Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 692472, 13 pages
http://dx.doi.org/10.1155/2014/692472

Research Article

The Optimal Selection for Restricted Linear Models with Average Estimator

Qichang Xie[1] and Meng Du[2]

[1] *School of Economics, Shandong Institute of Business and Technology, Yantai, Shandong 264005, China*
[2] *School of Finance, Dongbei University of Finance and Economics, Dalian, Liaoning 116025, China*

Correspondence should be addressed to Qichang Xie; qichangx@163.com

Received 5 March 2014; Revised 15 April 2014; Accepted 29 April 2014; Published 21 May 2014

Academic Editor: Qian Guo

The essential task of risk investment is to select an optimal tracking portfolio among various portfolios. Statistically, this process can be achieved by choosing an optimal restricted linear model. This paper develops a statistical procedure to do this, based on selecting appropriate weights for averaging approximately restricted models. The method of weighted average least squares is adopted to estimate the approximately restricted models under dependent error setting. The optimal weights are selected by minimizing a k-class generalized information criterion (k-GIC), which is an estimate of the average squared error from the model average fit. This model selection procedure is shown to be asymptotically optimal in the sense of obtaining the lowest possible average squared error. Monte Carlo simulations illustrate that the suggested method has comparable efficiency to some alternative model selection techniques.

1. Introduction

The essential task of risk investment aims to select an optimal tracking portfolio among numerous portfolios of stocks. Given a desired target and a series of stocks, a tracking portfolio is comprised by every nonempty subset of the given group of stocks so as to track the target to a certain degree. Because of the number of nonempty subsets of stocks, there exists a mass of possible tracking portfolios. Among all possible portfolios, we should find an optimal tracking portfolio whose return is closest to the targets. Statistically, a tracking portfolio is built by a group of stocks, which is equivalent to fitting a restricted linear model with the target's return as the dependent variable and returns on stocks in the group as the regressors. Since the coefficient of a regressor indicates the proportion of the investment in the corresponding stock within the total investment in the portfolio, the linear model is restricted such that all coefficients in the model sum to one. Thus, the task of choosing an optimal tracking portfolio can be accomplished by selecting an optimal restricted linear model.

In this paper, a model average technique is developed for examining the selection problem of restricted linear models. Model selection has played an important role in econometrics and statistics over the past decades. The goal of model selection is to choose a model which gives the well-posed fit for observational data. So, the investigation of model selection is an indispensable process in empirical analysis. This work proposes a procedure of minimizing k-class generalized information criterion to select the optimal weights for constrained linear models. Under some conditions, we examine the asymptotic behaviors of the selection program.

Various methods have been suggested to study the problems of model selection. Knight and Fu [1] discussed the lasso-type estimators with least squares methods. To simultaneously estimate parameters and select important variables, Fan and Peng [2] proposed a method of nonconcave penalized likelihood and demonstrated that this technique had an oracle property when the number of parameters was infinite. Zou and Yuan [3] investigated the oracle theory of model selection based on composite quantile regression. Caner [4] considered model selection by the generalized method of moments estimator. In the empirical likelihood framework, Tang and Leng [5] studied the parametric estimator and variable selection for diverging numbers of parameters. Jennifer et al. [6] explored the ability of automatic selection algorithms to handle the selection problems of both variables and principal components.

Model averaging is another popular and widely used technique for model selection. The method is to average the estimators corresponding to different candidate models. Bayesian and frequentist are two main perspectives of thought in model averaging. Although their spirit and objectives are similar, the two techniques are different in inference and selection of models. In view of the Bayesian model averaging, the basic paradigm was introduced by Leamer [7]. Owing to the difficulty of implementing, the approach was basically ignored until the 2000s. About recent developments of this method, the readers can refer to Brown et al. [8] and Rodney and Herman [9]. Compared with Bayesian model averaging, since the method of frequentist model averaging focused on model selection rather than model averaging, it has been considered by many authors, for instance, Hjort and Claeskens [10], Hansen [11], Liang et al. [12], Zhang and Liang [13], and Hansen and Racine [14].

Generally speaking, different methods of model selection need to construct distinct model selection criteria including AIC [15], Mallows' C_L [16], CV [17], BIC [18], GCV [19], GMM J-statistic [20], and FPE_α [21]. Zhang et al. [22] employed the generalized information criterion for selecting the regularization parameters. To choose basis functions of splines, Xu and Huang [23] showed the optimal property of a LsoCV criterion and designed an efficient Newton-type algorithm for this criterion. Focusing on the divergence measure of Kullback-Leibler, So and Ando [24] defined a generalized predictive information criterion using the bias correction of an expected weighted loglikelihood estimator. Groen and Kapetanios [25] examined the criteria of AIC and BIC to discuss consistent estimates of a factor-augmented regression.

The literature mentioned above pays more attention to the unconstrained models with independently and identically distributed random errors. Recently, Lai and Xie [26] discussed model selection for constrained models, which were limited to the homoscedastic cases. Instead of using unrestricted models or homoscedastic models, we develop a k-class generalized information criterion (k-GIC) to discuss the selecting problems of approximately constrained linear models with dependent errors. The k-GIC is an extension of the GIC_{λ_n} proposed by Shao [27] and includes some conventional model selection criteria, such as BIC and GIC. We employ the technique of weighted average least squares to estimate the approximately constrained models and choose the weights through minimizing the k-class generalized information criterion. Our main result demonstrates that the k-class generalized information criterion is asymptotically equivalent to the average squared error. In other words, the selected weights from k-GIC are asymptotically optimal. Moreover, we highlight two new results which enrich the works of Lai and Xie [26]. One is that an estimate of variance is given and the estimate is proved to be consistent. Another is that the selected weights from k-GIC are shown to be still asymptotically optimal, when the true variance is replaced by the suggested estimate. The finite sample properties of model selection are performed by Monte Carlo simulation. The results of simulation reveal that the proposed method of model selection is dominant over some alternative approaches.

The remainder of this paper begins with an illustration of the model set-up and average estimator in Section 2. Section 3 calculates the average squared error of the model average estimator. The k-GIC criterion is introduced and its asymptotic optimality is derived in Section 4. Section 5 states some results from simulation evidence and Section 6 is conclusions.

