A Bayesian Information Criterion for Portfolio Selection

Wei Lan[1], Hansheng Wang[1], and Chih-Ling Tsai[2]

[1] *Guanghua School of Management, Peking University &* [2] *Graduate School of Management, University of California, Davis.*

This version: May 8, 2011

Abstract

The mean-variance theory of Markowitz (1952) indicates that large investment portfolios naturally provide better risk diversification than small ones. However, due to parameter estimation errors, one may find ambiguous results in practice. Hence, it is essential to identify relevant stocks to alleviate the impact of estimation error in portfolio selection. To this end, we propose a linkage condition to link the relevant and irrelevant stock returns via their conditional regression relationship. Subsequently, we obtain a BIC selection criterion that enables us to identify relevant stocks consistently. Numerical studies indicate that BIC outperforms commonly used portfolio strategies in the literature.

KEY WORDS: Bayesian Information Criterion; Minimal Variance Portfolio; Portfolio Selection; Risk Diversification; Selection Consistency

[†]Both Wei Lan and Hansheng Wang (the corresponding author) are from the Guanghua School of Management, Peking University, Beijing, 100871, P. R. China. Chih-Ling Tsai is from the Graduate School of Management, University of California-Davis, California, 95616-8609, U.S.A. The research of Lan and Wang is supported in part by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China. It is also supported in part by a grant from PKU-Guanghua Harvest Institute for Investment and a grant from Center for Statistical Science at Peking University. The authors are grateful to the Editor, the Associate Editor, two anonymous referees, and also Professor Joseph Chen from University of California-Davis for their helpful comments, which lead a much improved manuscript.

1. INTRODUCTION

In financial risk analysis, Markowitz (1952) proposed mean-variance portfolio selection, and this landmark study earned him the 1990 Nobel Prize in Economic Sciences shared with Merton Miller and William Sharpe. Since then, risk diversification has become an increasingly important tool for analyzing investments; see for example, Jagannathan and Ma (2003), DeMiguel et al. (2009a) and DeMiguel et al. (2009b). This topic is particularly relevant given the current financial market turbulence, which motivates us to study portfolio selection.

In his seminal mean-variance theory, Markowitz (1952) uses two parameters to characterize a portfolio's performance: expected return and variance (i.e., risk). The investor, therefore, commonly optimizes his/her investment portfolio with an appropriate trade-off between expected return and risk. Because investors have different risk attitudes, the number of potentially optimal portfolios could be large. Among various optimal portfolios, the one with minimal variance is of particular interest for two reasons. First, the minimum-risk portfolio could be attractive to those investors with strong risk aversion characteristics (e.g., governments, pension funds). Second, although the criterion is minimal risk, the actual return remains competitive; see Jagannathan and Ma (2003), DeMiguel et al. (2009a), and DeMiguel et al. (2009b). This is because it is considerably more difficult to accurately estimate the mean of a stock's return than its variance; see Jorion (1986) and Jagannathan and Ma (2003). Testing a portfolio's mean-variance spanning is, therefore, important (Huberman and Kandel, 1987), and many researchers advocate for the minimal risk criterion in portfolio selection (Jagannathan and Ma, 2003).

To effectively employ the minimum-risk criterion in practice, one needs to accurately estimate covariance matrices. Following Markowitz (1952), diversification can reduce

the overall risk of an investment portfolio, and this strategy naturally leads us to favor larger portfolios. However, this notion induces high-dimension covariance matrices, which are difficult to estimate accurately; see Bickel and Levina (2008) and Rothman et al. (2009). Furthermore, past empirical evidence suggests that the sampling error in the covariance estimation process can significantly deteriorate a portfolio's out-of-sample performance. Moreover, estimation errors also lead to considerable portfolio instability. Accordingly, portfolio weights need to be adjusted frequently and appreciably, which yields non-negligible transaction costs. Hence, a good strategy should take into account estimation errors.

In the last decade, some empirical researchers (Goetzmann and Kumar, 2001; Polkovnichenko, 2003; Statman, 2004) have found that investors tend not to hold many stocks in their portfolios; average portfolio size is 3 or 4 stocks. From statistical consideration, this finding is sensible since a smaller portfolio size requires a fewer number of unknown parameters to be estimated. This in turn reduces the estimation instability and subsequently brings down the transaction or holding cost (Statman, 2004). These findings motivate us to consider alleviating the estimation error effect by controlling the size of the portfolio. To this end, we define relevant stocks (i.e, stocks that must be included in the portfolio) and irrelevant stocks (i.e., stocks that cannot provide any additional risk reduction given the existing relevant stocks). The optimal portfolio is established by choosing relevant stocks that balance diversification and estimation error. Therefore, the aim of this paper is to develop a selection criterion that enables us to consistently differentiate relevant and irrelevant stocks.

The rest of the article is organized as follows. Section 2 defines relevant and irrelevant stocks and proposes a linkage condition to link the relevant and irrelevant stock returns via their conditional regression relationship. Accordingly, we obtain the

Bayesian information criterion (BIC) and demonstrate its consistency (i.e., the capability to consistently differentiate relevant and irrelevant stocks). Section 3 presents numerical examples including Monte Carlo studies and an empirical analysis. The article concludes with a brief discussion in Section 4. All technical details are left to the Appendix.

2. BAYESIAN INFORMATION CRITERION

2.1. Relevant and Irrelevant Stocks

Let X_{tj} $(1 \leq j \leq d)$ be the return of the jth stock observed at time t and $X_t = (X_{t1}, \cdots, X_{td})^\top \in \mathbb{R}^d$, where d is the number of candidate stocks. We further assume that the X_t's are independent and identically distributed random variables with $E(X_t) = 0$ and $\mathrm{cov}(X_t) = \Sigma$ for $t = 1, \cdots, n$. To minimize the portfolio variance, one needs to find an optimal weight vector $\omega = (\omega_1, \cdots, \omega_d)^\top \in \mathbb{R}^d$, such that the variance $\mathrm{var}(\omega^\top X_t) = \omega^\top \Sigma \omega$ can be minimized under the constraint $\omega^\top \mathbf{1} = 1$, where $\mathbf{1} = (1, 1, \cdots, 1)^\top \in \mathbb{R}^d$. It has been shown that the optimal solution to this minimization problem is $\omega_0 = (\omega_{01}, \cdots, \omega_{0d})^\top = \Sigma^{-1}\mathbf{1}\{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}\}^{-1}$; see for example, Ledoit and Wolf (2003). To assess the out-of-sample performance, we consider $X_0 \in \mathbb{R}^d$ to be an independent copy of X_t. Then, the resulting portfolio's out-of-sample variance is $\mathrm{var}(\omega_0^\top X_0) = (\mathbf{1}^\top \Sigma^{-1}\mathbf{1})^{-1}$.

For the sake of convenience, we introduce generic notation $\mathcal{S} = \{j_1, \cdots, j_{\tilde{d}}\}$ to represent the portfolio that includes the j_1th, j_2th, $\cdots$, $j_{\tilde{d}}$th stocks. We denote its size as $|\mathcal{S}| = \tilde{d}$. Let $\mathcal{S}_F = \{1, 2, \cdots, d\}$ be the full-size portfolio that contains all candidate stocks. In addition, for any d-dimensional vector $\beta \in \mathbb{R}^d$ and $d \times d$ matrix $\Omega \in \mathbb{R}^{d \times d}$, let $\beta_{(\mathcal{S})}$ and $\Omega_{(\mathcal{S})}$ represent their corresponding sub-vector and sub-matrix. Accordingly, the return vector of the portfolio $\mathcal{S}$ at time t and its covariance matrix

are given by $X_{t(\mathcal{S})} = (X_{tj} : j \in \mathcal{S})^\top \in \mathbb{R}^{|\mathcal{S}|}$ and $\Sigma_{(\mathcal{S})} = (\sigma_{j_1 j_2})_{j_1, j_2 \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$, respectively. Moreover, for any two portfolios $\mathcal{S}_a$ and $\mathcal{S}_b$, we use the notation $\Sigma_{(\mathcal{S}_a, \mathcal{S}_b)}$ to represent the sub-matrix of Σ, where its rows and columns are determined by $\mathcal{S}_a$ and $\mathcal{S}_b$, correspondingly. For example, let $\hat{\Sigma} = n^{-1} \sum X_t X_t^\top = n^{-1}(\mathbb{X}^\top \mathbb{X})$ be the sample covariance matrix of the full-size portfolio $\mathcal{S}_F$, where $\mathbb{X} = (X_1, \cdots, X_n)^\top$. Then, $\hat{\Sigma}_{(\mathcal{S}_a, \mathcal{S}_b)} = (\hat{\sigma}_{j_1 j_2} : j_1 \in \mathcal{S}_a, j_2 \in \mathcal{S}_b) \in \mathbb{R}^{|\mathcal{S}_a| \times |\mathcal{S}_b|}$ is the sample covariance between $\mathcal{S}_a$ and $\mathcal{S}_b$. The difference between the subscript with parentheses and without parentheses is noteworthy. For example, $\omega_{0(\mathcal{S})}$ denotes a sub-vector of $\omega_0 \in \mathbb{R}^d$, where ω_0 is the optimal weight vector associated with the full-size portfolio $\mathcal{S}_F$. On the other hand, $\omega_{0\mathcal{S}} = \{\Sigma_{(\mathcal{S})}^{-1} \mathbf{1}_{(\mathcal{S})}\}\{\mathbf{1}_{(\mathcal{S})}^\top \Sigma_{(\mathcal{S})}^{-1} \mathbf{1}_{(\mathcal{S})}\}^{-1}$ is the optimal weight vector computed via the portfolio $\mathcal{S}$ only, which leads to $\omega_0 = \omega_{0\mathcal{S}_F}$.

Inspired by mean-variance spanning theory (Huberman and Kandel, 1987; Gibbons et al, 1989; Kan and Zhou, 2001), we next define a stock to be relevant (irrelevant) if its corresponding weight in ω_0 is non-zero (zero). Then the optimal portfolio is $\mathcal{S}_0 = \{j : \omega_{0j} \neq 0\}$ with size $d_0 = |\mathcal{S}_0|$, while its complement is $\mathcal{S}_0^c = \mathcal{S}_F \backslash \mathcal{S}_0$ with size $|\mathcal{S}_0^c| = d - d_0$. Although the relevant and irrelevant stocks are clearly defined, they are not directly useful for constructing the likelihood function of the portfolio. This is because the conditional regression relationship between the relevant and irrelevant stocks is not explicitly specified. To this end, we obtain the following theorem, whose detailed technical proof can be found in Appendix A.

