin{cases} \left((ssd_A)^{-1}\right)_{jk} & \text{if } j, k \in A \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The notation $[\mathbf{x}_A]^\Gamma$ means expanding the $N \times |A|$-dimensional matrix $\mathbf{x}_A$ to a $N \times V$ dimensional matrix.

$$\left([\mathbf{x}_A]^\Gamma\right)_{nk} = \begin{cases} x_{nk} & \text{if } k \in A \\ 0 & \text{for } k \notin A. \end{cases} \tag{6}$$

We denote $[\mathbf{x}'_A]^\Gamma := \left([\mathbf{x}_A]^\Gamma\right)'$. Clearly we have $[ssd_A]^\Gamma = [\mathbf{x}'_A]^\Gamma \cdot [\mathbf{x}_A]^\Gamma$.

2.2 Basic idea

This section presents the idea of the GLSE method and demonstrates its unbiased property under certain conditions.

Suppose that a decomposable graph g is given (known) which consists of several cliques $\mathcal{C} = \{C_1, \cdots, C_q\}$ and separators $\mathcal{S} = \{S_2, \cdots, S_q\}$. We can define an estimator for (3), associated with this graph g, as

$$\hat{\boldsymbol{\beta}}^g = \left[\sum_{C \in \mathcal{C}} [(ssd_C)^{-1}]^\Gamma - \sum_{S \in \mathcal{S}} [(ssd_S)^{-1}]^\Gamma\right] \mathbf{x}'\boldsymbol{y} \tag{7}$$

which is called GLSE estimator for $\boldsymbol{\beta}$. For the existence of the matrix inversions in the above formula, the following condition must hold.

Condition 2.1. *The sample size* $N > \max_{C \in \mathcal{C}}\{|C|\}$.

Then we can show the unbiasedness property of the estimator in (7) under the following Condition 2.2, which is presented in Theorem 2.1.

Condition 2.2. *The graph g is decomposable with a perfect ordering of cliques* $(C_1, \cdots, C_q)$ *and separators* $(S_2, \cdots, S_q)$*, such that*

(a) the following linear relationship holds,

$$\begin{aligned} \mathbf{x}_{C_1 \setminus S_2} &= \mathbf{x}_{S_2} \cdot \mathbf{r}_{S_2, C_1 \setminus S_2} + \boldsymbol{\xi}_1, \quad E(\boldsymbol{\xi}_1) = \mathbf{0}, \\ \mathbf{x}_{C_k \setminus S_k} &= \mathbf{x}_{S_k} \cdot \mathbf{r}_{S_k, C_k \setminus S_k} + \boldsymbol{\xi}_k, \quad E(\boldsymbol{\xi}_k) = \mathbf{0}, \quad k = 2, \cdots, q, \end{aligned} \tag{8}$$

where $\mathbf{r}_{S_k, C_k \setminus S_k}$ are constant matrices with dimensions $|s_k \times (c_k - s_k)|$;

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

$$(\boldsymbol{\xi}_k, \cdots, \boldsymbol{\xi}_q) \perp (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{k-1}) | \mathbf{x}_{S_k}. \tag{9}$$

Theorem 2.1. *Under Condition 2.2, the estimator in (7) is unbiased, i.e. $\mathbb{E}(\hat{\boldsymbol{\beta}}^g) = \boldsymbol{\beta}$.*

Proof. The proof of theorem 2.1 is given in Appendix A. □

Remark 1. Note that, equation (8) represents the partial correlation among the variables $\mathbf{x}$. For example, the first equation in (8) means that $\mathbf{x}_{\mathrm{C}_1 \setminus \mathrm{S}_2}$ are uncorrelated with $\mathbf{x}_{V \setminus \mathrm{C}_1}$, given $\mathbf{x}_{\mathrm{S}_2}$. Such partial correlation can be represented by the concentration matrix for multivariate normal distributions (Lauritzen, 1996). However, for non-Gaussian distributions, equation (8) is easier to use.

It can be noticed that the linear assumption, in Condition 2.2, does not limit the extent to which the GLSE can be applied, as variable transformation can be used to yield a linear relationship, if the variables are quantitative. In addition, since any non-linear relation can be approximated via a polynomial, the nonlinear dependence on $\boldsymbol{x}$ can be viewed as linear dependence on $\boldsymbol{x}, \boldsymbol{x}^2, \cdots$, (Laursen and Thorlund, 2010).

The GLSE method can give unbiased parameter estimates, which are often preferable (Raol et al., 2004). However, practically, the graph structure g is unknown and should be estimated based on the data set. As a consequence, the whole graph space has to be searched in order to find the graph that is closest to the true graph. In other words, with v covariates there are

$2^{v(v-1)/2}$ different graphs that should be searched for. This means that with large value of v, the searching process will incur a heavy computational cost. To overcome this challenge, the GLSE is extended to take a hybridized form, which will be discussed in the following section.

3 The hybridized GLSE (HGLSE)

3.1 HGLSE with two cliques

Suppose that the vertex set V can be partitioned into disjoint sets A, B and C, where variables in set A are important variables but variables in set B and C are less important. Also, assume that the covariance matrix $\mathbf{x}$ is factorised according to a given decomposable graph g of two cliques ($A \cup B$ and $B \cup C$) and a separator (B). Then the proposed estimator for $\boldsymbol{\beta}$ can be as follows

$$\hat{\boldsymbol{\beta}}^h = \left[\left\{ (ssd_{A\cup B})^{-1} \right\}^{\Gamma} + \left\{ (ssd_{B\cup C} + \lambda \mathbf{I}_{B\cup C})^{-1} \right\}^{\Gamma} - \left\{ (ssd_B)^{-1} \right\}^{\Gamma} \right] \mathbf{x}'\boldsymbol{y}, \tag{10}$$

where λ is the amount of penalty imposed on the variables with index set $B \cup C$. For the above estimator to exist, the following condition should be fulfilled,

Condition 3.1. *The sample size* $N > \{|A| + |B|\}$.

If we rewrite the covariate matrix $\mathbf{x}$ as $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)$ and the parameter $\boldsymbol{\beta}$ and its estimate $\hat{\boldsymbol{\beta}}^h$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}'_A, \boldsymbol{\beta}'_B, \boldsymbol{\beta}'_C)'$ and $\hat{\boldsymbol{\beta}}^h = \left(\left(\hat{\boldsymbol{\beta}}^h_A \right)', \left(\hat{\boldsymbol{\beta}}^h_B \right)', \left(\hat{\boldsymbol{\beta}}^h_C \right)' \right)'$, respectively, then based on the following Condition 3.2, the estimator $\hat{\boldsymbol{\beta}}^h_A$ in (10) can be shown to be unbiased for $\boldsymbol{\beta}_A$, while $\hat{\boldsymbol{\beta}}^h_B$ and $\hat{\boldsymbol{\beta}}^h_C$ are biased for $\boldsymbol{\beta}_B$ and $\boldsymbol{\beta}_C$, respectively.

Condition 3.2. *The sets A, B and C make a decomposition where B is the separator. The sets* $\mathbf{x}_A$ *and* $\mathbf{x}_C$ *are conditionally independent given* $\mathbf{x}_B$*, and* $\mathbf{x}_C$ *is such that*

$$\mathbf{x}_C = \mathbf{x}_B \cdot \mathbf{r}_{B,C} + \boldsymbol{\xi}_C, \;\; \mathbb{E}(\boldsymbol{\xi}_C) = \mathbf{0},$$

where $\mathbf{r}_{B,C}$ *is an* $a \times b$ *dimensional constant matrix and* $b = |B|$, $c = |C|$.

To summarize the above arguments, the following theorem is introduced.

Theorem 3.1. *Under Condition 3.2, the estimator in (10) is unbiased for the variables in set* A *while biased for those variables that are in set* $B \cup C$,

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_A^h) = \boldsymbol{\beta}_A,\ \mathbb{E}(\hat{\boldsymbol{\beta}}_B^h) \neq \boldsymbol{\beta}_{\boldsymbol{B}},\ \mathbb{E}(\hat{\boldsymbol{\beta}}_C^h) \neq \boldsymbol{\beta}_{\boldsymbol{C}}.$$

Proof. The proof of the theorem is given in Appendix B. □

3.2 The general form of HGLSE

We assume that the graph g has a set of cliques $\mathcal{C} = \{C_1, \cdots, C_q\}$ (a perfect ordering) with separators $\mathcal{S} = \{S_2, \cdots, S_q\}$. Consider that $C^* \in \mathcal{C}$, which can be a very large clique but the variables in it are less important than those not in it. Denote $\mathcal{C}' = \mathcal{C} \backslash \{C^*\}$ as the set of cliques excluding C^*. The general HGLSE is given by

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

For the HGLSE to exist, the following condition must be met.