2. Model Set-Up and Average Estimator

The core in risk investment is to build a tracking portfolio of stocks whose return mimics that of a chosen investment target. Let y_i be the return from investing in a selected target and $x_{i,j}$ be the historical return of the jth stock at time i. Assume that p stocks are available for building a tracking portfolio of the target. Then, a tracking portfolio consisting of all p stocks can be represented by

$$y_i = \sum_{j=1}^{p} x_{i,j}\theta_j + e_i, \tag{1}$$

where θ_j is unknown parameter and e_i is random error. The left-hand side of (1) is the return of investing one dollar in the target. The right-hand side is the return of investing one dollar in the portfolio consisting of all p stocks plus random noise.

Because each parameter θ_j stands for the proportion of investment on the corresponding stock to the total investment in the tracking portfolio, the sum of all parameters is one, namely,

$$\theta_1 + \cdots + \theta_p = 1, \tag{2}$$

which means the 100 percent of the whole investment.

In practice, there exist a large number of stocks. These stocks compose various portfolios that may track the target to some degree. Among all possible tracking portfolios, an ideal tracking portfolio should be the one whose return is closest to the target's return. Therefore, we need to find such an optimal tracking portfolio. Because of the dependance between a tracking portfolio and a restricted linear model, the aim of finding the optimal tracking portfolio can be accomplished by choosing an optimal restricted linear model.

In the following, we extend models (1) and (2) to consider a generalized constrained linear model for the problem of building an optimal tracking portfolio. Suppose that y_i, $i = 1, \ldots, n$ is a random observation at fixed value (x_i, z_i), where $x_i = (x_{i,1}, \ldots, x_{i,p})$ is fixed-dimensional explanatory variable and $z_i = (z_{i,1}, \ldots, z_{i,\infty})$ is countably infinite-dimensional explanatory variable. Consider the constrained linear model

$$y_i = \sum_{j=1}^{p} x_{i,j}\theta_j + \sum_{l=1}^{\infty} z_{i,l}\psi_l + e_i, \quad \text{subject to} \tag{3}$$

$$r_s\theta_s = \sum_{k=1}^{\infty} a_{s,k}\psi_k + b_s, \quad (s = 1, \ldots, p), \tag{4}$$

where p is a positive integer, θ_j and ψ_l are parameters, e_i is random error, r_s is a constant for restricting the sth parameter θ_s, and $a_{s,k}$ and b_s are some constants. Assume that $\phi(z_{i,l}, \psi_l) = \sum_{l=1}^{\infty} z_{i,l}\psi_l$ and $\phi(a_{s,k}, \psi_k) = \sum_{k=1}^{\infty} a_{s,k}\psi_k$ converge in mean square.

In (3), the explanatory variable $x_{i,j}$ is involved in the model on theoretical grounds or other reasons and $z_{i,l}$ is the additional explanatory variable that we need to make sure whether it should be included in the model. In the context of building tracking portfolio, the fixed explanatory variable $x_{i,j}$ stands for the historical return of the jth stock at time i, which must be selected by investors because of their personal preference or the stable earning of this stock. In a tracking portfolio, investors need to select some alternative stocks from numerous stocks to realize their expected return. So, the additional explanatory variable $z_{i,l}$ indicates the historical return of the lth alternative stock at time i. Since $z_{i,l}$ can be viewed as a series expansion, the identity (3) includes semiparametric models as special form. In fact, the model (3) generalizes the models considered by Lai and Xie [26] and Liang et al. [12]. In addition, the parameters θ_j and ψ_l denote, respectively, the proportion of the jth required stock and the lth alternative stock in a tracking portfolio. In (4), the parameter θ_s is adjusted by a linear combination of ψ_k and b_s. The economic significance of (4) is that the proportion of each fixed investment varies with the proportion of all alternative investments. When investors change their preference or have acquired new information on alternative stocks, they are capable of adjusting the proportion between required stocks and alternative stocks according to (4). This implies that the increase or decrease of alternative stocks can affect the proportion of each required stock in a portfolio. Besides, if we assume that $\psi_k = 0$, $k = 1, \ldots, \infty$, the model (4) becomes $r_s\theta_s = b_s$ which which has been discussed by Lai and Xie [26]. Particularly, if set $r_1 = \cdots = r_p = 1$, $a_{1,1} + \cdots + a_{p,1} = \cdots = a_{1,\infty} + \cdots + a_{p,\infty} = -1$, and $b_1 + \cdots + b_p = 1$, the restricted equation (4) turns into (2).

Denote an index set $\mathscr{U} = \{\mathscr{K}_1, \ldots, \mathscr{K}_M\}$, where M is a positive integer. Let $\Theta = (\theta_1, \ldots, \theta_p)^T$ and $\Psi = (\psi_1, \ldots, \psi_\infty)^T$, where "$T$" stands for the transpose operation. Due to the uncertain number of $z_{i,l}$ in formula (3), we consider a sequence of approximately restricted models (3) and (4) with $\mathscr{K}_m \in \mathscr{U}$, where the mth model includes the first $\mathscr{K}_m$ elements of z_i, that is, $z_{i,1}, \ldots, z_{i,\mathscr{K}_m}$, and the parameter of θ_s is restricted by a linear combination with $\mathscr{K}_m$ elements of Ψ and a constant b_s. Hence, the mth approximately restricted models (3) and (4) are

$$y_i = \sum_{j=1}^{p} x_{i,j}\theta_j + \sum_{l=1}^{\mathscr{K}_m} z_{i,l}\psi_l + d_i + e_i, \quad \text{subject to} \tag{5}$$

$$r_s\theta_s = \sum_{k=1}^{\mathscr{K}_m} a_{s,k}\psi_k + d_s^{\star} + b_s, \quad (s = 1, \ldots, p), \tag{6}$$

where $d_i = \sum_{l=\mathscr{K}_m+1}^{\infty} z_{i,l}\psi_l$ and $d_s^{\star} = \sum_{k=\mathscr{K}_m+1}^{\infty} a_{s,k}\psi_k$ are the approximation errors.