Theorem 1. *Assume that X_t follows a multivariate normal distribution for $t = 1, \cdots, n$. Then, a necessary and sufficient condition for $\mathcal{S} \supset \mathcal{S}_0$ is that, for any $k \notin \mathcal{S}$, we have $\sum_{j \in \mathcal{S}} \beta_{kj} = 1$, where β_{kj} are regression coefficients of X_{tk} on $\{X_{tj}, j \in \mathcal{S}\}$.*

The above theorem indicates that the condition $\sum_{j \in \mathcal{S}} \beta_{kj} = 1$ is crucial in determining whether $\mathcal{S} \supset \mathcal{S}_0$. To further understand this condition, we present an insightful

discussion below. For an arbitrary portfolio $\mathcal{S}$ and any given stock $k \notin \mathcal{S}$, we have

$$X_{tk} = \sum_{j \in \mathcal{S}} X_{tj}\beta_{kj} + \varepsilon_{tk}, \tag{2.1}$$

where (2.1) is stated in Appendix A for the proof of Theorem 1. It is noteworthy that the error term ε_{tk} is assumed to be independent of X_{tj} for $j \in \mathcal{S}$, and that such an assumption is crucial for the implementation of our proposed method. Similar assumption has been used in the mean-variance spanning literature; see for example Huberman and Kandel (1987). If $\sum_{j\in\mathcal{S}} \beta_{kj} = 1$, then we can treat $\{\beta_{kj} : j \in \mathcal{S}\}$ as a set of portfolio weights for $\mathcal{S}$. Thus, the return of $\mathcal{S}$ is $\sum_{j\in\mathcal{S}} X_{tj}\beta_{kj}$. This together with (2.1) implies that the expected return of the kth stock is exactly the same as that of the portfolio $\mathcal{S}$, i.e., $E(X_{tk}) = E(\sum_{j\in\mathcal{S}} X_{tj}\beta_{kj})$. However, the kth stock has a larger risk than that of $\mathcal{S}$, i.e., $\text{var}(X_{tk}) = \text{var}(\sum_{j\in\mathcal{S}} X_{tj}\beta_{kj}) + \text{var}(\varepsilon_{tk}) > \text{var}(\sum_{j\in\mathcal{S}} X_{tj}\beta_{kj})$. As a result, including the kth stock into portfolio $\mathcal{S}$ neither improves the portfolio's mean return nor reduces its risk. Consequently, we expect that $\mathcal{S} \supset \mathcal{S}_0$.

Remark 1. It is noteworthy that Britten-Jones (1999) and Kempf and Memmel (2006) also studied portfolios via the regression approach. Specifically, Britten-Jones (1999) employed the regression approach to test the weights of efficient portfolio, while Kempf and Memmel (2006) used it to obtain the distribution of the estimated weights of the global minimum variance portfolio and then they made statistical inferences for the efficient portfolio weights. In this article, however, our focus is to find the optimal portfolio by identifying the relevant stocks in a large set of risky assets.

2.2. Maximum Likelihood Estimation

In the context of model selection, one often assumes that the candidate model includes the true model. Adopting this assumption, the candidate portfolio consists of

the optimal portfolio (i.e., $\mathcal{S} \supset \mathcal{S}_0$). However, this heuristic assumption does not yield a useful likelihood function for portfolio selection. Therefore, we obtain the necessary and sufficient condition mentioned in Theorem 1, which links the relevant and irrelevant stocks via their conditional regression relationship. For the sake of simplicity, we name $\sum_{j\in\mathcal{S}} \beta_{kj} = 1$ the *linkage condition*.

Under the linkage condition, we are able to establish the joint likelihood function of $X_t = (X_{t(\mathcal{S})}^\top, X_{t(\mathcal{S}^c)}^\top)^\top \in \mathbb{R}^d$ $(t = 1, \cdots, n)$ for the candidate portfolio $\mathcal{S}$ in the following two steps. First, we obtain the marginal likelihood function of $X_{t(\mathcal{S})}$,

$$\ell(\Sigma_{(\mathcal{S})}) = \left(\frac{1}{2\pi}\right)^{n|\mathcal{S}|/2} \left|\Sigma_{(\mathcal{S})}\right|^{-n/2} \exp\left\{ -\frac{1}{2}\sum_{t=1}^{n} X_{t(\mathcal{S})}^\top \Sigma_{(\mathcal{S})}^{-1} X_{t(\mathcal{S})} \right\}. \tag{2.2}$$

We next get the conditional likelihood function of $X_{t(\mathcal{S}^c)}$ given $X_{t(\mathcal{S})}$,

$$\ell(\Sigma_{\mathcal{S}^c|\mathcal{S}}) = \left(\frac{1}{2\pi}\right)^{n|\mathcal{S}^c|/2} \left|\Sigma_{\mathcal{S}^c|\mathcal{S}}\right|^{-n/2}$$

$$\times \exp\left\{ -\frac{1}{2}\sum_{t=1}^{n} \left(X_{t(\mathcal{S}^c)} - B_{\mathcal{S}^c|\mathcal{S}}^\top X_{t(\mathcal{S})}\right)^\top \Sigma_{\mathcal{S}^c|\mathcal{S}}^{-1} \left(X_{t(\mathcal{S}^c)} - B_{\mathcal{S}^c|\mathcal{S}}^\top X_{t(\mathcal{S})}\right) \right\}, \tag{2.3}$$

where $B_{\mathcal{S}^c|\mathcal{S}}^\top$ is defined in Appendix A and $\Sigma_{\mathcal{S}^c|\mathcal{S}} = \mathrm{cov}(X_{t(\mathcal{S}^c)}|X_{t(\mathcal{S})})$. As a result, the joint likelihood function is $\ell(\Sigma_{(\mathcal{S})}, \Sigma_{(\mathcal{S}^c)}) = \ell(\Sigma_{(\mathcal{S})})\ell(\Sigma_{\mathcal{S}^c|\mathcal{S}})$.

Based on (2.2), we estimate the unknown parameter by maximizing $\ell(\Sigma_{(\mathcal{S})})$ with respect to $\Sigma_{(\mathcal{S})}$, which leads to $\hat{\Sigma}_{(\mathcal{S})} = n^{-1}\sum_{t=1}^{n} X_{t(\mathcal{S})} X_{t(\mathcal{S})}^\top$. Furthermore, under the constraint $B_{\mathcal{S}^c|\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})} = \mathbf{1}_{(\mathcal{S}^c)}$ from Theorem 1, we maximize $\ell(\Sigma_{\mathcal{S}^c|\mathcal{S}})$ in (2.2) with respect to parameters $B_{\mathcal{S}^c|\mathcal{S}}$ and $\Sigma_{\mathcal{S}^c|\mathcal{S}}$. The resulting constrained maximum likelihood estimators are

$$\tilde{B}_{\mathcal{S}^c|\mathcal{S}} = \left(\mathbb{X}_{(\mathcal{S})}^\top \mathbb{X}_{(\mathcal{S})}\right)^{-1} \left(\mathbb{X}_{(\mathcal{S})}^\top \mathbb{X}_{(\mathcal{S}^c)} - \mathbf{1}_{(\mathcal{S})}\tilde{\lambda}_{\mathcal{S}^c}^\top\right) = \hat{\Sigma}_{(\mathcal{S})}^{-1}\left(\hat{\Sigma}_{(\mathcal{S},\mathcal{S}^c)} - n^{-1}\mathbf{1}_{(\mathcal{S})}\tilde{\lambda}_{\mathcal{S}^c}^\top\right)$$

and

$$\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}} = n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}^c)} - \tilde{B}_{\mathcal{S}^c|\mathcal{S}}^{\top}X_{t(\mathcal{S})}\Big)\Big(X_{t(\mathcal{S}^c)} - \tilde{B}_{\mathcal{S}^c|\mathcal{S}}^{\top}X_{t(\mathcal{S})}\Big)^{\top}$$

$$= \hat{\Sigma}_{(\mathcal{S}^c)} - \hat{\Sigma}_{(\mathcal{S},\mathcal{S}^c)}^{\top}\hat{\Sigma}_{(\mathcal{S})}^{-1}\hat{\Sigma}_{(\mathcal{S},\mathcal{S}^c)} + n^{-2}\tilde{\lambda}_{\mathcal{S}^c}\mathbf{1}_{(\mathcal{S})}^{\top}\hat{\Sigma}_{(\mathcal{S})}^{-1}\mathbf{1}_{(\mathcal{S})}\tilde{\lambda}_{\mathcal{S}^c}^{\top}, \tag{2.4}$$

where

$$\tilde{\lambda}_{\mathcal{S}^c}^{\top} = \Big\{\mathbf{1}_{(\mathcal{S})}^{\top}\big(\mathbb{X}_{(\mathcal{S})}^{\top}\mathbb{X}_{(\mathcal{S})}\big)^{-1}\mathbb{X}_{(\mathcal{S})}^{\top}\mathbb{X}_{(\mathcal{S}^c)} - \mathbf{1}_{(\mathcal{S}^c)}^{\top}\Big\}\Big\{\mathbf{1}_{(\mathcal{S})}^{\top}\big(\mathbb{X}_{(\mathcal{S})}^{\top}\mathbb{X}_{(\mathcal{S})}\big)^{-1}\mathbf{1}_{(\mathcal{S})}\Big\}^{-1}$$

$$= n\Big\{\mathbf{1}_{(\mathcal{S})}^{\top}\hat{\Sigma}_{(\mathcal{S})}^{-1}\hat{\Sigma}_{(\mathcal{S},\mathcal{S}^c)} - \mathbf{1}_{(\mathcal{S}^c)}^{\top}\Big\}\Big\{\mathbf{1}_{(\mathcal{S})}^{\top}\hat{\Sigma}_{(\mathcal{S})}^{-1}\mathbf{1}_{(\mathcal{S})}\Big\}^{-1},$$

$\hat{\Sigma}_{(\mathcal{S}^c)} = n^{-1}\sum_{t=1}^{n}X_{t(\mathcal{S}^c)}X_{t(\mathcal{S}^c)}^{\top}$, and $\hat{\Sigma}_{(\mathcal{S},\mathcal{S}^c)} = n^{-1}\sum_{t=1}^{n}X_{t(\mathcal{S})}X_{t(\mathcal{S}^c)}^{\top}$. Accordingly, we find the estimator of the joint likelihood function, $\hat{\ell}(\Sigma_{(\mathcal{S})})\hat{\ell}(\Sigma_{\mathcal{S}^c|\mathcal{S}})$, by replacing those unknown parameters in $\ell(\Sigma_{(\mathcal{S})}, \Sigma_{(\mathcal{S}^c)})$ with their corresponding maximum likelihood estimators.