Condition 3.3. *The sample size* $N > \max_{C \in \mathcal{C}'}\{|C|\}$.

The we have the following result

Proposition 3.1. *Given that Condition 2.2 and Condition 3.3 hold true, the estimator in (11) is unbiased for all variables* $V \backslash C^*$ *and biased for those in* C^*.

Proof. Following Theorem 2.1 and Theorem 3.1, Proposition 3.1 is easily established. □

3.3 Bias and variance of HGLSE

As shown above, unbiased coefficient estimates are obtained for important variables, while for less important variables the estimates are biased. This follows the fact that the proposed HGLSE is partially unbiased under the given conditions. In order to estimate the variance of the unbiased estimates, the bootstrap method is used where samples of size N are repeatedly drawn from the observed sample using simple random sampling with replacement (Efron and Tibshirani, 1994). However, care should be taken to avoid the problem of multi-collinearity within a clique (except the clique C^*). For the less important covariates, the variance estimates are not desirable in that the variance is not very meaningful for biased estimates (Goeman et al., 2012).

4 Graph structure selection

Similar to the application of the GLSE method in Section 2.2, the unknown graph g needs to be estimated in order to apply the HGLSE method. A new iteration algorithm is proposed to search the space and select the graph that best fits the data under a regression framework, and to find the HGLSE.

The algorithm has two key stages in each iteration: 1. finding the best graph g^*, its associated estimate $\hat{\boldsymbol{b}}^g$ and the variances of $\hat{\boldsymbol{b}}^g$ based on given C^*, the set of less important variables; 2. selecting the less important variables C^* based on $\hat{\boldsymbol{b}}^g$ and its variance. The algorithm can start with a guess of C^*. A ridge penalty is then applied on this clique to make the inversion of the aforementioned clique possible.

In the first stage, the algorithm searches for the best graph structure g^* with the condition that the clique C^* is fixed, by minimizing the target function $\mathbb{T}(\hat{\boldsymbol{b}}^g, g)$ (sum of square errors)

given as follows:

$$\begin{aligned} g^* &= \arg\min_{g \in \mathcal{G}_{C^*}} \mathbb{T}(\hat{\boldsymbol{b}}^g, g), \\ \mathbb{T}(\hat{\boldsymbol{b}}^g, g) &= ||\boldsymbol{y} - \mathbf{x}\hat{\boldsymbol{b}}^g||^2 \end{aligned} \tag{12}$$

where $\hat{\boldsymbol{b}}^g$ is given in (11) and $\mathcal{G}_{C^*}$ denotes the space of all decomposable graphs with the clique C^* and satisfying Condition 3.3. Then bootstrapping is performed to calculate the variance and the significance level of the unbiased estimates for the important variables (those not in C^*).

In the second stage, we update C^*. Denote the variables (for those unbiased estimates only) above a certain significance threshold level, as $\mathcal{D} = \{D_1, \cdots, D_t\}$ where t denotes the total number of iterations . We then apply LASSO to the saturated regression model with variables $\mathbf{x}_{\mathcal{D}}$ fixed. The variables with coefficient estimates which are shrunk to 0 by LASSO are treated as less important variables and form the new clique C^*.

This process is repeated until convergence where the method selects the same sets of important variables twice. The algorithm is given as follows.

Algorithm 1 Pseudo-code of the hybridized iteration algorithm

1: **Initialization**: Select covariates C^* that are less important and $V \backslash C^*$ as the set for the important ones.

Iteration – repeated until the convergence with respect to the estimate $\hat{\boldsymbol{b}}^g$

2: Search the graph space and find the graph g^* and $\hat{\boldsymbol{b}}^g$ such that they minimize the target function $\mathbb{T}(.)$ (given in (12)), on the condition that C^* is a clique of g^*.

3: Based on g^*, do bootstrapping to find the variance and significance levels for the variables in $V \backslash C^*$.

4: Select variables in $V \backslash C^*$, which are above a certain significance threshold level (chosen as 0.5 in this paper). Denote the set of the selected variables as $\mathcal{D}$.

5: Consider the regression model with all covariates V and use LASSO, with the variables in $\mathcal{D}$ fixed, to do variable selection for all other variables in $V \backslash \mathcal{D}$. The non-selected variables by LASSO form the new C^* in the next iteration.

6: Go back to step 2.

In Algorithm 1, the second step can be parallelized to work if several central processing units are available in the computational environment. The computational cost can therefore be improved significantly via parallel computation.

5 Simulation

This section presents four simulation studies for the assessment of the proposed HGLSE in constructing an optimal portfolio. Scenario 1 is based on a case of normally distributed covariates, provided that the C^* is known and the response $\boldsymbol{y}$ is not a vector of 1s. Because we suppose C^* is known, we do not need the full Algorithm 1, but step 2 (finding the estimate and graph) and step 3 (finding the significant level and standard deviation). This toy simulation scenario is used to justify how good the standard deviation (or confidence intervals) of the estimate, given by the bootstrap method, is.

Scenario 2 is based on multivariate normally distributed predictor variables, where the important variables are supposed to be unknown and the response $\boldsymbol{y}$ is a vector of 1s, following the Markowitz model. Algorithm 1 is therefore applied.

Scenario 3 is similar to Scenario 2; however, the covariates do not follow multivariate normal distribution.

Likewise, Scenario 4 is also similar to Scenario 2, except for the fact that the data are generated from t-distribution.

In the four scenarios, we compare HGLSE with ridge regression. We did not compare with LASSO or LARS because these methods only provide sparse portfolios (not keeping all covariates in the model), which is not preferable in practice (the diversified portfolio risk via a saturated model is very important (Malkiel and Xu, 1997)).

5.1 Scenario 1

In this scenario, the model (3) is used to generate the response y where the random errors are normally distributed with mean 0 and standard error $\sigma = 1$. A total of $v = 20$ predictor variables, which follow a multivariate normal distribution, are considered and a total of $N = 15$ samples are generated. The true $\boldsymbol{\beta}$ values are given in Table 1, where five predictor variables are important, X_1, X_5, X_{10}, X_{15} and X_{20}, having large regression parameter values. The true graph used in this scenario to generate the data is given in Figure 2, where the partial correlations for the important variables are shown on the corresponding edges.

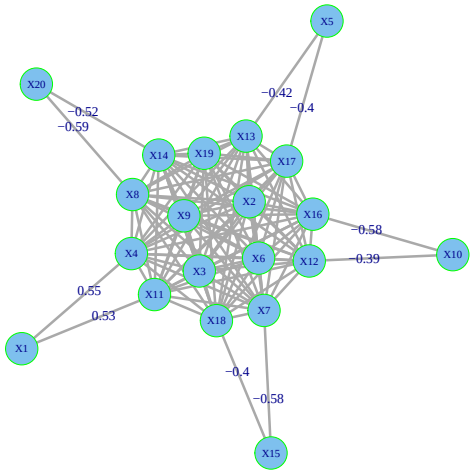


Figure 2: Graph structure for covariates under Scenario 1.

Table 1: Results from the 500 simulated data in Scenario 1.

		Ridge		HGLSE			
	$\boldsymbol{\beta}$	$Bias$[a]	SD[b]	$Bias$[a]	SD[b]	SE [c]	CP
X_1	0.9	-0.408	0.351	0.038	0.289	0.572	0.974
X_2	0.1	-0.056	0.567	-0.033	0.358	1.142	
X_3	0.05	-0.053	0.742	-0.044	0.359	0.961	
X_4	0.1	0.140	0.541	-0.100	0.258	0.400	
X_5	1.5	-0.836	0.418	0.016	0.300	0.586	0.976
X_6	0.1	-0.067	0.674	-0.105	0.162	0.265	
X_7	-0.02	-0.386	0.527	-0.040	0.045	0.041	
X_8	-0.05	-0.253	0.513	0.019	0.065	0.045	
X_9	0.01	-0.061	0.623	-0.017	0.038	0.033	
X_{10}	1.3	-0.820	0.337	0.066	0.261	0.468	0.966
X_{11}	0.1	0.146	0.625	-0.062	0.040	0.034	
X_{12}	-0.01	-0.086	0.444	-0.005	0.050	0.044	
X_{13}	-0.05	-0.062	0.438	0.026	0.055	0.045	
X_{14}	-0.03	-0.060	0.305	0.006	0.070	0.056	
X_{15}	2	-1.001	0.537	0.056	0.370	0.570	0.966
X_{16}	-0.01	-0.105	0.239	-0.034	0.084	0.070	
X_{17}	0.1	-0.141	0.507	-0.110	0.052	0.042	
X_{18}	-0.01	-0.199	0.530	-0.029	0.052	0.042	
X_{19}	0.1	0.196	0.274	0.016	0.079	0.068	
X_{20}	1.2	-0.706	0.351	0.075	0.297	0.536	0.960

[a] Bias$(\hat{\boldsymbol{\beta}}) = [\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]$.