Set $Y = (y_1, \ldots, y_n)^T$, $a_s^m = (a_{s,1}, \ldots, a_{s,\mathscr{K}_m})^T$, $A_{\aleph}^m = (a_1^{m^T}, \ldots, a_p^{m^T})^T$, $D = (d_1, \ldots, d_n)^T$, $D^{\star} = (d_1^{\star}, \ldots, d_p^{\star})^T$, $\Psi_m = (\psi_1, \ldots, \psi_{\mathscr{K}_m})^T$, $B = (b_1, \ldots, b_p)^T$, and $e = (e_1, \ldots, e_n)^T$. By matrix notation, the mth approximately restricted models (5) and (6) can be rewritten as

$$Y = X\Theta + Z_m\Psi_m + D + e, \quad \text{subject to} \tag{7}$$

$$R\Theta = A_{\aleph}^m\Psi_m + D^{\star} + B, \tag{8}$$

where X is a $n \times p$ matrix whose ijth element is $x_{i,j}$, Z_m is a $n \times \mathscr{K}_m$ matrix whose ijth element is $z_{i,j}$, and R is a $p \times p$ diagonal matrix whose ith diagonal element is r_i.

Hypothesize that $\{e_i\}_{i=1}^n$ satisfies $E(e_i \mid x_i, z_i) = 0$ and its conditional covariance matrix $\Omega_n = E(ee^T \mid X, Z_m)$ is

$$\Omega_n = \begin{pmatrix} \sigma_1^2 & \varrho_1 & \cdots & \varrho_{n-1} \\ \varrho_1 & \sigma_2^2 & \cdots & \varrho_{n-2} \\ \vdots & \vdots & \cdots & \vdots \\ \varrho_{n-1} & \varrho_{n-2} & \cdots & \sigma_n^2 \end{pmatrix}, \tag{9}$$

in which $E(e_ie_i \mid x_i, z_i) = \sigma_i^2$ and $E(e_ie_{i+j} \mid x_i, z_i) = \varrho_j$ with $i = 1, \ldots, n$ and $j = 1, \ldots, n-1$. Clearly, the random errors follow a heteroscedastic stationary Gaussian process.

Substituting (8) into (7), it yields

$$Y_u = Z_{m,u}\Psi_m + U, \tag{10}$$

where $Y_u = Y - XR^{-1}B$, $U = D + XR^{-1}D^{\star} + e$, and $Z_{m,u} = Z_m + XR^{-1}A_{\aleph}^m$. By the method of least squares, the estimator of Ψ_m is

$$\widehat{\Psi}_m = \left(Z_{m,u}^T Z_{m,u}\right)^{-1} Z_{m,u}^T Y_u, \tag{11}$$

where $(Z_{m,u}^T Z_{m,u})^{-1}$ denotes the inverse of $Z_{m,u}^T Z_{m,u}$. In the mth approximating model (7), we set $\mu_m = X\Theta + Z_m\Psi_m$, so that $\mu = E(Y \mid X, Z_m) = \mu_m + D$. Thus, the estimator of μ_m is

$$\widehat{\mu}_m = X\widehat{\Theta} + Z_m\widehat{\Psi}_m = \eta + P_mY, \tag{12}$$

where $\eta = (I - P_m)XR^{-1}B$ and $P_m = Z_{m,u}(Z_{m,u}^T Z_{m,u})^{-1} Z_{m,u}^T$ is the "hat" matrix.

Let $w = (w_1, \ldots, w_M)^T$ be a weight vector, where $M < n$. Define a weight set $\mathscr{W}$ as

$$\mathscr{W} = \left\{ w \mid w_m \in [0, 1], m = 1, \ldots, M, \sum_{m=1}^{M} w_m = 1 \right\}. \tag{13}$$

For all $m \le M$, the weighted average estimator of Ψ_m is

$$\widehat{\Psi}^w = \sum_{m=1}^{M} w_m\widehat{\Psi}_m = \sum_{m=1}^{M} w_m\left(Z_{m,u}^T Z_{m,u}\right)^{-1} Z_{m,u}^T Y_u. \tag{14}$$

Naturally, the weighted average estimator of $\widehat{\Theta}$ is

$$\widehat{\Theta}^w = R^{-1}B + R^{-1}A_{\aleph}^m\widehat{\Psi}^w. \tag{15}$$

Furthermore, the weighted average estimator of μ is

$$\widehat{\mu}(w) = \sum_{m=1}^{M} w_m\widehat{\mu}_m = \eta(w) + P(w)Y, \tag{16}$$

where $P(w) = \sum_{m=1}^{M} w_m P_m$ and $\eta(w) = (I - P(w))XR^{-1}B$. It can be seen that the weighted estimator $\hat{\mu}(w)$ is an average estimator of $\hat{\mu}_m$, $m = 1, \ldots, M$. The weighted "hat" matrix $P(w)$ depends on nonrandom regressor $Z_{m,u}$ and weight vector w. In general conditions, the matrix $P(w)$ is symmetric, but not idempotent.

For a positive integer $\mathscr{G}$, let ξ be the maximal value of σ_ι^2, $\iota = 1, \ldots, n$, and let $\overline{\lambda}_j$, $j = 1, \ldots, \mathscr{G}$ be the nonzero eigenvalue of Ω_n. Assume that both $\overline{\lambda}_j$ and ϱ_ι are summable, namely,

$$\Gamma = \sum_{j=1}^{\mathscr{G}} \overline{\lambda}_j < \infty,$$
$$\mathfrak{W} = \xi + 2\sum_{\iota=1}^{n-1} \left|\varrho_\iota\right| < \infty. \tag{17}$$

Since the covariance matrix Ω_n determines the algebraic structure of the model average estimator, we discuss its properties in the following.