2.3. A BIC Criterion

To study portfolio selection, we consider the joint likelihood function $\ell(\Sigma_{(\mathcal{S}^c)}, \Sigma_{(\mathcal{S})})$. Following Schwarz (1978), we then obtain the BIC criterion,

$$-2\Big(\log\hat{\ell}(\Sigma_{(\mathcal{S})}) + \log\hat{\ell}(\Sigma_{\mathcal{S}^c|\mathcal{S}})\Big)/n + df \times \log n/n,$$

where *df* is the number of unknown parameters involved in both (2.2) and (2.3). Specifically, we have

$$df = \frac{1}{2}\Big(|\mathcal{S}|+1\Big)|\mathcal{S}| + \frac{1}{2}\Big(d-|\mathcal{S}|+1\Big)\Big(d-|\mathcal{S}|\Big) + \Big(d-|\mathcal{S}|\Big)\Big(|\mathcal{S}|-1\Big),$$

where the first, second, and third terms correspond to the unknown parameters in $\Sigma_{(\mathcal{S})}$, $\Sigma_{\mathcal{S}^c|\mathcal{S}}$, and $B_{\mathcal{S}^c|\mathcal{S}}^{\top}$, respectively. It is noteworthy that the degrees of freedom associated

with $B_{\mathcal{S}^c|\mathcal{S}}^\top$ is calculated by taking into account the constraint $B_{\mathcal{S}^c|\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})} = \mathbf{1}_{(\mathcal{S}^c)}$. After algebraic simplification, the degrees of freedom becomes $df = d(d-1)/2 + |\mathcal{S}|$. Because $d(d-1)/2$ is a constant and can be ignored, we have $df = |\mathcal{S}|$. We further omit irrelevant constants from $\hat{\ell}(\Sigma_{(\mathcal{S})})$ and $\hat{\ell}(\Sigma_{\mathcal{S}^c|\mathcal{S}})$ and obtain

$$\mathrm{BIC}(\mathcal{S}) = \log \left|\hat{\Sigma}_{(\mathcal{S})}\right| + \log \left|\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}}\right| + |\mathcal{S}| \times \log n/n. \tag{2.5}$$

The optimal portfolio selected by BIC is $\hat{\mathcal{S}}_{\mathrm{BIC}} = \mathrm{argmin}_{\mathcal{S}\subset\mathcal{S}_F} \mathrm{BIC}(\mathcal{S})$.

Remark 2. The literature generally considers two categories of selection criteria: loss efficient (e.g., AIC, Akaike, 1973) and selection consistent (e.g., BIC, Schwartz, 1978). Loss efficient criteria choose the best candidate model with minimum expected loss in large samples when the true model is of infinite dimension. In contrast, selection consistent criteria select the true model with probability tending to one in large samples when the true model is of finite dimension and included in the set of candidate models. Detailed discussions on both criteria can be found in Shao (1997) and McQuarrie and Tsai (1998). In this paper, we assumed that the true model (i.e., optimal portfolio) $\mathcal{S}_0$ is of finite dimension and included in the list of candidate models. We further assumed that the optimal portfolio weights ω_0 is a fixed parameter and it is invariant as the sample size $n \to \infty$. Under the above settings, an ideal portfolio selection criterion should be able to identify $\mathcal{S}_0$ consistently. As a result, BIC becomes a preferable choice. However, if no true model $\mathcal{S}_0$ is assumed and/or the optimal portfolio weights ω_0 is changing with respect to the sample size n, the BIC might not be a good choice. Under this situation, the loss efficient criteria should be considered.

Remark 3. The linkage condition allows us to define the relevant and irrelevant stocks. As a result, we are able to construct a likelihood function, and then obtain the BIC

criterion to select relevant stocks. Since BIC relies on the likelihood function, it is not attainable without imposing this condition. After obtaining BIC, the optimal weight is subsequently determined by minimizing $\text{var}(\omega^{\top}_{(\hat{\mathcal{S}}_{\text{BIC}})} X_{i(\hat{\mathcal{S}}_{\text{BIC}})})$, which is $\hat{\omega}_{0\hat{\mathcal{S}}_{\text{BIC}}} = \{\hat{\Sigma}^{-1}_{(\hat{\mathcal{S}}_{\text{BIC}})} \mathbf{1}_{(\hat{\mathcal{S}}_{\text{BIC}})}\}\{\mathbf{1}^{\top}_{(\hat{\mathcal{S}}_{\text{BIC}})} \hat{\Sigma}^{-1}_{(\hat{\mathcal{S}}_{\text{BIC}})} \mathbf{1}_{(\hat{\mathcal{S}}_{\text{BIC}})}\}^{-1}$. Hence, $\hat{\mathcal{S}}_{\text{BIC}}$ and $\hat{\omega}_{0\hat{\mathcal{S}}_{\text{BIC}}}$ are simple to calculate. Furthermore, we can employ the joint likelihood function given in section 2.2 to develop various useful selection criteria (e.g., see McQuarrie and Tsai, 1998), which can be utilized to meet other purposes accordingly.

Remark 4. When the portfolio size, d, is large, it is impractical (often impossible) to consider all of the 2^d possible portfolios. To this end, we employ a backward elimination algorithm to generate a searching path. The optimal portfolio is then selected from this path. Furthermore, following an approach similar to that of Zhang and Wang (2011), we are able to prove theoretically that the resulting path is consistent for model selection.

To better understand the BIC portfolio, we show its asymptotic property. The detailed technical proof can be found in Appendix B.

Theorem 2. *Assume that X_t follows a multivariate normal distribution for $t = 1, \cdots, n$. Then, we have $P(\hat{\mathcal{S}}_{BIC} = \mathcal{S}_0) \to 1$ as $n \to \infty$.*

The above theorem indicates that BIC is able to identify the optimal portfolio $\mathcal{S}_0$ consistently, as long as the sample size is sufficiently large. In practice, however, a large data set may not be available. Hence, it is sensible to evaluate the finite sample performance of BIC empirically, which is done in the next section.

3. NUMERICAL STUDIES

To evaluate the finite sample performance of the BIC criterion, we present two numerical examples. The first one is a simulation study, which allows us to assess the

asymptotic result (i.e., Theorem 2). The second example is a real data set, for which the underlying "true" model structure is not known *a priori*.

To assess the performance of BIC as well as make comparisons, we consider five alternative methods that include NAIVE (the portfolio weights are equally distributed across each of the stocks), MV (the portfolio weights are determined by minimizing the in-sample variance), MU (the weights are determined by maximizing the in-sample expected utility), SHORT (the portfolio weights are determined by minimizing the in-sample risk under the shortsale constraint, $\omega_{0j} \geq 0$ for every $1 \leq j \leq d$), and LASSO (the portfolio weights are determined by minimizing the in-sample variance under an L_1-type constraint with the tuning parameter selected by two-fold cross-validation; see Tibshirani (1996) and DeMiguel et al. (2009a). Detailed illustrations regarding MV, MU, and SHORT can be found in DeMiguel et al. (2009a,b); they are also briefly described in the following remark. It is important to note that BIC, SHORT, and LASSO select stocks and determine their corresponding optimal portfolio weights, while NAIVE, MV, and MU only allocate optimal weights.

Remark 5. Based on the minimum variance portfolio (i.e., MV) approach, the optimal weight can be obtained by solving the equation $\min_\omega \omega^\top \hat{\Sigma} \omega \ s.t. \ \omega^\top \mathbf{1} = 1$, which yields $\omega_{MV} = \hat{\Sigma}^{-1}\mathbf{1}(\mathbf{1}^\top \hat{\Sigma}^{-1}\mathbf{1})^{-1}$. The setting of shortsale constraint portfolio (i.e., SHORT) is similar to MV, except for adding one more constraint $\omega_i \geq 0$ $(i = 1, \cdots, d)$. Furthermore, the optimal weight of the MU portfolio can be obtained by maximizing the in-sample expected utility, i.e., $\max_\omega \omega^\top \hat{\mu} - \kappa \omega^\top \hat{\Sigma} \omega$, where κ is the risk aversion parameter and $\hat{\mu}$ is the in-sample average return. Accordingly, the optimal weight is $\tilde{\omega} = \kappa^{-1}\hat{\Sigma}^{-1}\hat{\mu}$. Since we only focus on risky assets, the resulting optimal weight of the MU portfolio becomes $\omega_{MU} = \tilde{\omega}(\tilde{\omega}^\top \mathbf{1})^{-1} = (\mathbf{1}^\top \hat{\Sigma}^{-1}\hat{\mu})^{-1}\hat{\Sigma}^{-1}\mu$, where the normalization constant $\tilde{\omega}^\top \mathbf{1}$ is to make sure that the sum of weights on risky assets equals 1. It is

noteworthy that ω_{MU} is not related to the risk aversion parameter κ, and it is the same as the optimal weight obtained from the in-sample Sharpe Ratio (Chen et al, 2011).

Remark 6. Finding the optimal weight of the LASSO portfolio is different from those of MV, SHORT, and MV. It is the solution by minimizing the in-sample variance together with an additional penalty term, i.e., $\min_{\omega} \omega^{\top}\tilde{\Sigma}\omega + \lambda|\omega|_1 \quad s.t. \quad \omega^{\top}\mathbf{1} = 1$, where $\tilde{\Sigma}$ is the sample covariance for the given portfolio, $|\omega|_1$ denotes the L_1 norm of ω, and λ is a tuning parameter which can be selected by two-fold cross-validation. Specifically, for a given window with τ observations, we split them into two folds (one is for estimation and the other one is for validation), where each fold consists of $\tau/2$ sample returns. We then choose the LASSO parameter using the following steps. First, compute the sample covariance, $\tilde{\Sigma}$, from the first $\tau/2$ observations. Second, obtain the optimal weight of the LASSO portfolio, ω_{LASSO_λ}, for each given tuning parameter λ. Third, compute $r_{\lambda,t} = \omega_{LASSO_\lambda}^{\top}\mu_t$, where μ_t is the t-th return from the second $\tau/2$ observations. Subsequently, calculate the sample variance σ_λ^2 of $\{r_{\lambda,t} : t = \tau/2+1, \cdots, \tau\}$, and then choose the tuning parameter λ that minimizes σ_λ^2.

3.1. A Simulation Example

We consider three different sizes of the true model and the full model, $(d_0, d) = (5, 30)$, $(10, 60)$, and $(15, 90)$. For a given (d_0, d), we assume that the first d_0 stocks (i.e., X_{tj} with $1 \le j \le d_0$) are relevant and they are simulated from a d_0-dimensional multivariate normal distribution with mean 0 and covariance matrix $\Sigma_{(\mathcal{S}_0)} = \tau(A^{\top}A)$, where $\mathcal{S}_0 = \{1, 2, \cdots, d_0\}$, $A = (a_{j_1 j_2}) \in \mathbb{R}^{d_0 \times d_0}$, and $a_{j_1 j_2}$ are independently generated from a standard normal distribution. It is noteworthy that the constant $\tau \in \mathbb{R}^1$ satisfies $\mathbf{1}_{(\mathcal{S}_0)}^{\top}\Sigma_{(\mathcal{S}_0)}^{-1}\mathbf{1}_{(\mathcal{S}_0)} = \mathbf{1}^{\top}\Sigma^{*-1}\mathbf{1}$, where Σ^* is the sample covariance matrix computed from the real dataset in section 3.2. Hence, the theoretical minimum risk of this simulated portfolio imitates what we observe from the empirical example. Conditional on the first

d_0 stocks, the rest of the stock returns are generated by $X_{tk} = d_0^{-1} \sum_{j=1}^{d_0} X_{tj} + 0.1 \times \varepsilon_{tk}$ for every $d_0 < k \leq d$, where ε_{tk} is a standard normal random variable. According to Theorem 1, the stocks X_{tk} with $d_0 < k \leq d$ are irrelevant for risk minimization.