[b] SD: the Monte Carlo standard error for the 500 replicates.

[c] SE: the mean of the 500 standard error estimates; each estimate is based on 500 bootstrap samples.

[d] CP: Coverage Probability for 95% CI.

The bias values in Table 1 indicate that the HGLSE gives better results than the ridge method in estimating the coefficients for both important and less important variables. The severely biased ridge estimates for these important variables could reduce the overall return on the portfolio. The Monte Carlo estimates of standard errors for the HGLSE are smaller than those of the ridge. The mean square error (MSE) from the ridge estimate for $\boldsymbol{\beta}_1$ is $-0.408^2 + 0.351^2 = 0.290$; however, it is $0.038^2 + 0.289^2 = 0.085$ for the HGLSE, and this holds true to all important and most less important variable estimators. Therefore, the

performance of the HGLSE is superior to that of the ridge. Additionally, for the GLSE method, the mean of the 500 bootstrap standard error estimates (SE^c) is in general slightly larger than the Monte Carlo standard deviations (SD^b) of the 500 replicated estimates. This is because of the small sample size. However, when the sample size is increased ($N = 80$), the bootstrap standard error estimates (SE^c) of the important predictor coefficients are almost similar to the Monte Carlo standard deviations (SD^b) as demonstrated in Figure 3.

In terms of coverage probability of the 95% confidence interval, the results are reasonable for the 5 important variables. This implies that bootstrap estimation for the confidence intervals of the important predictor coefficients is acceptable. We did not present the coverage probability for the biased estimates (those less important variables), since confidence interval is of no use for biased estimates.

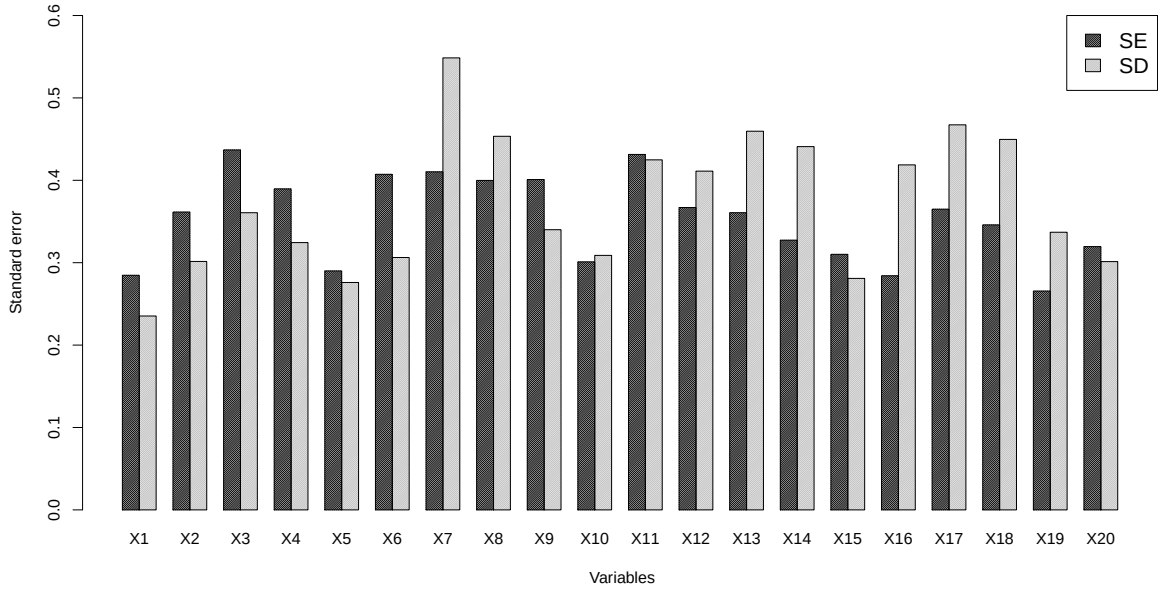


Figure 3: Monte Carlo and bootstrap standard error estimates for HGLSE with N=80.

5.2 Scenario 2

In this scenario, a total of $N = 48$ observations are generated on $v = 40$ variables from multivariate normal distribution with mean 0.01 and variance covariance matrix $\mathbf{\Sigma}$ (similar to Gilli et al. (2011)), where 36 observations (to mimic three-year monthly financial data) are used for estimating the regression coefficients and the portfolio performance (Sharpe ratios, expected returns and risk (measured by the standard deviations)) for the in-sample period, and the remaining 12 observations (to mimic one-year monthly financial data) for computing the portfolio's performance for the out-of-sample period. The true portfolio weight $\boldsymbol{w}$ is derived based on the true covariance matrix $\mathbf{\Sigma}$ (Britten-Jones, 1999). Then the regression parameters $\boldsymbol{\beta}$ are estimated on the basis that the response $\boldsymbol{y}$ is a vector of 1s and the estimated weights $\hat{\boldsymbol{w}}$ are derived from $\hat{\boldsymbol{\beta}}$ via equation (2).

The aim of this scenario is to study how the HGLSE performs in weight estimates and optimal portfolio construction. A total of 500 independent realizations are made in this simulation. The graph structure for covariates which are used in generating the data set under Scenario 2 is given in Figure 4.

Since there are too many covariates to display, we simply summarize the results in Table 2 and Figure 5. Table 2 shows the in-and-out-of-sample portfolio's Sharpe ratios, expected returns and Portfolio's risk from the 500 simulated data (in practice these characteristics are of more interests than the actual coefficient estimate). Figure 5 displays the bias of the tangent portfolio weight estimate.

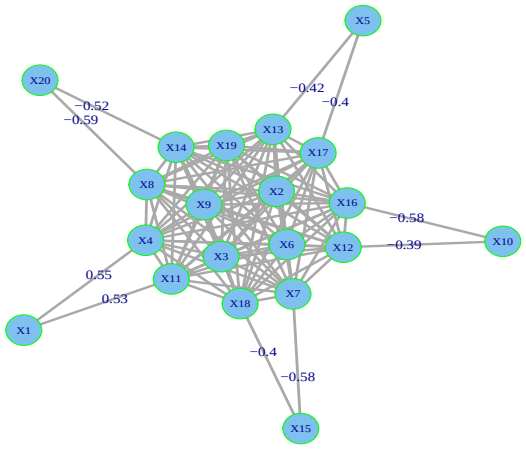
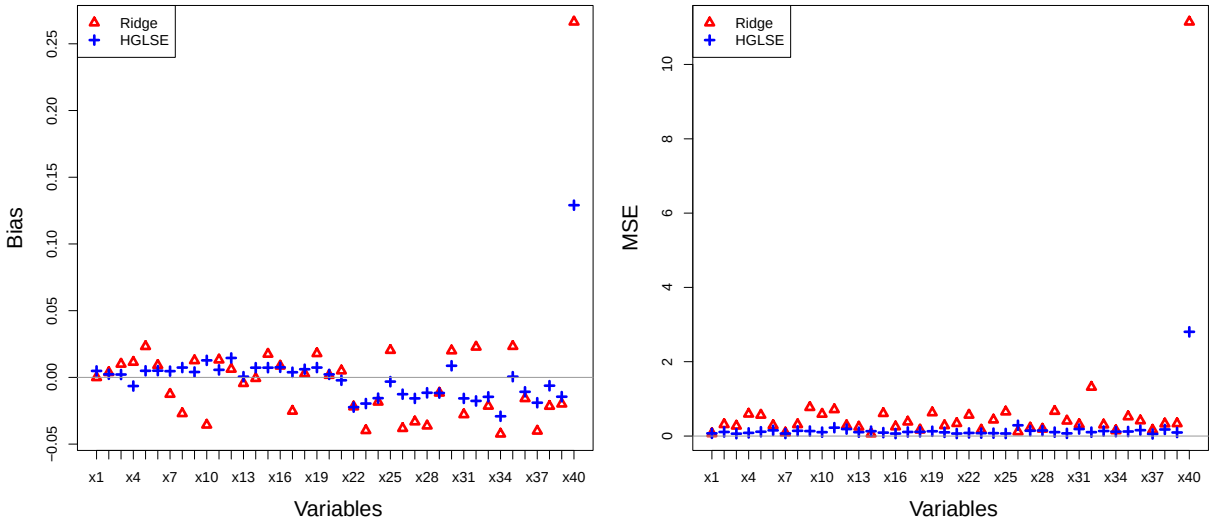


Figure 4: Graph structure for covariates under Scenarios 2.