Lemma 1. *For any n dimensional vectors $a = (a_1, \ldots, a_n)^T$ and $b = (b_1, \ldots, b_n)^T$, one has $|a^T\Omega_n b| \le \|a\|\|b\|\mathfrak{W}$, where $\|\cdot\|$ is the Euclidean norm.*

Proof. Through the definition of Ω_n, one has

$$\begin{aligned} \left|a^T\Omega_n b\right| &= \left|\sum_{l=1}^{n} a_l b_l \sigma_l^2 + \sum_{i=1}^{n-1} \varrho_i \sum_{l=1}^{n-i} \left(a_{l+i}b_l + a_l b_{l+i}\right)\right| \\ &\le \xi\left|\sum_{l=1}^{n} a_l b_l\right| + \left|\sum_{i=1}^{n-1} \varrho_i\right|\left|\sum_{l=1}^{n-i} \left(a_{l+i}b_l + a_l b_{l+i}\right)\right|. \end{aligned} \tag{18}$$

Applying Cauchy-Schwarz inequality, for $i \in \{1, \ldots, n-1\}$, one gets

$$\begin{aligned} \sum_{l=1}^{n-i} a_{l+i} b_l &\le \sqrt{\sum_{l=1}^{n-i} a_{l+i}^2 \sum_{l=1}^{n-i} b_l^2} \\ &\le \sqrt{\sum_{l=1}^{n} a_l^2 \sum_{l=1}^{n} b_l^2} = \|a\|\,\|b\|. \end{aligned} \tag{19}$$

Similarly, $\sum_{l=1}^{n-i} a_l b_{l+i} \le \|a\|\|b\|$ holds. Therefore,

$$\left|a^T\Omega_n b\right| \le \|a\|\,\|b\|\left(\xi + 2\sum_{i=1}^{n-1} \varrho_i\right) = \|a\|\,\|b\|\,\mathfrak{W}. \tag{20}$$

□

Because the matrix $P(w)$ takes an important role in analyzing the problems of model selection, we state some of its properties. We set $\tilde{\tau} = \max\{\mathscr{K}_1, \ldots, \mathscr{K}_M\}$. Let $\lambda(P(w))$ and $\lambda_{\max}(P(w))$ denote the eigenvalue and the largest eigenvalue of $P(w)$, respectively.

Lemma 2. *One has (i) $\operatorname{tr}(P(w)) \le \tilde{\tau}$ and (ii) $0 \le \lambda(P(w)) \le 1$, where $\operatorname{tr}(\cdot)$ denotes the trace operation.*

Proof. It follows from $\operatorname{tr}(P_m) = \mathscr{K}_m$ and $\operatorname{tr}(P(w)) = \sum_{m=1}^{M} w_m \mathscr{K}_m \le \tilde{\tau}$ that (i) is established. Next, we consider (ii). Without loss of generality, let $\lambda_{1,m} \ge \cdots \ge \lambda_{n,m}$ and $\varphi_{1,m}, \ldots, \varphi_{n,m}$ be the eigenvalues and standardly orthogonal eigenvectors of P_m, respectively. Observe that P_m is idempotent, which implies that $\lambda_{i,m} \in \{0, 1\}$, $i = 1, \ldots, n$. For any $\varpi \in \mathfrak{R}^n$, it can be seen that

$$\begin{aligned} \lambda\left(P(w)\right) &= \frac{\varpi^T P(w)\varpi}{\varpi^T\varpi} \\ &= \sum_{m=1}^{M} w_m \frac{\varpi^T P_m \varpi}{\varpi^T\varpi} \\ &= \sum_{m=1}^{M} w_m \frac{\chi^T \Xi_m \chi}{\chi^T\chi} \\ &= \sum_{m=1}^{M} w_m \sum_{i=1}^{n} \lambda_{i,m} \frac{\left|\chi_i\right|^2}{\chi^T\chi}, \end{aligned} \tag{21}$$

where $\Xi_m = \operatorname{diag}(\lambda_{1,m}, \ldots, \lambda_{n,m})$, $\varpi = \Phi_m\chi$, and $\Phi_m = (\varphi_{1,m}, \ldots, \varphi_{n,m})$. Due to $\lambda_{i,m} \in \{0, 1\}$, $i = 1, \ldots, n$, it means that

$$\begin{aligned} \lambda\left(P(w)\right) &\le \sum_{m=1}^{M} w_m \sum_{i=1}^{n} 1 \frac{\left|\chi_i\right|^2}{\chi^T\chi} = \sum_{m=1}^{M} w_m = 1, \\ \lambda\left(P(w)\right) &\ge \sum_{m=1}^{M} w_m \sum_{i=1}^{n} 0 \frac{\left|\chi_i\right|^2}{\chi^T\chi} = 0. \end{aligned} \tag{22}$$

□

From Lemma 2(ii), we know that $P(w)$ is nonnegative definite.

Lemma 3. *Let α denote the number of eigenvalues of $P(w)$. Then $(\alpha^{-1}\operatorname{tr}(P(w)))^2 \le \alpha^{-1}\operatorname{tr}(P^2(w))$.*

Proof. Let $\lambda_1, \ldots, \lambda_\alpha$ be the eigenvalues of $P(w)$. (i) If $\lambda_1 = \cdots = \lambda_\alpha = 0$, we know that Lemma 3 holds. (ii) Let $\lambda_1 \ne 0, \ldots, \lambda_{\overline{\alpha}} \ne 0$ and $\lambda_{\overline{\alpha}+1} = \cdots = \lambda_\alpha = 0$. It is easy to see that $\sum_{i=1}^{\overline{\alpha}} (\lambda_i - \overline{\lambda})^2 \ge 0$, where $\overline{\lambda} = \overline{\alpha}^{-1}\operatorname{tr}(P(w))$. Thus, it can be shown that

$$\begin{aligned} \sum_{i=1}^{\overline{\alpha}} \left(\lambda_i - \overline{\lambda}\right)^2 &= \sum_{i=1}^{\overline{\alpha}} \lambda_i^2 - 2\overline{\lambda}\sum_{i=1}^{\overline{\alpha}} \lambda_i + \overline{\alpha}\overline{\lambda}^2 \\ &= \operatorname{tr}\left(P^2(w)\right) - \overline{\alpha}\left(\overline{\alpha}^{-1}\operatorname{tr}\left(P(w)\right)\right)^2 \ge 0, \end{aligned} \tag{23}$$

which implies that $(\overline{\alpha}^{-1}\operatorname{tr}(P(w)))^2 \le \overline{\alpha}^{-1}\operatorname{tr}(P^2(w))$. □