In each of the above settings, 500 data sets are generated with sample sizes $n = 120$, 240 and 600. In each realization, we obtain $\hat{\mathcal{S}}$ via BIC as well as the five other selection criteria. Afterwards, we compute for each corresponding size $|\hat{\mathcal{S}}|$ and the resulting portfolio weights $\hat{\omega}_{\hat{\mathcal{S}}} = \hat{\Sigma}_{(\hat{\mathcal{S}})}^{-1} \mathbf{1}_{(\hat{\mathcal{S}})} \{\mathbf{1}_{(\hat{\mathcal{S}})}^\top \hat{\Sigma}_{(\hat{\mathcal{S}})}^{-1} \mathbf{1}_{(\hat{\mathcal{S}})}\}^{-1} \in \mathbb{R}^{|\hat{\mathcal{S}}|}$. Accordingly, the out-of-sample risks are given by $\hat{\sigma} = (\hat{\omega}_{\hat{\mathcal{S}}}^\top \Sigma_{(\hat{\mathcal{S}})} \hat{\omega}_{\hat{\mathcal{S}}})^{1/2}$. To assess the accuracy of BIC, SHORT and LASSO in portfolio selection, we further compute the correct fit (CF = $I(\hat{\mathcal{S}} = \mathcal{S}_0)$), the true rate (TR=$|\hat{\mathcal{S}} \bigcap \mathcal{S}_0|/|\mathcal{S}_0|$), and the false rate (FR=$|\hat{\mathcal{S}} \bigcap \mathcal{S}_0^c|/|\mathcal{S}_0^c|$). Note that TR examines model underfitting, whereas FR measures model overfitting. Finally, the average for each of the five performance measures across the 500 simulation realizations are reported.

We begin by assessing the finite sample performance of BIC and then comparing its selection ability with SHORT and LASSO. Table 1 demonstrates that, for a fixed (d_0, d)-specification, both CF and TR of BIC steadily increase as the sample size n increases, while FR steadily decreases. In fact, our unreported simulation results show that CF can be close to 100% when the sample size is sufficiently large. All of the findings confirm that BIC is a consistent criterion, as stated in Theorem 2. Accordingly, the portfolio size $\hat{\mathcal{S}}$ selected by BIC is close to that of the true model. Table 1 also indicates that BIC is considerably better than SHORT via the underfitting measure TR, while it is somewhat inferior to SHORT via the overfitting measure FR in the case of $(d_0, d) = (15, 90)$. Moreover, although LASSO underfits slightly less often than BIC, it overfits substantially more often than BIC. As a result, BIC outperforms SHORT and LASSO in identifying the true model correctly. In sum, BIC performs best.

In addition to examining portfolio selection, we next compare BIC with the other five approaches via the out-of-sample risk (i.e., $\hat{\sigma}$). Because the simulation results across different sample sizes are qualitatively similar, we mainly focus on the results with $n = 120$. Table 1 shows that NAIVE and MU perform poorly, and SHORT is better than both of them. It also indicates that MV and LASSO are superior to the above three methods. However, BIC outperforms MV by a considerable margin. The relative improvement can be as large as $(5.97 - 3.78)/3.78 = 57.9\%$ when $d = 90$. It is remarkable that such outstanding performance is achieved with a very small portfolio size of 19.31, which accounts for only $19.31/90 = 21.5\%$ of the total candidate stocks. Moreover, LASSO is less competitive than BIC, although it performs better than MV. Interestingly, LASSO selects the larger portfolio to attain the minimum risk, while SHORT might choose the smaller portfolio to mitigate risk. In conclusion, BIC balances risk diversification and estimation error, which enables it to perform best.

3.2. An Empirical Example

We next consider a real dataset that records the monthly return (i.e., monthly closing prices) of 100 US risky assets named "100 Portfolios Formed on Size and Book-to-Market Ratio" collected from K. French's online data library. After eliminating the assets with missing values, we obtain a sample of 87 risky assets during the period of 01/1969–12/1998. The total number of observations is $N = 30 \times 12 = 360$. For the sake of convenience, we denote returns by $\{X_t : 121 - N \leq t \leq 120\}$, where $t = 0$ corresponds to 12/1988. Then, we employ $\{X_t : -n + 1 \leq t \leq 0\}$ as the in-sample training data with $n = 120$ and 240. Moreover, the last 10 years of data ($\{X_t : 1 \leq t \leq 120\}$) are reserved for out-of-sample testing. It is also noteworthy that the return in the training data is centralized. Following DeMiguel et al. (2009a,b), we assume that the investor updates the portfolio weights in a monthly manner. Specifically, the investor selects an

optimal portfolio (i.e., $\hat{\mathcal{S}}$) based on the training data $\{X_{1-t} : 1 \leq t \leq n\}$, where n is the sample size. Once $\hat{\mathcal{S}}$ is selected, it is fixed throughout the testing period (i.e., 01/1989–12/1998). However, for every observation in the testing period (i.e., $1 \leq t \leq 120$), we obtain an updated optimal portfolio weight based on the mostly recent n observations, i.e., $\{X_{t-s} : 1 \leq s \leq n\}$. Accordingly, the resulting estimate $\hat{\omega}^t_{\hat{\mathcal{S}}}$ yields its realized return $r_t = X_t^\top \hat{\omega}^t_{\hat{\mathcal{S}}}$ for $1 \leq t \leq 120$.

To assess the out-of-sample performance of the selected portfolio, we consider three performance measures, namely the portfolio size $|\hat{\mathcal{S}}|$, the sample mean ($\hat{\mu}^*$) of r_t, and the sample standard deviation ($\hat{\sigma}^*$) of r_t ($1 \leq t \leq 120$). Because the portfolio weights are updated periodically, we further compute the transaction cost explained below. Let the portfolio weight be $\hat{\omega}^t_{\hat{\mathcal{S}}} = (\hat{\omega}^t_{j,\hat{\mathcal{S}}} : j \in \hat{\mathcal{S}}) \in \mathbb{R}^{|\hat{\mathcal{S}}|}$ for $1 \leq t \leq 120$. After the portfolio is held for a month, the actual weight assigned to the jth stock has changed from $\hat{\omega}^t_{j,\hat{\mathcal{S}}}$ to $\hat{\omega}^t_{j,\hat{\mathcal{S}}}(1 + X_{tj})$ for every $j \in \hat{\mathcal{S}}$, while the updated portfolio weight becomes $\hat{\omega}^{t+1}_{j,\hat{\mathcal{S}}}$ obtained via the newly updated data. To adjust the portfolio weight from $\hat{\omega}^t_{j,\hat{\mathcal{S}}}(1+X_{tj})$ to $\hat{\omega}^{t+1}_{j,\hat{\mathcal{S}}}$, the amount of change is $|\hat{\omega}^t_{j,\hat{\mathcal{S}}}(1 + X_{tj}) - \hat{\omega}^{t+1}_{j,\hat{\mathcal{S}}}|$. Hence, the total change across the entire portfolio $\hat{\mathcal{S}}$ is $\sum_j |\hat{\omega}^t_{j,\hat{\mathcal{S}}}(1 + X_{tj}) - \hat{\omega}^{t+1}_{j,\hat{\mathcal{S}}}|$. As a result, the average change (i.e., Average Turnover, DeMiguel et al. (2009a,b)) across the entire testing period is

$$\mathrm{AT} = \frac{1}{119} \sum_{t=1}^{119} \sum_{j=1}^{d} \left| \hat{\omega}^t_{j,\hat{\mathcal{S}}}(1 + X_{tj}) - \hat{\omega}^{t+1}_{j,\hat{\mathcal{S}}} \right|.$$

Using $\hat{\mu}^*$, $\hat{\sigma}^*$, and AT, we define the last performance measure, the Sharpe Ratio (Sharpe et al., 2001), which is $\mathrm{SR} = (\hat{\mu}^* - \gamma \times \mathrm{AT})/\hat{\sigma}^*$, where γ is the parameter that characterizes the transaction cost. For the sake of comparison, we consider three different values of γ, which correspond to low ($\gamma = 0.25\%$), median ($\gamma = 0.5\%$), and high ($\gamma = 0.75\%$) transaction costs, respectively, although $\gamma = 0.5\%$ seems to be a reasonable choice (Blanchett, 2007).

Table 2 presents six portfolio strategies with seven performance measures. Because NAIVE's diversification is independent of the sample size, its associated performance measures are identical across the two different sample sizes (i.e., $n = 120$ and $n = 240$). In addition, the portfolio weights of NAIVE are constant. Hence, it is not surprising that NAIVE yields the smallest AT, and its Sharpe Ratio is above 30%. Our results corroborate the empirical findings of DeMiguel et al. (2009b). We next consider the MV approach. Although it minimizes the in-sample risk, Table 2 shows that MV fails to extend its outstanding in-sample performance to the out-of-sample validation. Its standard deviation is larger than that of NAIVE when $n = 120$. Moreover, the average turnover of MV is appreciable. Accordingly, if the sample size is small (i.e., $n = 120$) or the transaction cost is high (i.e., $\gamma = 0.75\%$), then MV's Sharpe Ratios are worse than those of NAIVE; see DeMiguel et al. (2009b) for a useful discussion.

In contrast to MV, MU yields the largest standard deviation although its average return is very high when $n = 240$. This is because the mean return in an efficient capital market (e.g., the US market) is difficult to estimate accurately and the resulting estimation errors lead to unstable portfolio weights. Accordingly, the associated turnover and transaction costs are very high. See Jorion (1986) for an excellent discussion and references therein. It is, therefore, not surprising to find that MU's out-of-sample Sharpe Ratios in Table 2 are very low.