The results given in Table 2 reveal that the HGLSE gives higher means of in-and-out-of-sample portfolio's Sharpe ratios and lower risk from the 500 simulated data. In spite of the fact that the returns of the ridge portfolio for the in-sample period are higher than the HGLSE ones, the standard deviations of the returns for the ridge are much higher than the ones for the HGLSE. This indicates that the HGLSE method is likely to yield more stable returns than the ridge can do. In addition, the out-of-sample expected returns are positive for the HGLSE but negative for the ridge, which is essential in the world of finance. On the other hand, as far as the Sharpe ratio is concerned, it can be seen from Table 2 that the standard deviations for the HGLSE method are slightly higher than those for the ridge, and this difference is still relatively small when compared to the differences in standard deviations for expected returns and error. As for the weight estimates, Figure 5 indicates that the HGLSE outperforms the ridge in estimating the weights, as the bias in most of the weights found by the HGLSE is closer to zero than that in the ridge. In addition, Figure 5 also shows that the mean square errors of HGLSE are much lower than those of the ridge, which means that HGLSE is more consistent than the ridge in estimating the weights of the portfolio.

Table 2: The means and standard deviations of in-and-out-of-sample portfolio's Sharpe ratios, expected returns and risk from the 500 simulated data (partially consistent with Ledoit and Wolf (2014)).

	Ridge		HGLSE	
	In-sample	Out of sample	In-sample	Out of sample
Sharpe ratio	0.257	0.218	0.567	0.456
Expected returns	0.176	-0.075	0.158	0.055
Portfolio's risk	1.556	1.486	0.744	0.778
	Standard deviation		Standard deviation	
	In-sample	Out of sample	In-sample	Out of sample
Sharpe ratio	0.297	0.349	0.438	0.460
Expected returns	2.500	4.139	0.571	1.005
Portfolio's risk	9.492	8.367	2.336	2.819



(*a*) The bias in estimating weights (*b*) Mean square errors

Figure 5: Comparison of the bias in estimating weights and mean square errors by ridge and HGLSE methods from data sets generated in Scenario 2.

5.3 Scenario 3

This scenario is exactly similar to Scenario 2 in all aspects except in terms of predictor variables, $\mathbf{x}$, which do not follow a multivariate normal distribution. The idea of this simulation is to find out how the method performs in more general cases. The covariates are generated

from a set of regression models with uniform random errors, which is given in detail in the supplementary file. Similar to Scenario 2, the true portfolio weight $\boldsymbol{w}$ is also derived based on the true covariance matrix $\boldsymbol{\Sigma}$, whose concentration matrix $\boldsymbol{\Sigma}^{-1}$ is given in Figure 6.

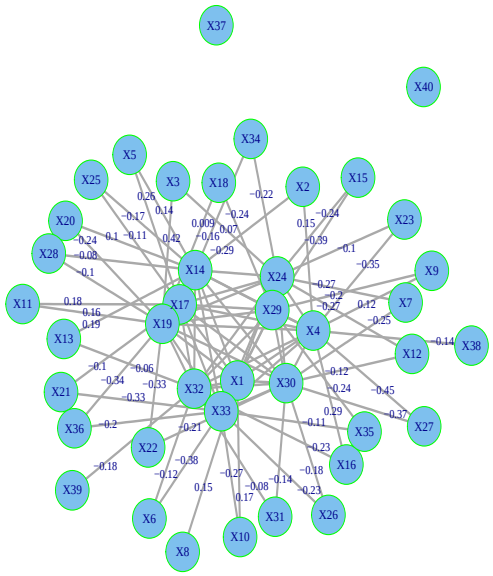


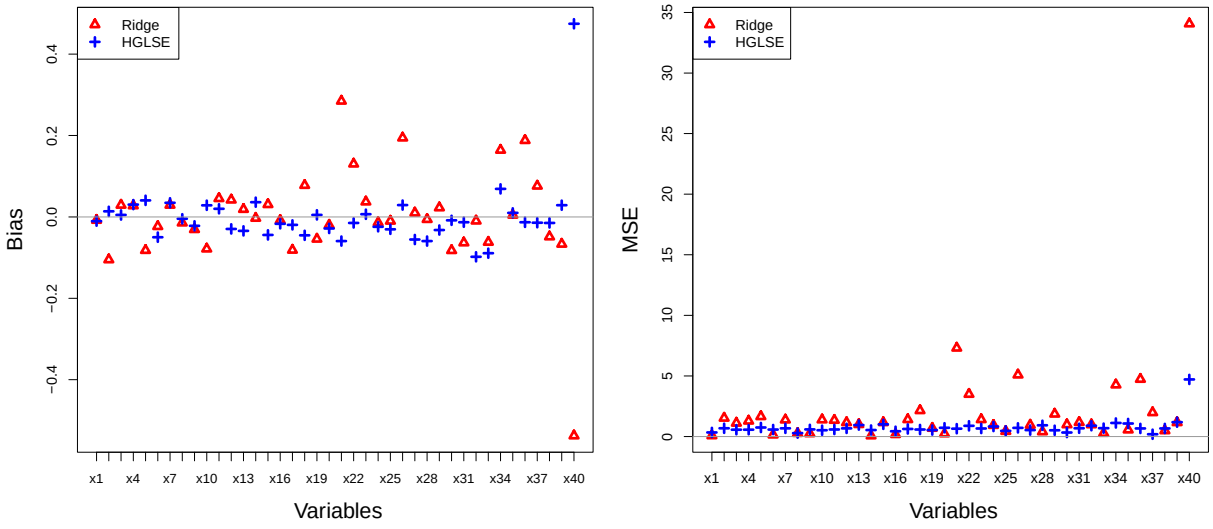
Figure 6: Graph structure for covariates under Scenario 3.

The results based on Scenario 3 are given in Table 3 and Figure 7. It can be seen from Table 3 that, out of the 500 simulated data, the HGLSE yields higher means of the portfolio's Sharpe ratio and lower risk (measured by the standard deviations) than the ridge does for the in-sample period. However, for the out-of-sample period, the ridge gives higher means of the portfolio's Sharpe ratio than the HGLSE does. It should be noted, though, that this may not be very accurate due to the fact that the mean standard deviations of the portfolio's risk and expected returns are very high compared to those under the HGLSE. In addition, the out-of-sample expected returns are negative for the ridge but positive for the HGLSE, which is desirable in finance. With regards to the weight estimates, Figure 7 shows that the HGLSE is better able to estimate the weights than the ridge can do, since the bias detected by the HGLSE in most of the weights is more stable and closer to zero than that detected by the ridge. Moreover, the mean square error of HGLSE is much lower than that of the ridge, which indicates that the performance of HGLSE is superior to that of the ridge.

This is due the reason that the proposed method can find the interactions among variables effectively, while the ridge disregards them. This implies that even if the distribution of predictor variables does not follow multivariate normal distribution, unbiased estimates for the regression coefficients can still be obtained by the HGLSE.

Table 3: The means and standard deviations of in-and-out-of-sample portfolio's Sharpe ratios, expected returns and risk from the 500 simulated data (partially consistent with Ledoit and Wolf (2014)).

	Ridge		HGLSE	
	In-sample	Out of sample	In-sample	Out of sample
Sharpe ratio	0.024	0.026	0.047	0.019
Expected returns	0.430	-1.200	0.092	0.128
Portfolio's risk	4.418	4.723	2.753	2.713
	Standard deviation		Standard deviation	
	In-sample	Out of sample	In-sample	Out of sample
Sharpe ratio	0.160	0.312	0.164	0.332
Expected returns	9.270	26.972	1.326	1.315
Portfolio's risk	59.607	67.214	7.068	7.254



(*a*) The bias in estimating weights (*b*) Mean square errors

Figure 7: Comparison of mean square errors and the bias in estimating weights by ridge and HGLSE methods from data sets generated in Scenario 3.

5.4 Scenario 4

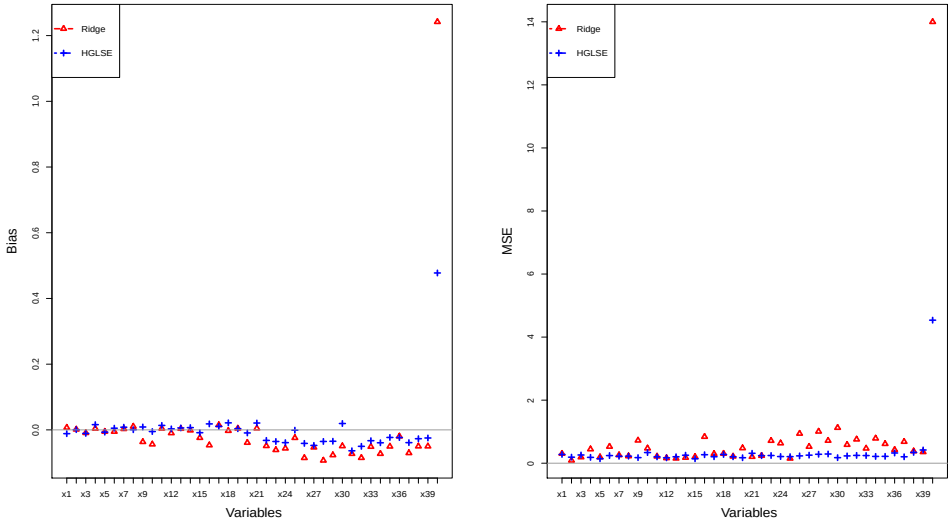
This scenario is similar to Scenario 2 in most aspects. However, the predictor variables $\mathbf{x}$ in this scenario are generated from t-distribution with a degree of freedom equal to 4. The target of this simulation is to examine how the method performs when the predictor variables are generated with tails that are fatter than those of a normal distribution.