Lemma 4. *For any $w \in \mathscr{W}$, there exists $\mathscr{C}_1 = \mathscr{C}'^2$ such that $\operatorname{tr}(P(w)\Omega_n) \le \mathscr{C}'\Gamma$, $\operatorname{tr}(\Omega_n P(w))^2 \le \mathscr{C}_1\Gamma^2$ and $\operatorname{tr}(\Omega_n P^T(w)P(w))^2 \le \mathscr{C}_1\Gamma\operatorname{tr}(P(w)\Omega_n P^T(w))$, in which $\mathscr{C}' = sup_{w\in\mathscr{W}}\lambda_{\max}(P(w))$.*

Proof. Notice that $\lambda_{\min}(\mathscr{A})\lambda_{\max}(\mathscr{B}) \le \lambda_{\max}(\mathscr{A}\mathscr{B}) \le \lambda_{\max}(\mathscr{A})\lambda_{\max}(\mathscr{B})$ and $\operatorname{tr}(\mathscr{A}\mathscr{B}) \le \lambda_{\max}(\mathscr{A})\operatorname{tr}(\mathscr{B}) \le$

$\operatorname{tr}(\mathscr{A})\operatorname{tr}(\mathscr{B})$, where $\mathscr{A}$ and $\mathscr{B}$ are any positive semidefinite matrices.

Since $P(w)$ and Ω_n are symmetric and nonnegative definite, the first inequality is established by $\operatorname{tr}(P(w)\Omega_n) \le \lambda_{\max}(P(w))\operatorname{tr}(\Omega_n)$. Let λ_i be eigenvalue of $P(w)$. From the symmetric property of $P(w)$, we know that there exists a $n \times n$ orthogonal matrix $\overline{\Psi}$ such that $P(w) = \overline{\Psi}\Upsilon\overline{\Psi}^T$, where $\Upsilon = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Then, it yields $\operatorname{tr}(\Omega_n P(w)) = \operatorname{tr}(\Upsilon^{1/2}\overline{\Psi}^T\Omega_n\overline{\Psi}\Upsilon^{1/2})$, where $\Upsilon^{1/2}\overline{\Psi}^T\Omega_n\overline{\Psi}\Upsilon^{1/2}$ is symmetric and nonnegative definite. The second inequality holds because

$$\begin{aligned} \operatorname{tr}\left(\Omega_n P(w)\right)^2 &\le \left[\operatorname{tr}\left(\Omega_n P(w)\right)\right]^2 \\ &\le \left[\lambda_{\max}\left(P(w)\right)\operatorname{tr}(\Omega_n)\right]^2 \\ &\le \mathscr{C}_1\Gamma^2. \end{aligned} \tag{24}$$

For the last formula, we note that

$$\begin{aligned} &\operatorname{tr}\left(\Omega_n P^T(w) P(w)\right)^2 \\ &\quad = \operatorname{tr}\left(P(w)\Omega_n P^T(w)\right)^2 \\ &\quad \le \lambda_{\max}\left(P(w)\Omega_n P^T(w)\right)\operatorname{tr}\left(P(w)\Omega_n P^T(w)\right) \\ &\quad \le \lambda_{\max}\left(P^2(w)\right)\lambda_{\max}\left(\Omega_n\right)\operatorname{tr}\left(P(w)\Omega_n P^T(w)\right) \\ &\quad \le \lambda_{\max}^2\left(P(w)\right)\lambda_{\max}\left(\Omega_n\right)\operatorname{tr}\left(P(w)\Omega_n P^T(w)\right) \\ &\quad \le \mathscr{C}_1\Gamma\operatorname{tr}\left(P(w)\Omega_n P^T(w)\right). \end{aligned} \tag{25}$$

This completes the proof of Lemma 4. □

Lemma 5. *Let $\mathscr{A}$ be a symmetric matrix and X be a random vector with zero expectation. Then, one has* $\operatorname{Var}(X^T\mathscr{A}X) = 2\operatorname{tr}(\operatorname{Var}(X)\mathscr{A}\operatorname{Var}(X)\mathscr{A})$*, where* $\operatorname{Var}(\cdot)$ *is the operation of variance.*

Proof. The proof of this lemma is provided in Lai and Xie [26]. □

3. Average Squared Error

Denote an average squared error by

$$L_n(w) = \frac{1}{n}\left\|\mu - \widehat{\mu}(w)\right\|_k^2 = \frac{\left(\mu - \widehat{\mu}(w)\right)^T k\left(\mu - \widehat{\mu}(w)\right)}{n}, \tag{26}$$

where k is a fixed positive integer which is often used to eliminate the boundary effect. Andrews [28] suggested that the k can take the value of σ^{-2}, when errors obeyed an independent and identical distribution with variance σ^2. The most common situation is $k = 1$. The average squared error $L_n(w)$ can be viewed as a measure of accuracy between $\widehat{\mu}(w)$ and μ. Obviously, an optimal estimator can obtain the minimum value of $L_n(w)$. In other words, we should select a weight vector w from $\mathscr{W}$ to make that the average squared error $L_n(w)$ takes value that is as small as possible.

In order to investigate the problem of weight selection, we give the expression of conditional expected average squared error as follows:

$$R_n(w) = E\left(L_n(w) \mid X, Z_m\right). \tag{27}$$

From the definition of $R_n(w)$, the following lemma can be obtained.

Lemma 6. *The conditional expected average squared error can be rewritten as*

$$R_n(w) = \frac{\|M(w)\mu - \eta(w)\|_k^2}{n} + \frac{k\operatorname{tr}\left(P(w)\Omega_n P^T(w)\right)}{n}, \tag{28}$$

where $M(w) = I - P(w)$.

Proof. A straight calculation of $L_n(w)$ leads to

$$\begin{aligned} &nL_n(w) \\ &\quad = \left(\mu - \widehat{\mu}(w)\right)^T k\left(\mu - \widehat{\mu}(w)\right) \\ &\quad = \left(\mu - \eta(w) - P(w)(\mu + e)\right)^T \\ &\qquad \times k\left(\mu - \eta(w) - P(w)(\mu + e)\right) \\ &\quad = \left[\mu^T M^T(w) - \eta^T(w) - e^T P^T(w)\right] \\ &\qquad \times k\left[M(w)\mu - \eta(w) - P(w)e\right] \\ &\quad = \|M(w)\mu - \eta(w)\|_k^2 - k\left(\mu - XR^{-1}B\right)^T M^T(w)P(w)e \\ &\qquad + ke^T P^T(w)P(w)e - ke^T P^T(w)M(w)\left(\mu - XR^{-1}B\right). \end{aligned} \tag{29}$$

Using $E(e^T P^T(w)P(w)e \mid X, Z_m) = \operatorname{tr}(P(w)\Omega_n P^T(w))$ and taking conditional expectations on both sides of the above equality give rise to Lemma 6. □

4. The k-Class Generalized Information Criterion and Asymptotic Optimality

As the value of μ is unknown, the average squared error $L_n(w)$ cannot be used directly to select the weight vector w. Thus, we suggest a k-class generalized information criterion (k-GIC) to choose the weight vector. Further, we will prove that the selected weight vector from k-GIC minimizes the average squared error $L_n(w)$.