According to the theoretical analysis of Jagannathan and Ma (2003), imposing the short sale constraint in the portfolio optimization process can lead to risk reduction. Table 2 confirms their findings and indicates that SHORT has the third smallest risk, $\hat{\sigma}^*$. However, its mean returns are considerably worse than those of BIC. As a result, the Sharpe Ratio measures are not the best. It is of interest to note that LASSO yields the best out-of-sample risk performance. However, its average turnover is higher

that of BIC. This is mainly because LASSO's size is larger than the size of BIC. Hence, the resulting Sharpe Ratio measures of LASSO are inferior to those of BIC. This is particularly true when the transaction cost is high. The above findings are more discernible when the sample size is 120. Furthermore, Table 2 indicates that BIC yields the second smallest risk across the two different sample sizes, while yielding the highest and second highest returns, respectively, for $n = 120$ and 240. Moreover, the sizes of BIC are small (i.e., 11 for $n = 120$ and 8 for $n = 240$) so that the resulting AT values are reasonable.

Finally, Table 3 presents the pairwise comparison between BIC and each of the other five portfolio methods via three measures. The first measure is $\tilde{\mu}$, which stands for the difference between the BIC portfolio and one of its competing methods (e.g., Naive) in terms of the out-of-sample mean return. The second is $\tilde{\sigma}$; it reflects the difference in standard deviation. The third measure is SR, which represents the difference in terms of Sharpe Ratio with $\gamma = 0.50\%$. Based on these three measures, BIC generally performs better than the rest of methods in terms of both the mean return and standard deviation measures. In particular, the Sharpe Ratio differences between BIC and each of the other methods are statistically significant at the 5% level. Consequently, BIC performs best in Sharpe Ratios.

4. CONCLUSIONS

In portfolio selection, we introduce relevant and irrelevant stocks to balance diversification and estimation error. In addition, we propose the linkage condition to establish a connection between the relevant and irrelevant stocks, which allows us to obtain a Bayesian information criterion. To broaden the usefulness of this proposed strategy, we identify three avenues for future research. The first is to adopt the ideas of Bayesian model averaging (Hoeting et al., 1999) and adaptive mixing (Yang, 2001)

to mitigate the instability of portfolio selection. Thus, the resulting strategy would take into account not only the estimation error but also the selection process. The second potential avenue would be extending Bayesian information criterion to study high dimensional portfolios with $d \gg n$ (Wang, 2009). The third avenue would be to relax the independence assumption, which was used to facilitate the theoretical developments of BIC; see also (Kan and Zhou, 2007) and Fan et al (2008). This area of research is particularly important given the high degree of serial dependence found in stock returns; see DeMiguel et al. (2009a) for a recent discussion of this issue). We believe efforts in these areas would strengthen the field of portfolio selection.

APPENDIX: TECHNICAL DETAILS

Appendix A. The Proof of Theorem 1

SUFFICIENCY. To show sufficiency, we need to demonstrate that including any additional stocks in the portfolio would increase the portfolio risk, i,e., the risk of the new portfolio is larger than $\sigma_{\mathcal{S}}^2 = (\mathbf{1}_{(\mathcal{S})}\Sigma_{(\mathcal{S})}^{-1}\mathbf{1}_{(\mathcal{S})})^{-1}$. To this end, we consider an arbitrary portfolio weight vector $\omega \in \mathbb{R}^d$, and then evaluate its corresponding risk as follows,

$$\mathrm{var}\Big(\omega^\top X_t\Big) = \mathrm{var}\Big(\omega_{(\mathcal{S})}^\top X_{t(\mathcal{S})} + \omega_{(\mathcal{S}^c)}^\top X_{t(\mathcal{S}^c)}\Big),$$

where $X_{t(\mathcal{S}^c)}$ and $\omega_{(\mathcal{S}^c)}$ are the complements of $X_{t(\mathcal{S})}$ and $\omega_{(\mathcal{S})}$, respectively. It is noteworthy that, under the normality assumption of X_t, we are able to show that $X_{tk} = \sum_{j\in\mathcal{S}} X_{tj}\beta_{kj} + \varepsilon_{tk}$ for any $k \notin \mathcal{S}$, where β_{kj} are some regression coefficients, ε_{tk} is a normally distributed random variable with mean zero and positive variance, and ε_{tk} is independent of $X_{t(\mathcal{S})}$. Accordingly, we have

$$\omega_{(\mathcal{S}^c)}^\top X_{t(\mathcal{S}^c)} = \omega_{(\mathcal{S}^c)}^\top B_{\mathcal{S}^c|\mathcal{S}}^\top X_{t(\mathcal{S})} + \omega_{(\mathcal{S}^c)}^\top \varepsilon_{t\mathcal{S}^c},$$

where $B_{\mathcal{S}^c|\mathcal{S}}^\top = \{\beta_{kj} : k \in \mathcal{S}^c, j \in \mathcal{S}\} \in \mathbb{R}^{|\mathcal{S}^c|\times|\mathcal{S}|}$ and $\varepsilon_{t\mathcal{S}^c} = \{\varepsilon_{tk} : k \in \mathcal{S}^c\} \in \mathbb{R}^{|\mathcal{S}^c|}$. Subsequently,

$$\begin{aligned}\mathrm{var}\Big(\omega^\top X_t\Big) &= \mathrm{var}\Big\{\Big(\omega_{(\mathcal{S})}^\top + \omega_{(\mathcal{S}^c)}^\top B_{\mathcal{S}^c|\mathcal{S}}^\top\Big)X_{t(\mathcal{S})} + \omega_{(\mathcal{S}^c)}^\top \varepsilon_{t\mathcal{S}^c}\Big\} \\ &= \mathrm{var}\Big\{\Big(\omega_{(\mathcal{S})}^\top + \omega_{(\mathcal{S}^c)}^\top B_{\mathcal{S}^c|\mathcal{S}}^\top\Big)X_{t(\mathcal{S})}\Big\} + \mathrm{var}\Big(\omega_{(\mathcal{S}^c)}^\top \varepsilon_{t\mathcal{S}^c}\Big). \qquad \text{(A.1)}\end{aligned}$$

Under the condition of $B_{\mathcal{S}^c|\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})} = \mathbf{1}_{(\mathcal{S}^c)}$, we obtain

$$\Big(\omega_{(\mathcal{S})}^\top + \omega_{(\mathcal{S}^c)}^\top B_{\mathcal{S}^c|\mathcal{S}}^\top\Big)\mathbf{1}_{(\mathcal{S})} = \omega_{(\mathcal{S})}^\top \mathbf{1}_{(\mathcal{S})} + \omega_{(\mathcal{S}^c)}^\top \mathbf{1}_{(\mathcal{S}^c)} = 1.$$

Hence, $\omega_{\mathcal{S}}^{*\top} = \omega_{(\mathcal{S})}^\top + \omega_{(\mathcal{S}^c)}^\top B_{\mathcal{S}^c|\mathcal{S}}^\top$ is also a portfolio weight (i,e., $\omega_{\mathcal{S}}^{*\top}\mathbf{1}_{(\mathcal{S})} = 1$). This, together with (A.1), implies that the risk of $\omega^\top X_t$ is strictly larger than that of $\omega_{\mathcal{S}}^{*\top} X_{t(\mathcal{S})}$, unless $\omega_{(\mathcal{S}^c)} = 0$. In other words, if $\omega_{(\mathcal{S}^c)} \neq 0$, we must have $\mathrm{var}(\omega^\top X_t) > \mathrm{var}(\omega_{\mathcal{S}}^{*\top} X_{t(\mathcal{S})}) \geq \sigma_{\mathcal{S}}^2$, where the last inequality is due to the definition of $\sigma_{\mathcal{S}}^2$. Consequently, portfolio $\mathcal{S}$ is sufficient for risk minimization (i.e, $\mathcal{S} \supset \mathcal{S}_0$), and the proof for sufficiency is complete.

Necessity. Assume $\sum \beta_{kj} \neq 1$ for some $k \notin \mathcal{S}$, and we then show that $\mathcal{S}$ is not sufficient for the risk minimization. We define $\mathcal{S}^* = \mathcal{S}\bigcup\{k\}$, $X_t^* = X_{t(\mathcal{S}^*)}$, and $\omega^* = \{(1-\omega_k)\omega_{0\mathcal{S}}^\top, \omega_k)^\top \in \mathbb{R}^{|\mathcal{S}|+1}$, where ω_k is the weight assigned to the kth stock. We next compute the variance of the portfolio $\mathcal{S}^*$ below. After algebraic simplification,

$$\begin{aligned}\mathrm{var}\Big(\omega^{*\top} X_t^*\Big) &= \Big\{\beta_{k,\mathcal{S}}^\top \Sigma_{(\mathcal{S})}\beta_{k,\mathcal{S}} + \mathrm{var}(\varepsilon_{tk}) + \sigma_{\mathcal{S}}^2 - 2\sigma_{\mathcal{S}}^2 \beta_{k,\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})}\Big\}\omega_k^2 - 2\sigma_{\mathcal{S}}^2\Big(1 - \beta_{k,\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})}\Big)\omega_k + \sigma_{\mathcal{S}}^2 \\ &= \Big\{\beta_{k,\mathcal{S}}^\top \Sigma_{(\mathcal{S})}^{1/2}\Big(I_{(\mathcal{S})} - M_{(\mathcal{S})}\Big)\Sigma_{(\mathcal{S})}^{1/2}\beta_{k,\mathcal{S}} + \sigma_{\mathcal{S}}^2\Big(1 - \beta_{k,\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})}\Big)^2 + \mathrm{var}(\varepsilon_{tk})\Big\}\omega_k^2 \\ &\qquad -2\sigma_{\mathcal{S}}^2\Big(1 - \beta_{k,\mathcal{S}}^\top \mathbf{1}_{(\mathcal{S})}\Big)\omega_k + \sigma_{\mathcal{S}}^2, \qquad \text{(A.2)}\end{aligned}$$

where $I_{(\mathcal{S})}$ is the $|\mathcal{S}|\times|\mathcal{S}|$ identity matrix and $M_{(\mathcal{S})} = \Sigma_{(\mathcal{S})}^{-1/2}\mathbf{1}_{(\mathcal{S})}(\mathbf{1}_{(\mathcal{S})}^\top\Sigma_{(\mathcal{S})}^{-1}\mathbf{1}_{(\mathcal{S})})^{-1}\mathbf{1}_{(\mathcal{S})}^\top\Sigma_{(\mathcal{S})}^{-1/2}$. Because the coefficient of ω_k^2 is positive, we obtain the minimum of $\mathrm{var}(\omega^{*\top}X_t^*)$ at

$$\hat{\omega}_k = \frac{\sigma_{\mathcal{S}}^2(1-\beta_{k,\mathcal{S}}^\top\mathbf{1}_{(\mathcal{S})})}{\beta_{k,\mathcal{S}}^\top\Sigma_{(\mathcal{S})}\beta_{k,\mathcal{S}}^\top + \mathrm{var}(\varepsilon_{tk}) + \sigma_{\mathcal{S}}^2 - 2\sigma_{\mathcal{S}}^2\beta_{k,\mathcal{S}}^\top\mathbf{1}_{(\mathcal{S})}}.$$

Under the condition $\sum\beta_{kj}\neq 1$, we have that $\hat{\omega}_k\neq 0$. However, (A.2) indicates that $\omega_k = 0$ leads to $\sigma_{\mathcal{S}}^2$. This implies that $\mathrm{var}(\hat{\omega}^{*\top}X_t^*) < \sigma_{\mathcal{S}}^2$, where $\hat{\omega}^* = (\hat{\omega}_{\mathcal{S}}^\top, \hat{\omega}_k)^\top$ and $\hat{\omega}_{\mathcal{S}} = \big(1-\hat{\omega}_k\big)\omega_{0\mathcal{S}}$. Consequently, the risk of $\mathcal{S}$ can be further reduced by including the kth stock into the portfolio. This completes the proof.