The results based on this scenario are given in Table 4 and Figure 8. Table 4 reveals that, for the in-and-out-of-sample period, the HGLSE gives higher means of the portfolio's Sharpe ratio, expected returns and lower risk than the ridge does. Moreover, the expected returns for the in-and-out-of-sample period are negative for the ridge but positive for the HGLSE.

In terms of the weight estimates, Figure 8 indicates that the HGLSE outperforms the ridge in estimating the weights. Additionally, the mean square error of HGLSE is much lower than that of the ridge, which suggests that the performance of HGLSE is better than that of the ridge. This indicates that HGLSE can still give unbiased estimates even when the predictor variables are generated with tails that are fatter than those of normal distribution.

Table 4: The means and standard deviations of in-and-out-of-sample portfolio's Sharpe ratios, expected returns and risk from the 500 simulated data (partially consistent with Ledoit and Wolf (2014)).

	Ridge		HGLSE	
	In-sample	Out of sample	In-sample	Out of sample
Sharpe ratio	0.021	-0.014	0.074	0.019
Expected returns	-0.140	-0.067	0.146	0.095
Portfolio's risk	6.344	6.011	5.229	4.990
	Standard deviation		Standard deviation	
	In-sample	Out of sample	In-sample	Out of sample
Sharpe ratio	0.166	0.302	0.142	0.300
Expected returns	8.286	3.318	1.737	2.569
Portfolio's risk	29.054	30.403	10.171	9.258



(a) The bias in estimating weights (b) Mean square errors

Figure 8: Comparison of mean square errors and the bias in estimating weights by ridge and HGLSE methods from data sets generated in Scenario 4.

5.5 Discussion on the convergence of Algorithm 1

The above algorithm reaches convergence at the average about three iterations, as illustrated in Scenarios 2, 3 and 4 in the simulation section. We here present a typical example in

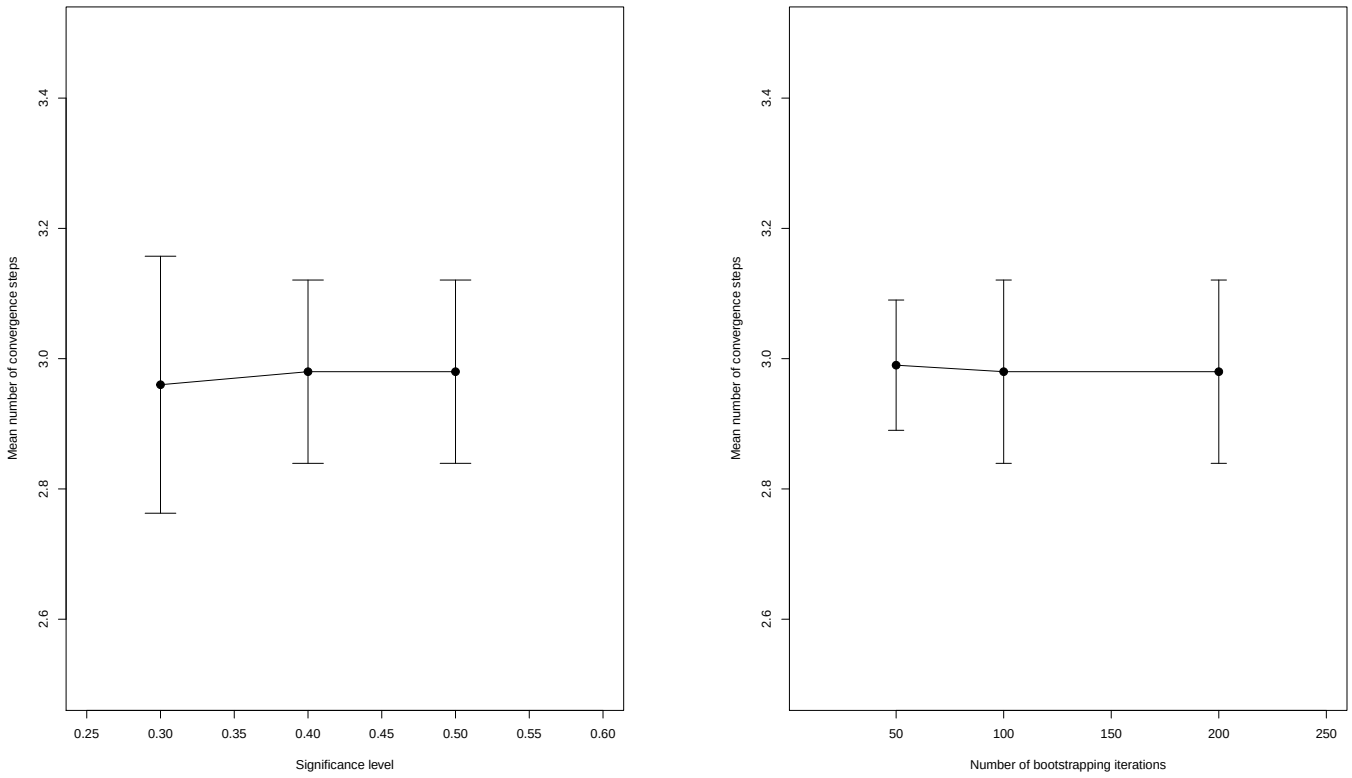
our simulation based on a single data set, where LASSO selected 11 variables in the first step. Based on this our HGLSE selected 7 variables as important variables. In the second step, LASSO fixed these 7 variables and do further selection among the rest, choosing 16 variables (including the seven fixed ones). Then we reapply HGLSE and select 12 important variables. In the third step, LASSO selects 14 variables including the 12 important variables from the previous step. Then we reapply our HGLSE and selected the same 12 variables as in the previous step. This means that the algorithm reaches convergence. Although only 12 variables will be treated as important ones, our HGLSE actually included all variables in the model, not discarding the less important ones. We distinguish important and less important variables, only because it can simplify the graphical search.

To further clarify the algorithm convergence based on the significance of threshold level and the amount of bootstrapping, Figure 9 shows the outcome of implementing a simulation scenario, which is a repetition of Scenario 2 with a significance threshold level of 0.3, 0.4 and 0.5. Moreover, Figure 9 also displays the results of different bootstrapping amounts of $50, 100$ and 200.

It can be seen that when the significance threshold level changes (using 200 bootstraps), the algorithm convergence remains the same (three on average). Although there are some variations when the significance threshold is changed (0.3 has the highest variation), these variations are marginal. Note that the important or significant variable coefficients correspond to low p-values (< 0.3), while the unimportant or nonsignificant variables have higher p-values (≥ 0.5). Therefore, when changing the significance threshold level, both important and unimportant variables are unchanged. However, when a low threshold value (higher significance value) is used, the likelihood of missing important variables will increase, which can consequently reduce the weight of those assets in the portfolio, and this in turn can increase the portfolio risk and decrease its returns.

Another justification for the convergence of the algorithm is that LASSO might have an effect on the convergence due to its characteristic of selecting a maximum of N variables. In each

iteration, LASSO first selects non-zero variables, which are then filtered by HGLSE, and some of these are chosen as important variables $\mathcal{D}$. These important variables remain fixed by Lasso until all iterations are completed. The process of fixing the important variables is repeated in each iteration until LASSO is saturated (i.e. selects the same set of variables twice). Once LASSO is saturated (in the last iteration) and no more variables can be selected by it, HGLSE selects the important variables among these fixed ones. Moreover, HGLSE might select the same important variables twice, even though LASSO may give different variables in each iteration (i.e. when LASSO is not saturated yet).