The k-GIC for the restricted model average estimator is

$$C_n(w) = \frac{\|Y - \widehat{\mu}(w)\|_k^2}{n} + \frac{2\hbar}{n}\operatorname{tr}\left(P(w)\Omega_n\right), \tag{30}$$

where $\hbar$ is larger than one and satisfies assumption (35) mentioned below. The k-class generalized information criterion extends the generalized information criterion (GIC_{λ_n}) proposed by Shao [27]. Because the $\hbar$ can take different values, the k-GIC includes some common information criteria for model selection such as the Mallows C_L criterion ($k = \hbar = 1$), the GIC criterion ($k = 1, \hbar \to \infty$), the FPE_α criterion ($k = 1, \hbar > 1$), and the BIC criterion ($k = 1, \hbar = \log(n)$).

Lemma 7. *If $k = \hbar$, we have $E(C_n(w) \mid X, Z_m) = R_n(w) + k\Gamma/n$.*

Proof. Recalling the definition of $C_n(w)$, one gets

$$\begin{aligned} nC_n(w) &= (Y - \widehat{\mu}(w))^T k (Y - \widehat{\mu}(w)) + 2k\,\mathrm{tr}(P(w)\Omega_n) \\ &= [M(w)Y - \eta(w)]^T k [M(w)Y - \eta(w)] \\ &\quad + 2k\,\mathrm{tr}(P(w)\Omega_n) \\ &= 2k\,\mathrm{tr}(P(w)\Omega_n) + ke^T M^T(w) M(w) e \\ &\quad + k(M(w)\mu - \eta(w))^T M(w) e \\ &\quad + ke^T M^T(w)(M(w)\mu - \eta(w)) \\ &\quad + \|M(w)\mu - \eta(w)\|_k^2. \end{aligned} \tag{31}$$

Observe that

$$\begin{aligned} ke^T M^T(w) M(w) e &= e^T (I - P(w))^T k (I - P(w)) e \\ &= ke^T e - ke^T P^T(w) e - ke^T P(w) e \\ &\quad + ke^T P^T(w) P(w) e. \end{aligned} \tag{32}$$

Notice that $E(ke^T e \mid X, Z_m) = k\,\mathrm{tr}(\Omega_n)$ and $E(ke^T P^T(w) e \mid X, Z_m) = k\,\mathrm{tr}(P^T(w)\Omega_n)$. Moreover, we have that $E(ke^T P(w) e \mid X, Z_m) = k\,\mathrm{tr}(P(w)\Omega_n)$ and $E(ke^T P^T(w) P(w) e \mid X, Z_m) = k\,\mathrm{tr}(P(w)\Omega_n P^T(w))$.

Thus, taking conditional expectations on both sides of (31), we obtain

$$\begin{aligned} nE(C_n(w) \mid X, Z_m) &= k\,\mathrm{tr}(\Omega_n) + \|M(w)\mu - \eta(w)\|_k^2 \\ &\quad + k\,\mathrm{tr}(P(w)\Omega_n P^T(w)). \end{aligned} \tag{33}$$

It follows from (33) and Lemma 6 that Lemma 7 is established as desired. □

Lemma 7 shows that $C_n(w)$ is equivalent to the conditional expected average squared error plus an error bias. Particularly, when n approaches to infinite, $C_n(w)$ is an unbiased estimation of $R_n(w)$.

The k-GIC criterion is defined so as to select the optimal weight vector $\widehat{w}$. The optimal weight vector $\widehat{w}$ is chosen by minimizing $C_n(w)$.

Obviously, the well-posed estimators of parameters are $\widehat{\Theta}^w$ and $\widehat{\Psi}^w$ with the weight vector $\widehat{w}$. Under some regular conditions, we intend to demonstrate that the selection procedure is asymptotically optimal in the following sense:

$$\frac{L_n(\widehat{w})}{\inf_{w\in\mathscr{W}} L_n(w)} \xrightarrow{p} 1. \tag{34}$$

If the formula (34) holds, we know that the selected weight vector $\widehat{w}$ from $C_n(w)$ can realize the minimum value of $L_n(w)$. In other words, the weight vector $\widehat{w}$ is equivalent to the selected weight vector by minimizing $L_n(w)$ and is the optimal weight vector for $\widehat{\mu}(w)$. The asymptotic optimality of $C_n(w)$ can be established under the following assumptions.

Assumptions 1. Write $\overline{\alpha}_n = \inf_{w\in\mathscr{W}} nR_n(w)$. As $n \to \infty$, we assume that

$$\underline{\lim}\inf_{w\in\mathscr{W}} \frac{\hbar^2}{\overline{\alpha}_n} \longrightarrow 0, \tag{35}$$

$$\sum_{w\in\mathscr{W}} \frac{1}{(nR_n(w))^{N'}} \longrightarrow 0 \tag{36}$$

for a large $\hbar$ and any positive integer $1 \le N' < \infty$.

The above assumptions have been employed by many literatures of model selection. For instance, the expression (35) was used by Shao [27] and the formula (36) was adopted by Li [29], Andrews [28], Shao [27], and Hansen [11].

The following lemma offers a bridge for proving the asymptotic optimality of $C_n(w)$.