Appendix B. The Proof of Theorem 2

For any candidate portfolio $\mathcal{S}$, we define $\mathscr{Q}_+ = \{\mathcal{S} : \mathcal{S}\supset\mathcal{S}_0, \mathcal{S}\neq\mathcal{S}_0\}$ and $\mathscr{Q}_- = \{\mathcal{S} : \mathcal{S}\not\supset\mathcal{S}_0\}$. To prove the theorem, we consider two different cases, namely overfitted portfolios and underfitted portfolios, as given below.

Case 1. Overfitted (i.e., $\mathcal{S}\in\mathscr{Q}_+$). The joint likelihood function of (X_t) for the overfitted portfolio $\mathcal{S}$ is $\ell(\Sigma_{(\mathcal{S})},\Sigma_{(\mathcal{S}^c)}) = \ell(\Sigma_{(\mathcal{S})})\ell(\Sigma_{\mathcal{S}^c|\mathcal{S}})$, where $\ell(\Sigma_{(\mathcal{S})})$ and $\ell(\Sigma_{\mathcal{S}^c|\mathcal{S}})$ are given in equations (2.2) and (2.3), respectively. Because $\mathcal{S}$ is overfitted, we can further decompose the marginal likelihood function as follows,

$$\ell(\Sigma_{(\mathcal{S})}) = \ell\Big(\Sigma_{(\mathcal{S}_0)}\Big)\ell\Big(\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big),$$

where

$$\log\ell\Big(\Sigma_{(\mathcal{S}_0)}\Big) = -\frac{1}{2}\Big(n\log|\Sigma_{(\mathcal{S}_0)}| + \sum_{t=1}^n X_{t(\mathcal{S}_0)}^\top\Sigma_{(\mathcal{S}_0)}^{-1}X_{t(\mathcal{S}_0)}\Big), \tag{A.3}$$

$$\log\ell\Big(\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big) = -\frac{1}{2}\Bigg\{n\log\Big|\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big|$$

$$+\sum_{t=1}^n\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - B_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}^\top X_{t(\mathcal{S}_0)}\Big)^\top\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}^{-1}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - B_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}^\top X_{t(\mathcal{S}_0)}\Big)\Bigg\}, \tag{A.4}$$

$B^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}$ is a $|\mathcal{S}\setminus\mathcal{S}_0| \times |\mathcal{S}|$ matrix with unknown regression coefficients, and some irrelevant constants are omitted from equations (A.3) and (A.4). Accordingly,

$$\ell(\Sigma_{(\mathcal{S})}, \Sigma_{(\mathcal{S}^c)}) = \ell\Big(\Sigma_{(\mathcal{S}_0)}\Big)\ell\Big(\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big)\ell\Big(\Sigma_{\mathcal{S}^c|\mathcal{S}}\Big). \tag{A.5}$$

Applying the first two likelihood functions on the right-hand side of (A.5), we obtain their corresponding maximum likelihood estimators,

$$\hat{\Sigma}_{(\mathcal{S}_0)} = n^{-1}\sum_{t=1}^{n} X_{t(\mathcal{S}_0)}X^{\top}_{t(\mathcal{S}_0)}, \tag{A.6}$$

and

$$\hat{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)^{\top}, \tag{A.7}$$

where $\hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = (\mathbb{X}^{\top}_{(\mathcal{S}_0)}\mathbb{X}_{(\mathcal{S}_0)})^{-1}(\mathbb{X}^{\top}_{(\mathcal{S}_0)}\mathbb{X}_{(\mathcal{S}\setminus\mathcal{S}_0)})$. Furthermore, maximizing the third term on the right-hand side of (A.5) with the constraint $B^{\top}_{\mathcal{S}^c|\mathcal{S}}\mathbf{1}_{(\mathcal{S})} = \mathbf{1}_{(\mathcal{S}^c)}$ obtained from Theorem 1 yields the constrained maximum likelihood estimator, $\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}}$, which is given on equation (2.4). After algebraic simplification and then omitting the irrelevant terms, the $-2\log/n$ of the maximum of $\ell(\Sigma_{(\mathcal{S})}, \Sigma_{(\mathcal{S}^c)})$ is

$$\begin{aligned}\mathcal{L}(\mathcal{S}) &= -2\log\Big[\ell\Big(\hat{\Sigma}_{(\mathcal{S}_0)}\Big)\ell\Big(\hat{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big)\ell\Big(\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}}\Big)\Big]/n \\ &= \log\big|\hat{\Sigma}_{(\mathcal{S}_0)}\big| + \log\big|\hat{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\big| + \log\big|\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}}\big|. \end{aligned} \tag{A.8}$$

We next consider the joint likelihood function of X_t for the optimal portfolio $\mathcal{S}_0$, $\ell(\Sigma_{(\mathcal{S}_0)}, \Sigma_{(\mathcal{S}_0^c)}) = \ell(\Sigma_{(\mathcal{S}_0)})\ell(\Sigma_{\mathcal{S}_0^c|\mathcal{S}_0})$, where $\ell(\Sigma_{(\mathcal{S}_0)})$ and $\ell(\Sigma_{\mathcal{S}_0^c|\mathcal{S}_0})$ are given in equations (2.2) and (2.3), respectively, except for replacing $\mathcal{S}$ with $\mathcal{S}_0$. Because $\mathcal{S} \supset \mathcal{S}_0$ and

$\mathcal{S} \neq \mathcal{S}_0$, we can further decompose the conditional likelihood function as follows,

$$\ell(\Sigma_{\mathcal{S}_0^c|\mathcal{S}_0}) = \ell\Big(\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big)\ell\Big(\Sigma_{\mathcal{S}^c|\mathcal{S}}\Big).$$

Accordingly,

$$\ell(\Sigma_{(\mathcal{S}_0)}, \Sigma_{(\mathcal{S}_0^c)}) = \ell\Big(\Sigma_{(\mathcal{S}_0)}\Big)\ell\Big(\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big)\ell\Big(\Sigma_{\mathcal{S}^c|\mathcal{S}}\Big). \tag{A.9}$$

It is noteworthy that (A.9) represents the joint likelihood function of (X_t) for the portfolio $\mathcal{S}_0$, while (A.5) is the joint likelihood function of (X_t) for the portfolio $\mathcal{S}$. Hence, the resulting parameter estimator of $\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}$ obtained from the second term on the right-hand side of the above equation is different from that obtained from (A.5), although the other two parameter estimators of $\Sigma_{(\mathcal{S}_0)}$ and $\Sigma_{\mathcal{S}^c|\mathcal{S}}$ are the same as those given in equations (A.6) and (2.4), respectively. Maximizing the conditional likelihood function $\ell\Big(\Sigma_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big)$ with the constraint $B_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}^\top \mathbf{1}_{(\mathcal{S}_0)} = \mathbf{1}_{(\mathcal{S}\setminus\mathcal{S}_0)}$ obtained from Theorem 1 leads to the constrained maximum likelihood estimator

$$\tilde{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \tilde{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}^\top X_{t(\mathcal{S}_0)}\Big)\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \tilde{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}^\top X_{t(\mathcal{S}_0)}\Big)^\top, \tag{A.10}$$

where

$$\tilde{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = \Big(\mathbb{X}_{(\mathcal{S}_0)}^\top \mathbb{X}_{(\mathcal{S}_0)}\Big)^{-1}\Big(\mathbb{X}_{(\mathcal{S}_0)}^\top \mathbb{X}_{(\mathcal{S}\setminus\mathcal{S}_0)} - \mathbf{1}_{(\mathcal{S})}\tilde{\lambda}_{(\mathcal{S}\setminus\mathcal{S}_0)}^\top\Big)$$

and

$$\tilde{\lambda}_{(\mathcal{S}\setminus\mathcal{S}_0)}^\top = \Big\{\mathbf{1}_{(\mathcal{S}_0)}^\top\Big(\mathbb{X}_{(\mathcal{S}_0)}^\top \mathbb{X}_{(\mathcal{S}_0)}\Big)^{-1}\mathbb{X}_{(\mathcal{S}_0)}^\top \mathbb{X}_{(\mathcal{S}\setminus\mathcal{S}_0)} - \mathbf{1}_{(\mathcal{S}\setminus\mathcal{S}_0)}^\top\Big\}\Big\{\mathbf{1}_{(\mathcal{S}_0)}^\top\Big(\mathbb{X}_{(\mathcal{S}_0)}^\top \mathbb{X}_{(\mathcal{S}_0)}\Big)^{-1}\mathbf{1}_{(\mathcal{S}_0)}\Big\}^{-1}.$$

After algebraic simplification and ignoring the irrelevant terms, the $-2\log/n$ of the

maximum of $\ell(\Sigma_{(\mathcal{S}_0)}, \Sigma_{(\mathcal{S}_0^c)})$ is

$$\mathcal{L}(\mathcal{S}_0) = -2\log\Big[\ell\Big(\hat{\Sigma}_{(\mathcal{S}_0)}\Big)\ell\Big(\tilde{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\Big)\ell\Big(\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}}\Big)\Big]/n$$

$$= \log\big|\hat{\Sigma}_{(\mathcal{S}_0)}\big| + \log\big|\tilde{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\big| + \log\big|\tilde{\Sigma}_{\mathcal{S}^c|\mathcal{S}}\big|. \tag{A.11}$$

Applying (A.7), (A.8), (A.10), (A.11), and BIC in (2.5), we obtain that

$$\mathrm{BIC}(\mathcal{S}) - \mathrm{BIC}(\mathcal{S}_0) = \Big\{\mathcal{L}(\mathcal{S}) - \mathcal{L}(\mathcal{S}_0)\Big\} + \Big(|\mathcal{S}| - |\mathcal{S}_0|\Big) \times \log(n)/n$$

$$\geq \Big\{\mathcal{L}(\mathcal{S}) - \mathcal{L}(\mathcal{S}_0)\Big\} + \log(n)/n$$

$$= \log\Big|n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)^{\top}\Big|$$

$$- \log\Big|n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \tilde{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \tilde{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)^{\top}\Big|$$

$$+ \log(n)/n. \tag{A.12}$$

Following the definitions of $\hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}$ and $\tilde{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}$, we have that

$$\tilde{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = \hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}$$

$$-\Big(\mathbb{X}^{\top}_{(\mathcal{S}_0)}\mathbb{X}_{(\mathcal{S}_0)}\Big)^{-1}\mathbf{1}_{(\mathcal{S}_0)}\Big\{\mathbf{1}^{\top}_{(\mathcal{S}_0)}\Big(\mathbb{X}^{\top}_{(\mathcal{S}_0)}\mathbb{X}_{(\mathcal{S}_0)}\Big)^{-1}\mathbf{1}_{(\mathcal{S}_0)}\Big\}^{-1}$$

$$\times\Big(\mathbf{1}^{\top}_{(\mathcal{S}_0)}\hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} - \mathbf{1}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)}\Big). \tag{A.13}$$