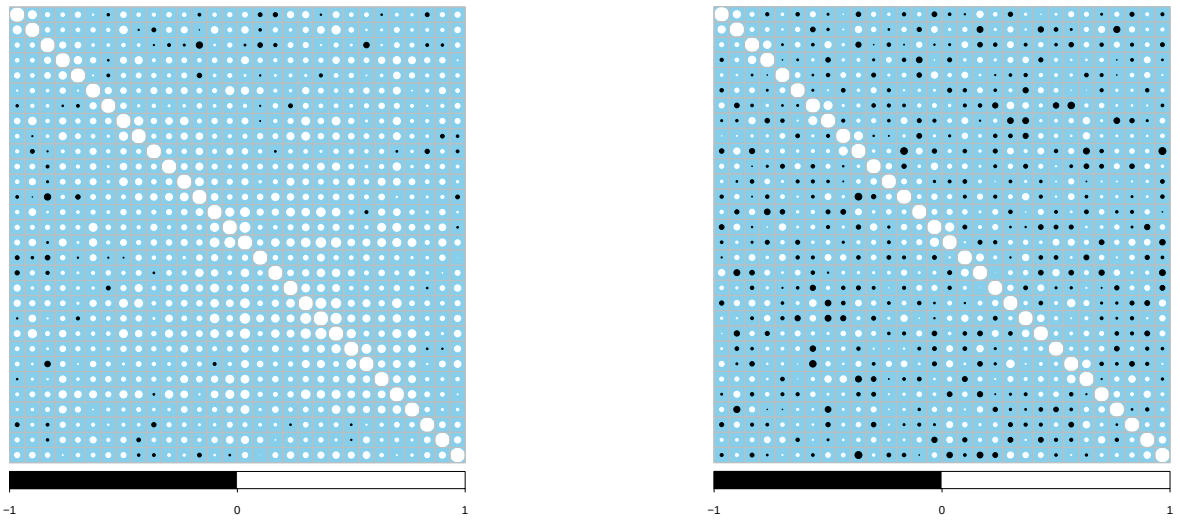


(a) Significance threshold levels (with 200 bootstraps) (b) Number of bootstrapping iterations (0.5 threshold)

Figure 9: Plots representing numbers of convergence steps for the LASSO-GLSE iterations in the algorithm, based on different significant thresholds and bootstrapping sample size.

6 Real data analysis

Monthly returns of more than 850 stocks are used in this study (which have a complete return history over the last 10 years on London Stock Exchange) starting on 31/07/2005 and ending on 31/07/2015 (120 months). The year 2008, which is the peak period of the financial crisis (Ivashina and Scharfstein, 2010), has been removed from the data in order not to influence the analysis. The marginal and conditional correlations of the stocks are shown in Figure 10, via a small sample of 30 stocks which are selected randomly. Figure 10 displays that most of the marginal correlations range from weak to moderate, while the conditional ones vary between very weak and weak with a few moderate ones. This means that there is an underlying graph structure that needs to be identified. In the world of finance, this may imply that the portfolio's risk can be lowered by taking into consideration the correlations between assets in the portfolio (Levi, 2009).



(a) Marginal correlations (b) Conditional correlations

Figure 10: Marginal and conditional correlations plots for randomly chosen 30 assets of the data.

Out of the 850 stocks, 100, 200, 300 and 400 have been selected at random to construct

portfolios using ridge and HGLSE methods. Those portfolios are constructed by using the past 96 months (8 years) of stock return. Subsequently, the portfolios are kept for one year, and their monthly out-of-sample returns are observed. The in-sample period for of portfolios is from 31/07/2005 to 31/12/2007 and from 01/01/2009 to 31/07/2014, and the out-of-sample period lasts from 01/08/2014 to 31/07/2015.

For ridge, cross validation is used for obtaining the penalty parameter that gives the smallest mean squared error. The Sharpe ratios, expected returns and risk for both in-and-out-of-sample periods are computed, which is used to evaluate the performance of different sizes of the obtained portfolios. The results in Table 5 reveal that the HGLSE method outperforms the ridge method, in terms of Sharpe ratio, expected rate of returns and the risk (measured by the standard deviations) of portfolios for both in-and-out-of-sample periods. Moreover, it can also be noticed that the HGLSE method reduces the risk of portfolios for both in-and-out of sample periods more than ridge, when the size of the portfolio increases; except for the 200 sample portfolio, the risk is almost the same. Thus, it can be argued that the HGLSE method may be a good strategy in large data sets as it could construct a diversified portfolio with smaller risk. The 200 sample portfolio has a high number of small companies which have more risk investments than large companies (Vassalou and Xing, 2004). However, the Sharp ratios of the 200 sample portfolio constructed by the HGLSE is significantly much higher compared to the ones produced by the ridge method for both in-and-out-sample periods.

In the first 100-assets portfolio, HGLSE selects 29 out of 100 as important assets based on the significance of their estimated coefficients. These assets have a significant impact on the returns of the portfolio, as they contribute by more than 50% of the total portfolio returns for the in-sample period and approximately 85% of the total portfolio returns for the out-of-sample period. In the rest of portfolios (200, 300, and 400 stocks), the HGLSE method selects 40, 45 and 60 assets, respectively, as important assets. These assets yield a considerable effect on the total portfolios' return as they account for in-sample periods by nearly 44%, 53% and 53% and for out-of-sample periods by approximately 47%, 85% and

88%, respectively. Although the contribution of 47% of the 200-assets portfolio's return by 40 important assets, for the out-of-sample period, appears to be a substantial percentage, this value is still relatively small if compared with the 85% and 88% accounted for by 45 and 60 assets, respectively, in other portfolios. This is due to the same reason mentioned above regarding the fact that the 200 sample belong to companies most of which are small firms, which is deemed as a risk investment.

Moreover, the 29 assets in the case of a 100-assets portfolio are distributed among 29 cliques separated by important and less important assets, which is consistent with the low conditional correlations in Figure 10. These low conditional correlations in Figure 10 are very likely to be responsible for the sparseness of the cliques (i.e. most of the conditional correlations are zero). This implies that the risk of important assets will not be dominant due to the absence of a direct relation between these assets and the fact that most of these assets come from various market sectors, which suggests that the HGLSE has diversified the sources of the involved assets. In other words, if one market sector declines or collapses, this is unlikely to have a substantial impact on the performance of the portfolio. However, if all these assets have a direct relation with one another within a clique or come from one market sector, this will mean that a decline of one asset will potentially affect the rest of the assets in the clique.

Table 5: Portfolios' sizes and in-and-out-of-sample portfolio's Sharpe ratios, expected returns and risk found by ridge and HGLSE.

Portfolio size	Methods	Sharpe ratio	Expected returns	Portfolio's risk
100 stocks (in sample)	Ridge	1.044	0.050	0.048
	HGLSE	1.164	0.054	0.046
100 stocks (out of sample)	Ridge	0.376	0.020	0.053
	HGLSE	0.672	0.034	0.050
200 stocks (in sample)	Ridge	1.195	0.051	0.043
	HGLSE	1.721	0.073	0.042
200 stocks (out of sample)	Ridge	0.302	0.017	0.057
	HGLSE	0.447	0.025	0.056
300 stocks (in sample)	Ridge	1.579	0.060	0.038
	HGLSE	1.954	0.064	0.033
300 stocks (out of sample)	Ridge	0.394	0.023	0.058
	HGLSE	0.754	0.035	0.046
400 stocks (in sample)	Ridge	1.537	0.051	0.033
	HGLSE	2.002	0.062	0.031
400 stocks (out of sample)	Ridge	0.538	0.020	0.037
	HGLSE	0.715	0.024	0.033

7 Conclusion

We have proposed a hybridized GLSE procedure with the idea of imposing a ridge type penalty on less important variables, while maintaining unbiased coefficient estimates of the important covariates/assets. In the proposed method, The paper has also showed that the general form of GLSE gives unbiased estimators for all variables under certain conditions and proved that the proposed HGLSE method is only biased for less important covariates/assets due to the ridge type penalty. The new method is applied to portfolio optimization problem when $N < v$. The results from the data analysis and simulations indicate that the HGLSE

can always outperform the ridge method in maximizing the portfolio's Sharpe ratios and expected rates of returns and minimizing their risk. The proposed method gives a better diversified portfolio risk in small and large data sets. For regression problems where unbiased estimates or saturated regression models are more important, the new method provides a good alternative to existing regularization methods.

The proposed method uses the idea of graphical models, but it does not require a particular distribution assumption for $\mathbf{x}$ but uses a linear relationship assumption among predictors. This is an advantage of the proposed method, as existing graphical model theory needs the assumption that $\mathbf{x}$ follows a multivariate Gaussian distribution, which may not be true for the asset returns. In this particular application, the response and covariates are all continuous variables, which is actually a condition of applying the proposed method. It is worth studying further how to extend this idea to more general regression frameworks.

References

S. Aldahmani and H. Dai. Unbiased estimation for linear regression when $n < v$. *International Journal of Statistics and Probability*, 4(3):p61, 2015.