Lemma 8. *We have*

$$\begin{aligned} \|M(w)Y - \eta(w)\|_k^2 &= nL_n(w) - 2\langle e^T, P(w)e\rangle_k \\ &\quad + 2\langle e^T, M(w)\mu - \eta(w)\rangle_k + \|e\|_k^2. \end{aligned} \tag{37}$$

Proof. Using $\widehat{\mu}(w) = P(w)Y + \eta(w)$, $Y = \mu + e$, and the definition of $L_n(w)$, one obtains

$$\begin{aligned} \|M(w)Y - \eta(w)\|_k^2 &= \|\mu + e - P(w)Y - \eta(w)\|_k^2 \\ &= \|e\|_k^2 + \|\mu - P(w)Y - \eta(w)\|_k^2 \\ &\quad + 2\langle e^T, \mu - \eta(w) - P(w)Y\rangle_k \\ &= \|e\|_k^2 + nL_n(w) \\ &\quad + 2\langle e^T, \mu - \eta(w) - P(w)(\mu + e)\rangle_k \\ &= \|e\|_k^2 + nL_n(w) \\ &\quad + 2\langle e^T, (I - P(w))\mu - \eta(w) - P(w)e\rangle_k \\ &= \|e\|_k^2 + nL_n(w) \\ &\quad + 2\langle e^T, M(w)\mu - \eta(w)\rangle_k \\ &\quad - 2\langle e^T, P(w)e\rangle_k. \end{aligned} \tag{38}$$

This completes the proof of Lemma 8. □

Lemma 8 and $\|Y - \widehat{\mu}(w)\|_k^2 = \|M(w)Y - \eta(w)\|_k^2$ imply that

$$\begin{aligned} C_n(w) &= L_n(w) + \frac{2\langle e^T, M(w)\mu - \eta(w)\rangle_k}{n} \\ &\quad + \frac{\|e\|_k^2}{n} + \frac{2\hbar\,\mathrm{tr}(P(w)\Omega_n)}{n} - 2\frac{\langle e^T, P(w)e\rangle_k}{n}. \end{aligned} \tag{39}$$

The goal is to choose $\widehat{w}$ by minimizing $C_n(w)$. From (39), one only needs to select $\widehat{w}$ through minimizing

$$L_n(w) + \frac{2\left\langle e^T, M(w)\mu - \eta(w)\right\rangle_k}{n} + \frac{2\hbar\,\mathrm{tr}\left(P(w)\Omega_n\right) - 2\left\langle e^T, P(w)e\right\rangle_k}{n}, \tag{40}$$

where $w \in \mathscr{W}$.

Compared with $L_n(w)$, it is sufficient to establish that $n^{-1}\langle e^T, M(w)\mu - \eta(w)\rangle_k$ and $n^{-1}\hbar\,\mathrm{tr}(P(w)\Omega_n) - n^{-1}\langle e^T, P(w)e\rangle_k$ are uniformly negligible for any $w \in \mathscr{W}$. More specifically, to prove (34), we need to check

$$\sup_{w\in\mathscr{W}} \frac{\left\langle e^T, M(w)\mu - \eta(w)\right\rangle_k}{nR_n(w)} \xrightarrow{p} 0, \tag{41}$$

$$\sup_{w\in\mathscr{W}} \frac{\left|\hbar\,\mathrm{tr}\left(P(w)\Omega_n\right) - \left\langle e^T, P(w)e\right\rangle_k\right|}{nR_n(w)} \xrightarrow{p} 0, \tag{42}$$

$$\sup_{w\in\mathscr{W}} \left|\frac{L_n(w) - R_n(w)}{R_n(w)}\right| \xrightarrow{p} 0, \tag{43}$$

where "$\xrightarrow{p}$" denotes the convergence in probability.

Following the idea of Li [29], we testify the main result of our work that the minimizing of k-class generalized information criterion is asymptotically optimal. Now, we state the main theorem.

Theorem 9. *Under assumptions* (35) *and* (36), *the minimizing of k-class generalized information criterion $C_n(w)$ is asymptotically optimal, namely,* (34) *holds.*

Proof. The asymptotically optimal property of k-GIC needs to show that (41), (42), and (43) are valid.

Firstly, we prove that (41) holds. For any $\zeta > 0$, by Chebyshev's inequality, one has

$$\begin{aligned} \Pr&\left\{\sup_{w\in\mathscr{W}} \frac{\left|\left\langle e^T, M(w)\mu - \eta(w)\right\rangle_k\right|}{nR_n(w)} > \zeta\right\} \\ &\leq \sum_{w\in\mathscr{W}} \frac{E\left[\left\langle e^T, M(w)\mu - \eta(w)\right\rangle_k\right]^2}{\left(\zeta nR_n(w)\right)^2}, \end{aligned} \tag{44}$$

which, by Lemma 1, is no greater than

$$\zeta^{-2}k\mathfrak{W}\sum_{w\in\mathscr{W}} n^{-2}\|M(w)\mu - \eta(w)\|_k^2\left(R_n(w)\right)^{-2}. \tag{45}$$

Recalling the definition of $R_n(w)$, we get $\|M(w)\mu - \eta(w)\|_k^2 \leq nR_n(w)$. Then, (45) does not exceed $\zeta^{-2}k\mathfrak{W}\sum_{w\in\mathscr{W}}(nR_n(w))^{-1}$. By assumption (36), one knows that (45) tends to zero in probability. Thus, (41) is established.

To prove (42), it suffices to testify that

$$\frac{k\,\mathrm{tr}\left(P(w)\Omega_n\right) - \left\langle e^T, P(w)e\right\rangle_k}{nR_n(w)} \xrightarrow{p} 0, \tag{46}$$

$$\frac{(\hbar - k)\,\mathrm{tr}\left(P(w)\Omega_n\right)}{nR_n(w)} \xrightarrow{p} 0. \tag{47}$$

By Chebyshev's inequality and Lemmas 4 and 5, for any $\epsilon > 0$, we have

$$\begin{aligned} \Pr&\left\{\sup_{w\in\mathscr{W}}\left|\frac{\mathrm{tr}\left(P(w)\Omega_n\right) - \left\langle e^T, P(w)e\right\rangle}{nR_n(w)}\right| > \epsilon\right\} \\ &\leq \sum_{w\in\mathscr{W}} \frac{E\left[\mathrm{tr}(P(w)\Omega_n) - \left\langle e^T, P(w)e\right\rangle\right]^2}{\left(\epsilon nR_n(w)\right)^2} \\ &= \frac{1}{\epsilon^2}\sum_{w\in\mathscr{W}} \frac{\mathrm{Var}\left(e^TP(w)e\right)}{\left(nR_n(w)\right)^2} \\ &\leq \frac{2\mathscr{C}_1\Gamma^2}{\epsilon^2}\sum_{w\in\mathscr{W}} \frac{1}{\left(nR_n(w)\right)^2}. \end{aligned} \tag{48}$$

From (36), we know that (48) is close to zero. Thus, (46) is reasonable.