Furthermore, $\hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} - B_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = O_p(n^{-1/2})$. This together with $B^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\mathbf{1}_{(\mathcal{S}_0)} = \mathbf{1}_{(\mathcal{S}\setminus\mathcal{S}_0)}$ leads to $\mathbf{1}^{\top}_{(\mathcal{S}_0)}\hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} - \mathbf{1}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)} = O_p(n^{-1/2})$. Moreover

$$\Big(\mathbb{X}^{\top}_{(\mathcal{S}_0)}\mathbb{X}_{(\mathcal{S}_0)}\Big)^{-1}\mathbf{1}_{(\mathcal{S}_0)}\Big\{\mathbf{1}^{\top}_{(\mathcal{S}_0)}\big(\mathbb{X}^{\top}_{(\mathcal{S}_0)}\mathbb{X}_{(\mathcal{S}_0)}\big)^{-1}\mathbf{1}_{(\mathcal{S}_0)}\Big\}^{-1} \to_p \Sigma^{-1}_{(\mathcal{S}_0)}\mathbf{1}_{(\mathcal{S}_0)}\Big(\mathbf{1}^{\top}_{(\mathcal{S}_0)}\Sigma^{-1}_{(\mathcal{S}_0)}\mathbf{1}_{(\mathcal{S}_0)}\Big)^{-1},$$

as $n \to \infty$. Substituting the above results into (A.13), we have that $\tilde{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = \hat{B}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} + O_p(n^{-1/2})$. Hence,

$$\tilde{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} = n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \tilde{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \tilde{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)^{\top}$$

$$= n^{-1}\sum_{t=1}^{n}\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)\Big(X_{t(\mathcal{S}\setminus\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)^{\top}$$

$$+n^{-1}\sum_{t=1}^{n}\Big(\tilde{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)\Big(\tilde{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)} - \hat{B}^{\top}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}X_{t(\mathcal{S}_0)}\Big)^{\top}$$

$$= \hat{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} + O_p(1/n), \tag{A.14}$$

where the cross term in the derivation of the above equation vanishes and is omitted. Consequently,

$$\mathrm{BIC}(\mathcal{S}) - \mathrm{BIC}(\mathcal{S}_0) \geq \log\left|\hat{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0}\right| - \log\left|\hat{\Sigma}_{(\mathcal{S}\setminus\mathcal{S}_0)|\mathcal{S}_0} + O_p(1/n)\right| + \log(n)/n$$

$$\geq O_p(1/n) + \log(n)/n,$$

which indicates that, with probability tending to 1, $\mathrm{BIC}(\mathcal{S}) > \mathrm{BIC}(\mathcal{S}_0)$ for any overfitted portfolio $\mathcal{S}$. Thus, we conclude that $P\{\inf_{\mathcal{S}\in\mathscr{Q}_+}\mathrm{BIC}(\mathcal{S}) > \mathrm{BIC}(\mathcal{S}_0)\} \to 1$.

Case 2. Underfitted (i.e., $\mathcal{S} \in \mathscr{Q}_-$). Define $\mathcal{S}^* = \mathcal{S}\bigcup\mathcal{S}_0$. Then, applying the same techniques as used in obtaining (A.5), we have the joint likelihood function of X_t for the portfolio $\mathcal{S}^*$ as follows,

$$\ell(\Sigma_{(\mathcal{S}^*)}, \Sigma_{(\mathcal{S}^{*c})}) = \ell(\Sigma_{(\mathcal{S}^*)})\ell(\Sigma_{\mathcal{S}^{*c}|\mathcal{S}^*})$$

$$= \ell\Big(\Sigma_{(\mathcal{S})}\Big) \times \ell\Big(\Sigma_{(\mathcal{S}^*\setminus\mathcal{S})|\mathcal{S}}\Big) \times \ell\Big(\Sigma_{(\mathcal{S}^*)^c|\mathcal{S}^*}\Big), \tag{A.15}$$

where

$$\log \ell\Big(\Sigma_{(\mathcal{S})}\Big) = -\frac{1}{2}\Big(n \log |\Sigma_{(\mathcal{S})}| + \sum_{t=1}^{n} X_{t(\mathcal{S})}^{\top} \Sigma_{(\mathcal{S})}^{-1} X_{t(\mathcal{S})}\Big),$$

$$\log \ell\Big(\Sigma_{(\mathcal{S}^*\setminus\mathcal{S})|\mathcal{S}}\Big) = -\frac{1}{2}\Bigg(n \log \Big|\Sigma_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\Big|$$

$$+ \sum_{t=1}^{n} \Big(X_{t(\mathcal{S}^*\setminus\mathcal{S})} - B_{(\mathcal{S}^*\setminus\mathcal{S})|\mathcal{S}}^{\top} X_{t(\mathcal{S})}\Big)^{\top} \Sigma_{(\mathcal{S}^*\setminus\mathcal{S})|\mathcal{S}}^{-1} \Big(X_{t(\mathcal{S}^*\setminus\mathcal{S})} - B_{(\mathcal{S}^*\setminus\mathcal{S})|\mathcal{S}}^{\top} X_{t(\mathcal{S})}\Big)\Bigg),$$

$$\log \ell\Big(\Sigma_{\mathcal{S}^{*c}|\mathcal{S}^*}\Big) = -\frac{1}{2}\Bigg(n \log \Big|\Sigma_{\mathcal{S}^{*c}|\mathcal{S}^*}\Big|$$

$$+ \sum_{t=1}^{n} \Big(X_{t(\mathcal{S}^{*c})} - B_{\mathcal{S}^{*c}|\mathcal{S}^*}^{\top} X_{t(\mathcal{S}^*)}\Big)^{\top} \Sigma_{\mathcal{S}^{*c}|\mathcal{S}^*}^{-1} \Big(X_{t(\mathcal{S}^{*c})} - B_{\mathcal{S}^{*c}|\mathcal{S}^*}^{\top} X_{t(\mathcal{S}^*)}\Big)\Bigg),$$

and some irrelevant constants are omitted from the above equations. Following the same discussion as given in (A.8), we obtain that

$$\mathcal{L}(\mathcal{S}^*) = \log \Big|\hat{\Sigma}_{(\mathcal{S})}\Big| + \log \Big|\hat{\Sigma}_{(\mathcal{S}^*\setminus\mathcal{S})|\mathcal{S}}\Big| + \log \Big|\tilde{\Sigma}_{\mathcal{S}^{*c}|\mathcal{S}^*}\Big|,$$

where $\hat{\Sigma}_{(\mathcal{S})}$, $\hat{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}$, and $\tilde{\Sigma}_{\mathcal{S}^{*c}|\mathcal{S}^*}$ are defined as in equations (A.6), (A.7), and (2.4), except for replacing $\mathcal{S}_0$ and $\mathcal{S}$ in those equations with $\mathcal{S}$ and $\mathcal{S}^*$, respectively. Applying the same approach used in the derivation of equation (A.11), we have that

$$\mathcal{L}(\mathcal{S}) = \log \Big|\hat{\Sigma}_{(\mathcal{S})}\Big| + \log \Big|\tilde{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\Big| + \log \Big|\tilde{\Sigma}_{\mathcal{S}^{*c}|\mathcal{S}^*}\Big|,$$

where $\tilde{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}$ is defined as in (A.10), except for replacing $\mathcal{S}_0$ and $\mathcal{S}$ in that equation with $\mathcal{S}$ and $\mathcal{S}^*$, respectively.

We next follow the same technique as in the derivation of (A.14) and obtain that

$$\tilde{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}} = \hat{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}} + \hat{C}, \tag{A.16}$$

where

$$\hat{C} = \Big(\hat{B}^{\top}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\mathbf{1}_{(\mathcal{S})} - \mathbf{1}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\Big)\mathbf{1}^{\top}_{(\mathcal{S})}\hat{\Sigma}^{-1}_{(\mathcal{S})}\mathbf{1}_{(\mathcal{S})}\Big(\mathbf{1}^{\top}_{(\mathcal{S})}\hat{B}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}} - \mathbf{1}^{\top}_{(\mathcal{S}*\setminus\mathcal{S})}\Big).$$

Furthermore, with probability tending to 1

$$\hat{C} \to_p C = \Big(\ B^{\top}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\mathbf{1}_{(\mathcal{S})} - \mathbf{1}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\Big)\mathbf{1}^{\top}_{(\mathcal{S})}\Sigma^{-1}_{(\mathcal{S})}\mathbf{1}_{(\mathcal{S})}\Big(\mathbf{1}^{\top}_{(\mathcal{S})}\ B_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}} - \mathbf{1}^{\top}_{(\mathcal{S}*\setminus\mathcal{S})}\Big).$$

Because $\mathcal{S} \not\supset \mathcal{S}_0$, we have $\mathbf{1}^{\top}_{(\mathcal{S})}\ B_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}} \neq \mathbf{1}^{\top}_{(\mathcal{S}*\setminus\mathcal{S})}$. Hence, C is a non-zero positive semi-definite matrix. Thus, there exists a positive constant c_0 that satisfies

$$\log\left|\tilde{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\right| - \log\left|\hat{\Sigma}_{(\mathcal{S}*\setminus\mathcal{S})|\mathcal{S}}\right| \geq c_0.$$

As a result,

$$\mathrm{BIC}(\mathcal{S}) - \mathrm{BIC}(\mathcal{S}^*) = \Big\{\mathcal{L}(\mathcal{S}) - \mathcal{L}(\mathcal{S}^*)\Big\} - \Big(|\mathcal{S}^*| - |\mathcal{S}|\Big) \times \log(n)/n$$

$$\geq c_0 - \Big(|\mathcal{S}^*| - |\mathcal{S}|\Big) \times \log(n)/n \to_p \geq c_0.$$

Consequently, with probability tending to 1, we have that

$$\mathrm{BIC}(\mathcal{S}) > \mathrm{BIC}(\mathcal{S}^*) \tag{A.17}$$

for any underfitted portfolio $\mathcal{S}$.