M. Britten-Jones. The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671, 1999.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

M. Carrasco and N. Noumon. Optimal portfolio selection using regularization. Technical report, University of Montreal, 2011.

V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

C. Ennew, T. Watkins, and M. Wright. *Cases in marketing financial services*. Oxford, Butterworth-Heinemann Limited, 2014.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

B. Fastrich, S. Paterlini, and P. Winker. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, 12(3):417–434, 2015.

M. Gilli, D. Maringer, and E. Schumann. *Numerical methods and optimization in finance*. Academic Press, 2011.

J. Goeman, R. Meijer, and N. Chaturvedi. L1 and l2 penalized regression models. *cran. r-project. or*, 2012.

M. C. Golumbic. *Algorithmic graph theory and perfect graphs*, volume 57. Elsevier, 2004.

T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

V. Ivashina and D. Scharfstein. Bank lending during the financial crisis of 2008. *Journal of Financial economics*, 97(3):319–338, 2010.

T. L. Lai, H. Xing, and Z. Chen. Mean-variance portfolio optimization when means and covariances are unknown. *The Annals of Applied Statistics*, 5(2A):798–823, 2011.

S. L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

G. Laursen and J. Thorlund. *Business analytics for managers: Taking business intelligence beyond reporting*, volume 40. John Wiley & Sons, 2010.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

O. Ledoit and M. Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *SSRN 2383361*, 2014.

C.-F. Lee and J. Lee. *Handbook of quantitative finance and risk management*. Springer Science and Business Media, 2010.

M. D. Levi. *International Finance 5th Edition*. Routledge, 2009.

J. Li. Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business and Economic Statistics*, 33(3):381–392, 2014.

X. Long, K. Solna, and J. Xin. Three l_1 based nonconvex methods in constructing sparse mean reverting portfolios. *Journal of Scientific Computing*, 75(2):1156–1186, 2018.

B. G. Malkiel and Y. Xu. Risk and return revisited. *The Journal of Portfolio Management*, 23(3):9–14, 1997.

H. Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

M. Norouzirad, S. Hossain, and M. Arashi. Shrinkage and penalized estimators in weighted least absolute deviations regression models. *Journal of Statistical Computation and Simulation*, 88(8):1557–1575, 2018.

P. Radchenko, G. M. James, et al. Improved variable selection with forward-lasso adaptive shrinkage. *The Annals of Applied Statistics*, 5(1):427–448, 2011.

J. R. Raol, G. Girija, and J. Singh. *Modelling and parameter estimation of dynamic systems*, volume 65. Iet, 2004.

S. Still and I. Kondor. Regularizing portfolio optimization. *New Journal of Physics*, 12(7): 075034, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 2011.

M. Vassalou and Y. Xing. Default risk in equity returns. *The Journal of Finance*, 59(2): 831–868, 2004.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

A Proof of Theorem 2.1

The proof for a very special case of the GLSE method (the graph with only two cliques) has been shown in Aldahmani and Dai (2015). Here we provide a more general proof with respect to Theorem 2.1, applicable to GLSE with more than two cliques.

Proof. The GLSE in (7) is associated with a graph g with cliques $C_1, ..., C_q$ as a perfect sequence and separators $S_2, ..., S_q$. For $\kappa = 2, \cdots, q$, define the index sets

$$\begin{aligned} H_\kappa &= \cup_{i=1}^{\kappa} C_i, \\ H_\kappa^c &:= H_{\kappa+1} \backslash H_\kappa = C_{\kappa+1} \backslash S_{\kappa+1} \\ J_\kappa &= H_\kappa \backslash S_{\kappa+1} \end{aligned} \tag{13}$$

with $H_q = V$ and $H_q^c = \Phi$ (the empty set) and define the matrix

$$\mathbf{K}_\kappa = \sum_{i=1}^{\kappa} [(ssd_{C_i})^{-1}]^\Gamma - \sum_{i=2}^{\kappa-1} [(ssd_{S_i})^{-1}]^\Gamma. \tag{14}$$

Note that the elements in the ith rows and elements in the jth columns of $\mathbf{K}_\kappa$, for $i, j \in H_\kappa^c$, are all 0s. Therefore if we write

$$\mathbf{x}'_{H_{\kappa+1}} \mathbf{x}_{H_{\kappa+1}} = \begin{pmatrix} ssd_{H_\kappa} & \mathbf{x}'_{H_\kappa} \mathbf{x}_{H_\kappa^c} \\ \mathbf{x}'_{H_\kappa^c} \mathbf{x}_{H_\kappa} & ssd_{H_\kappa^c} \end{pmatrix} \tag{15}$$

we then have

$$\mathbf{K}_\kappa \cdot [\mathbf{x}'_{H_{\kappa+1}} \mathbf{x}_{H_{\kappa+1}}]^\Gamma = \mathbf{K}_\kappa \cdot [\mathbf{x}'_{H_\kappa} \mathbf{x}_{H_\kappa}]^\Gamma + \mathbf{K}_\kappa \cdot [\mathbf{x}'_{H_\kappa}]^\Gamma [\mathbf{x}_{H_\kappa^c}]^\Gamma. \tag{16}$$

From (7) and $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ we know $\mathrm{E}(\hat{\boldsymbol{\beta}}^g) = \mathrm{E}\,[\mathbf{K}_q \cdot \mathbf{x}'\mathbf{x} \cdot \boldsymbol{\beta}]$. Therefore we only need to show

that

$$\mathrm{E}\left[\mathbf{K}_q \cdot \mathbf{x}'\mathbf{x}\right] = \mathbf{I}. \tag{17}$$

Now we prove (17) via mathematical induction. Suppose for $\kappa < q$, that

$$\mathrm{E}\left[\mathbf{K}_\kappa \cdot [\mathbf{x}'_{H_\kappa}\mathbf{x}_{H_\kappa}]^\Gamma\right] = [\boldsymbol{I}_{H_\kappa}]^\Gamma. \tag{18}$$

Then from (14), (16) and the mathematical induction assumption (18) we have that, for $\kappa+1$,

$$\begin{aligned}
& \mathrm{E}\left[\mathbf{K}_{\kappa+1} \cdot [\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^\Gamma\right] \\
= \; & \mathrm{E}\left[\left\{\mathbf{K}_\kappa + \left[(ssd_{C_{\kappa+1}})^{-1}\right]^\Gamma - \left[(ssd_{S_{\kappa+1}})^{-1}\right]^\Gamma\right\} \cdot [\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^\Gamma\right] \\
= \; & \mathrm{E}\left[\mathbf{K}_\kappa[\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^\Gamma\right] + \mathrm{E}\left[\left\{\left[(ssd_{C_{\kappa+1}})^{-1}\right]^\Gamma - \left[(ssd_{S_{\kappa+1}})^{-1}\right]^\Gamma\right\} \cdot [\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^\Gamma\right] \\
= \; & [\mathbf{I}_{H_\kappa}]^\Gamma + \mathrm{E}\left[\mathbf{K}_\kappa \cdot [\mathbf{x}'_{H_\kappa}]^\Gamma[\mathbf{x}_{H_\kappa^c}]^\Gamma\right] + \mathrm{E}\left[\left\{\left[(ssd_{C_{\kappa+1}})^{-1}\right]^\Gamma - \left[(ssd_{S_{\kappa+1}})^{-1}\right]^\Gamma\right\} \cdot [\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^\Gamma\right].
\end{aligned} \tag{19}$$

Condition 2.2 gives $\mathbf{x}_{H_\kappa^c} = \mathbf{x}_{S_{\kappa+1}} \cdot \mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}} + \boldsymbol{\xi}_{H_\kappa^c}$, for some $|s_{\kappa+1} \times (c_{\kappa+1} - s_{\kappa+1})|$ dimensional matrix $\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}}$ and some independent zero-mean residual matrix $\boldsymbol{\xi}^c_{H_\kappa}$. We can write $\mathbf{r}$ in terms of its elements via,

$$\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}} = (r_{ij})_{i\in S_{\kappa+1}, j\in C_{\kappa+1}\backslash S_{\kappa+1}}. \tag{20}$$

If we define the expansion for this $\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}}$ to an $V \times V$ matrix $[\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}}]^\Gamma$, as

$$\left([\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}}]^\Gamma\right)_{ij} = \begin{cases} r_{ij} & \text{if } i \in S_{\kappa+1}, j \in C_{\kappa+1}\backslash S_{\kappa+1} \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

Then Condition 2.2 further implies

$$[\mathbf{x}_{H_\kappa^c}]^\Gamma \;=\; [\mathbf{x}_{H_\kappa}]^\Gamma \cdot [\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\backslash\mathrm{S}_{\kappa+1}}]^\Gamma + [\boldsymbol{\xi}_{H_\kappa^c}]^\Gamma. \tag{22}$$