Because $\mathcal{S}^* \supset \mathcal{S}_0$, we must have either $\mathcal{S}^* = \mathcal{S}_0$ or both $\mathcal{S}^* \supset \mathcal{S}_0$ and $\mathcal{S}^* \neq \mathcal{S}_0$. When $\mathcal{S}^* = \mathcal{S}_0$, $\mathrm{BIC}(\mathcal{S}^*) = \mathrm{BIC}(\mathcal{S}_0)$ which leads to $\mathrm{BIC}(\mathcal{S}) > \mathrm{BIC}(\mathcal{S}_0)$. In the case of $\mathcal{S}^* \supset \mathcal{S}_0$ and $\mathcal{S}^* \neq \mathcal{S}_0$, we employ the result of Case 1 and obtain $\mathrm{BIC}(\mathcal{S}^*) > \mathrm{BIC}(\mathcal{S}_0)$ with probability tending to 1. This together with (A.17) leads to $\mathrm{BIC}(\mathcal{S}) > \mathrm{BIC}(\mathcal{S}_0)$.

In sum, we obtain that $P(\inf_{\mathcal{S}\in\mathscr{Q}_-} \mathrm{BIC}(\mathcal{S}) > \mathrm{BIC}(\mathcal{S}_0)) \to 1$. The results of CASE 1 and CASE 2 imply that $P(\hat{\mathcal{S}}_{\mathrm{BIC}} = \mathcal{S}_0) \to 1$ as $n \to \infty$, and the proof is complete.

REFERENCES

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle." *Proceeding of the Second International Symposium on Information Theory* (eds B. N. Petrov and F. Csaki), 267-281. Budapest: Akademiai Kiado.

Bickel, P. J., and Levina, E. (2008), "Covariance regularization by thresholding," *The Annals of Statistics*, 36, 2577–2604.

Blanchett, D. (2007), "The pre-tax costs of portfolio turnover," *Journal of Indexes*, May/June, 35–39.

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

Chen, H. H., Tsai, H. T., and Lin, K. J. (2011), "Optimal mean-variance portfolio selection using Cauchy-Schwarz maximization," *Applied Economics*, 43, 1–7.

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

DeMiguel, V., Garlappi, L., and Uppal, R. (2009b), "Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy?" *Review of Financial Studies*, 22, 1915–1953.

DeMiguel, V., Uppal, R., and Nogales, F. J. (2010), "Stock return serial dependence and out-of-sample portfolio performance," *Working Paper, London Business School.*

Fan, J., Fan, Y., and Lv, J. (2008), "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147, 186–197.

Gibbons, M. R., Ross, S. A., and Shanken, J. (1989), "A test of the efficiency of a given portfolio," *Econometrica*, 57, 1121–1152.

Goetzmann, W., and Kumar, A. (2001), "Equity portfolio diversification," *Working Paper 8686, National Bureau of Economic Research.*

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999), "Bayesian model averaging," *Statistical Science*, 14, 382–401.

Huberman, G., and Kandel, S. (1987), "Mean-variance spanning," *Journal of Finance*, 42, 873–888.

Jagannathan, R., and Ma, T. (2003), "Risk reduction in large portfolios: why imposing the wrong constraints helps," *Journal of Finance*, 58, 1651–1683.

Jorion, P. (1986), "Bayes-Stein estimation for portfolio analysis," *The Journal of Financial and Quantitative Analysis*, 21, 279–292.

Kan, R., and Zhou, G. (2007), "Optimal portfolio choice with parameter uncertainty," *Journal of Financial and Quantitative analysis*, 42, 621–656.

Kan, R., and Zhou, G. (2001), "Tests for mean-variance spanning," *Working Paper, Washington University.*

Kempf, A., and Memmel, C. (2006), "Estimating the global minimum variance portfolio," *Schmalenbach Business Reviews*, 58, 332–348.

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

Markowitz, H. M. (1952), "Portfolio selection," *Journal of Finance*, 7, 77–91.

McQuarrie, A. D. R., and Tsai, C. L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific.

Polkovnichenko, V. (2003), "Household portfolio diversification," *Working Paper, University of Minnesota*.

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, 104, 177–186.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997), "An asymptotic theory for linear model selection (with discussion)," *Statistica Sinica*, 7, 221–264.

Sharpe, W. F., Alexander, G. J., and Bailey, J. V. (2001), *Investments*, New York: Prentice Hall.

Statman, M. (2004), "The diversification puzzle," *Financial Analysts Journal*, 60, 44–53.

Tibshirani, R. J. (1996), "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H. (2009), "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104, 1512–1524.

Yang, Y. (2001), "Adaptive regression by mixing," *Journal of the American Statistical Association*, 96, 574–588.

Zhang, Q., and Wang, H. (2011), "On BIC's selection consistency for discriminant analysis," *Statistica Sinica*, To Appear.

Table 1: Detailed simulation results with a sparse structure.

Portfolio Construction Method	$(d_0, d) = (5, 30)$ $\lvert\hat{\mathcal{S}}\rvert$	$\hat{\sigma}$ (%)	CF (%)	TR (%)	FR (%)	$(d_0, d) = (10, 60)$ $\lvert\hat{\mathcal{S}}\rvert$	$\hat{\sigma}$ (%)	CF (%)	TR (%)	FR (%)	$(d_0, d) = (15, 90)$ $\lvert\hat{\mathcal{S}}\rvert$	$\hat{\sigma}$ (%)	CF (%)	TR (%)	FR (%)
						$n = 120$									
NAIVE	30.00	37.48	0.00	100.00	100.00	60.00	49.04	0.00	100.00	100.00	90.00	38.52	0.00	100.00	100.00
MV	30.00	3.40	0.00	100.00	100.00	60.00	4.16	0.00	100.00	100.00	90.00	5.97	0.00	100.00	100.00
MU	30.00	55.42	0.00	100.00	100.00	60.00	215.79	0.00	100.00	100.00	90.00	137.20	0.00	100.00	100.00
SHORT	7.87	14.01	0.00	71.52	17.16	10.35	21.71	0.00	68.80	6.95	13.01	18.01	0.00	68.49	3.65
LASSO	14.00	3.22	1.40	88.48	38.31	19.57	3.75	0.20	82.44	22.66	23.37	4.24	0.60	80.25	15.11
BIC	5.48	3.10	23.40	87.44	4.44	11.56	3.33	3.40	86.94	5.73	19.31	3.78	1.20	86.15	8.51
						$n = 240$									
NAIVE	30.00	28.94	0.00	100.00	100.00	60.00	38.33	0.00	100.00	100.00	90.00	50.23	0.00	100.00	100.00
MV	30.00	3.14	0.00	100.00	100.00	60.00	3.40	0.00	100.00	100.00	90.00	3.71	0.00	100.00	100.00
MU	30.00	139.92	0.00	100.00	100.00	60.00	115.15	0.00	100.00	100.00	90.00	221.51	0.00	100.00	100.00
SHORT	7.64	11.86	0.00	73.32	15.89	9.48	17.19	0.00	69.52	5.06	12.08	23.35	0.00	69.44	2.22
LASSO	15.27	3.33	3.20	92.12	42.64	24.68	3.30	1.20	89.50	31.47	30.89	3.43	0.20	88.43	23.50
BIC	5.20	3.00	40.20	92.04	2.38	10.62	3.09	13.00	90.16	3.22	16.13	3.16	4.80	90.65	3.37
						$n = 600$									
NAIVE	30.00	36.24	0.00	100.00	100.00	60.00	56.00	0.00	100.00	100.00	90.00	50.31	0.00	100.00	100.00
MV	30.00	3.02	0.00	100.00	100.00	60.00	3.10	0.00	100.00	100.00	90.00	3.19	0.00	100.00	100.00
MU	30.00	83.39	0.00	100.00	100.00	60.00	93.27	0.00	100.00	100.00	90.00	125.00	0.00	100.00	100.00
SHORT	6.76	15.16	0.00	72.60	12.53	8.55	21.54	0.00	70.58	2.98	11.73	22.36	0.00	70.17	1.61
LASSO	15.68	2.99	4.00	95.44	43.64	28.18	3.07	1.60	94.40	37.48	38.15	3.15	1.40	93.00	32.27
BIC	5.08	2.96	53.00	93.56	1.59	10.23	2.99	28.00	93.76	1.71	15.20	3.01	15.00	93.41	1.58

Table 2: The results of analysis for the real data example.

Sample Size	Portfolio Construction Method	Standard Performance Measures				The Sharpe Ratio with γ in (%)		
		$\lvert\hat{\mathcal{S}}\rvert$	$\hat{\sigma}^*$ (%)	μ^* (%)	AT (%)	0.25	0.50	0.75
n=120	NAIVE	87.00	4.20	1.30	3.94	30.81	30.58	30.34
	MV	87.00	5.64	1.49	420.51	7.76	-10.89	-29.54
	MU	87.00	10.79	1.44	888.10	-7.25	-27.83	-48.40
	SHORT	7.42	3.72	1.34	11.84	35.21	34.41	33.62
	LASSO	14.71	3.37	1.44	113.71	34.37	25.95	17.52
	BIC	11.00	3.49	1.89	21.75	52.63	51.07	49.51
n=240	NAIVE	87.00	4.20	1.30	3.94	30.81	30.58	30.34
	MV	87.00	3.91	1.99	115.25	43.53	36.15	28.77
	MU	87.00	6.67	2.42	259.45	26.58	16.85	7.12
	SHORT	8.08	3.66	1.55	7.46	41.77	41.26	40.75
	LASSO	17.04	3.23	1.91	44.80	55.49	52.03	48.56
	BIC	8.00	3.43	2.21	12.84	63.55	62.61	61.68

Table 3: The pairwise comparisons between BIC and other five methods, where $\tilde{\mu}$, $\tilde{\sigma}$, and SR stands for, respectively, the differences in terms of mean return, standard deviation, and Sharpe Ratio with $\gamma = 0.50\%$.

Sample Size	Method	$\tilde{\mu}$(%)	p-value	$\tilde{\sigma}$(%)	p-value	SR(%)	p-value
n=120	BIC-Naive	0.59	0.10	-0.71	0.07	20.49	0.03
	BIC-MV	0.40	0.37	-2.15	0.00	61.96	0.00
	BIC-MU	0.45	0.64	-7.30	0.00	78.90	0.00
	BIC-SC	0.55	0.04	-0.23	0.48	16.66	0.03
	BIC-LASSO	0.36	0.07	0.18	0.47	19.42	0.00
n=240	BIC-Naive	0.91	0.01	-0.76	0.07	32.03	0.00
	BIC-MV	0.23	0.32	-0.47	0.09	26.46	0.00
	BIC-MU	-0.21	0.72	-3.23	0.00	45.76	0.00
	BIC-SC	0.67	0.01	-0.22	0.46	21.36	0.00
	BIC-LASSO	0.36	0.01	0.17	0.32	13.08	0.00