Therefore

$$\mathrm{E}\left(\mathbf{K}_{\kappa}\cdot[\mathbf{x}'_{H_{\kappa}}]^{\Gamma}[\mathbf{x}_{H_{\kappa}^{c}}]^{\Gamma}\right)=[\mathbf{I}_{H_{\kappa}}]^{\Gamma}[\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\setminus\mathrm{S}_{\kappa+1}}]^{\Gamma}. \tag{23}$$

We also have the following terms,

$$\begin{aligned}&\mathrm{E}\left[\left[(ssd_{C_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^{\Gamma}\right]\\ =\;&[\mathbf{I}_{C_{\kappa+1}}]^{\Gamma}+\mathrm{E}\left[\left[(ssd_{C_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{C_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]\end{aligned} \tag{24}$$

and

$$\begin{aligned}&-\mathrm{E}\left[\left[(ssd_{S_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^{\Gamma}\right]\\ =\;&-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}-\mathrm{E}\left[\left[(ssd_{S_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{S_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]-\mathrm{E}\left[\left[(ssd_{S_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{S_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{H_{\kappa}^{c}}]^{\Gamma}\right]\\ =\;&-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}-\mathrm{E}\left[\left[(ssd_{S_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{S_{\kappa+1}}][\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}[\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\setminus\mathrm{S}_{\kappa+1}}]^{\Gamma}\end{aligned} \tag{25}$$

Therefore subsitituting (23), (24) and (25) into (19) we know that

$$\begin{aligned}&\mathrm{E}\left[\mathbf{K}_{\kappa+1}\cdot[\mathbf{x}'_{H_{\kappa+1}}\mathbf{x}_{H_{\kappa+1}}]^{\Gamma}\right]\\ =\;&[\mathbf{I}_{H_{\kappa}}]^{\Gamma}+[\mathbf{I}_{H_{\kappa}}]^{\Gamma}[\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\setminus\mathrm{S}_{\kappa+1}}]^{\Gamma}+[\mathbf{I}_{C_{\kappa+1}}]^{\Gamma}+\mathrm{E}\left[\left[(ssd_{C_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{C_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]\\ &-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}-\mathrm{E}\left[\left[(ssd_{S_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{S_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}[\mathbf{r}_{\mathrm{S}_{\kappa+1},\mathrm{C}_{\kappa+1}\setminus\mathrm{S}_{\kappa+1}}]^{\Gamma}\\ =\;&[\mathbf{I}_{H_{\kappa}}]^{\Gamma}+[\mathbf{I}_{C_{\kappa+1}}]^{\Gamma}-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}=[\mathbf{I}_{H_{\kappa+1}}]^{\Gamma},\end{aligned} \tag{26}$$

where the result uses $[\mathbf{I}_{H_{\kappa}}]^{\Gamma}[\mathbf{r}_{S_{\kappa+1},C_{\kappa+1}\setminus S_{\kappa+1}}]^{\Gamma}-[\mathbf{I}_{S_{\kappa+1}}]^{\Gamma}[\mathbf{r}_{S_{\kappa+1},C_{\kappa+1}\setminus S_{\kappa+1}}]^{\Gamma}=\mathbf{0}$ and

$$\mathrm{E}\left[\left[(ssd_{C_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{C_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]-\mathrm{E}\left[\left[(ssd_{S_{\kappa+1}})^{-1}\right]^{\Gamma}\cdot[\mathbf{x}'_{S_{\kappa+1}}]^{\Gamma}[\mathbf{x}_{J_{\kappa}}]^{\Gamma}\right]=\mathbf{0}. \tag{27}$$

Equation (27) is from the fact that Condition 2.2 implies $\mathbf{x}_{J_{\kappa}}=\mathbf{x}_{S_{\kappa+1}}\cdot\mathbf{q}_{\mathrm{S}_{\kappa+1},\mathrm{J}_{\kappa}}+\boldsymbol{\eta}_{J_{\kappa}}$, for

some $|S_{\kappa+1}| \times |J_\kappa|$ dimensional matrix $\mathbf{q}_{\mathrm{S}_{\kappa+1},\mathrm{J}_\kappa}$ with $\mathrm{E}(\boldsymbol{\eta}_{J_\kappa}) = \mathbf{0}$ and further

$$[\mathbf{x}_{J_\kappa}]^\Gamma = [\mathbf{x}_{S_{\kappa+1}}]^\Gamma \cdot [\mathbf{q}_{\mathrm{S}_{\kappa+1},\mathrm{J}_\kappa}]^\Gamma + [\boldsymbol{\eta}_{J_\kappa}]^\Gamma = [\mathbf{x}_{C_{\kappa+1}}]^\Gamma \cdot [\mathbf{q}_{\mathrm{S}_{\kappa+1},\mathrm{J}_\kappa}]^\Gamma + [\boldsymbol{\eta}_{J_\kappa}]^\Gamma. \tag{28}$$

The theorem is then proved by mathematical induction with the fact that for $\kappa = 1$, $\mathrm{E}\left[\mathbf{K}_2 \cdot [\mathbf{x}'_{H_2}\mathbf{x}_{H_2}]^\Gamma\right] = [\mathbf{I}_{H_2}]^\Gamma$ (from the results in Aldahmani and Dai (2015) for a graph with only two cliques) and $[\mathbf{x}'_{H_q}\mathbf{x}_{H_q}]^\Gamma = \mathbf{x}'\mathbf{x}$. □

B Proof of Theorem 3.1

Proof. We have

$$\begin{aligned}
& \mathrm{E}(\hat{\beta}^h|\mathbf{x}_B) \\
= \; & E\left(\left([(ssd_{A\cup B})^{-1}]^\Gamma + [(ssd_{B\cup C} + \lambda\mathbf{I}_{B\cup C})^{-1}]^\Gamma - [(ssd_B)^{-1}]^\Gamma\right)\mathbf{x}'\mathbf{x}\boldsymbol{\beta}|\mathbf{x}_B\right) \\
= \; & \mathrm{E}\left[\begin{pmatrix} \mathbf{I}_A & \mathbf{0} & \\ \mathbf{0} & \mathbf{I}_B & (ssd_{A\cup B})^{-1}\mathbf{x}'_{A\cup B}\mathbf{x}_C \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}\right. \\
& + \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (ssd_{B\cup C} + \lambda\mathbf{I}_{B\cup C})^{-1}\mathbf{x}'_{B\cup C}\mathbf{x}_A & (ssd_{B\cup C} + \lambda\mathbf{I}_{B\cup C})^{-1}\mathbf{x}'_{B\cup C}\mathbf{x}_B & (ssd_{B\cup C} + \lambda\mathbf{I}_{B\cup C})^{-1}\mathbf{x}'_{B\cup C}\mathbf{x}_C \end{pmatrix} \\
& \left. - \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_A & \mathbf{I}_B & (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_C \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}\right| \mathbf{x}_B\Bigg]\boldsymbol{\beta}.
\end{aligned} \tag{29}$$

Using Condition 3.2, we have

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

therefore,

$$\begin{aligned} & \mathrm{E}\left[(ssd_{A\cup B})^{-1}\mathbf{x}'_{A\cup B}\mathbf{x}_C\right] \\ = \; & \mathrm{E}\left[(ssd_{A\cup B})^{-1}\mathbf{x}'_{A\cup B}\mathbf{x}_{A\cup B}\begin{pmatrix}\mathbf{0}\\ \boldsymbol{r}_{B,C}\end{pmatrix} - \begin{pmatrix}\mathbf{0}\\ (ssd_B)^{-1}\mathbf{x}'_B\mathbf{x}_B\boldsymbol{r}_{B,C}\end{pmatrix}\right] \\ = \; & \mathrm{E}\left[\begin{pmatrix}\mathbf{0}\\ \boldsymbol{r}_{B,C}\end{pmatrix} - \begin{pmatrix}\mathbf{0}\\ \boldsymbol{r}_{B,C}\end{pmatrix}\right] = \mathbf{0}. \end{aligned} \tag{30}$$

Then equation (29) becomes

$$\mathrm{E}(\hat{\beta}^h|\mathbf{x}_B) \;=\; \begin{pmatrix} \mathbf{I}_A & \mathbf{0} & \mathbf{0} \\ * & * & * \\ * & * & * \end{pmatrix}\boldsymbol{\beta} \tag{31}$$

and further $\boldsymbol{E}(\hat{\boldsymbol{\beta}}^h_A) = \boldsymbol{\beta}_A$, but $\boldsymbol{E}(\hat{\boldsymbol{\beta}}^h_B) \neq \boldsymbol{\beta}_B$, $\boldsymbol{E}(\hat{\boldsymbol{\beta}}^h_C) \neq \boldsymbol{\beta}_C$. □