Recalling Lemma 4 and (35), it yields

$$\frac{\left|(\hbar - k)\,\mathrm{tr}\left(P(w)\Omega_n\right)\right|}{nR_n(w)} \leq \frac{\left|\mathscr{C}'(\hbar - k)\Gamma\right|}{nR_n(w)} \longrightarrow 0, \tag{49}$$

which derives (47). Then, (42) is proved.

Next, we show that the expression (43) also holds. A straightforward calculation leads to

$$\begin{aligned} \|\mu &- \widehat{\mu}(w)\|_k^2 - \|M(w)\mu - \eta(w)\|_k^2 \\ &= \|\mu - \eta(w) - P(w)Y\|_k^2 - \|M(w)\mu - \eta(w)\|_k^2 \\ &= \|\mu - \eta(w) - P(w)\mu - P(w)e\|_k^2 - \|M(w)\mu - \eta(w)\|_k^2 \\ &= \|M(w)\mu - \eta(w) - P(w)e\|_k^2 - \|M(w)\mu - \eta(w)\|_k^2 \\ &= \|P(w)e\|_k^2 - 2\left\langle \mu^TM^T(w) - \eta^T(w), P(w)e\right\rangle_k. \end{aligned} \tag{50}$$

From the identity (50), we know that

$$\begin{aligned} L_n(w) - R_n(w) = &-\frac{2\left\langle \mu^TM^T(w) - \eta^T(w), P(w)e\right\rangle_k}{n} \\ &+ \frac{\|P(w)e\|_k^2 - k\,\mathrm{tr}\left(P(w)\Omega_nP^T(w)\right)}{n}. \end{aligned} \tag{51}$$

Obviously, the proof of (43) needs to verify that

$$\begin{gathered}\sup_{w\in\mathscr{W}} \frac{\|P(w)e\|_k^2 - k\operatorname{tr}\left(P(w)\,\Omega_n P^T(w)\right)}{nR_n(w)} \xrightarrow{p} 0,\\ \sup_{w\in\mathscr{W}} \frac{\left\langle \mu^T M^T(w) - \eta^T(w), P(w)e\right\rangle_k}{nR_n(w)} \xrightarrow{p} 0.\end{gathered} \tag{52}$$

To prove (52), it should be noticed that $\operatorname{tr}(P(w)\Omega_n P^T(w)) = E(e^T P^T(w)P(w)e)$ and $\|P(w)e\|_k^2 = \langle e^T P^T(w), P(w)e\rangle_k$. In addition,

$$\begin{aligned}\left\|P(w)\left(M(w)\mu - \eta(w)\right)\right\|^2 &\le \lambda^2_{\max}(P(w))\left\|M(w)\mu - \eta(w)\right\|^2\\ &\le \left\|M(w)\mu - \eta(w)\right\|^2.\end{aligned} \tag{53}$$

Thus, there exist any $\zeta' > 0$ and $\epsilon' > 0$ such that

$$\begin{aligned}\Pr&\left\{\sup_{w\in\mathscr{W}} \frac{\|P(w)e\|_k^2 - k\operatorname{tr}\left(P(w)\,\Omega_n P^T(w)\right)}{nR_n(w)} > \zeta'\right\}\\ &\le \sum_{w\in\mathscr{W}} \frac{E\left[\|P(w)e\|_k^2 - kE\left(e^T P^T(w)\,P(w)\,e\right)\right]^2}{\left(nR_n(w)\,\zeta'\right)^2}\\ &= \frac{k^2}{\zeta'^2}\sum_{w\in\mathscr{W}} \frac{\operatorname{Var}\left(e^T P^T(w)\,P(w)\,e\right)}{\left(nR_n(w)\right)^2}\\ &\le \frac{2k\mathscr{C}_1\Gamma}{\zeta'^2}\sum_{w\in\mathscr{W}} \frac{k\operatorname{tr}\left(P(w)\,\Omega_n P^T(w)\right)}{\left(nR_n(w)\right)^2}\\ &\le \frac{2k\mathscr{C}_1\Gamma}{\zeta'^2}\sum_{w\in\mathscr{W}} \frac{1}{\left(nR_n(w)\right)},\end{aligned} \tag{54}$$

$$\begin{aligned}\Pr&\left\{\sup_{w\in\mathscr{W}} \frac{\left\langle \mu^T M^T(w) - \eta^T(w), P(w)e\right\rangle_k}{nR_n(w)} > \epsilon'\right\}\\ &\le \sum_{w\in\mathscr{W}} \frac{E\left[\left\langle \mu^T M^T(w) - \eta^T(w), P(w)e\right\rangle_k\right]^2}{\left(nR_n(w)\epsilon'\right)^2}\\ &\le \sum_{w\in\mathscr{W}} \frac{\|P(w)(M(w)\mu - \eta(w))\|_k^2 k\mathfrak{W}}{\left(nR_n(w)\epsilon'\right)^2}\\ &\le \epsilon'^{-2}k\mathfrak{W}\sum_{w\in\mathscr{W}} n^{-2}\|M(w)\mu - \eta(w)\|_k^2\left(R_n(w)\right)^{-2}.\end{aligned} \tag{55}$$

Combining (36) and (45), one knows that both (54) and (55) tend to zero. In other words, (43) is confirmed. We conclude that the expressions (41), (42), and (43) are reasonable. This completes the proof of Theorem 9. □

In practice, the covariance of errors is usually unknown and needs to be estimated. However, it is difficult to build a good estimate for Ω_n in virtue of the special structure of Ω_n. In the special case, when the random errors are independent and identical distribution with variance σ^2, the consistent estimator of σ^2 can be built for the constrained models (7) and (8). Let $m = Q$ in $\widehat{\mu}_m$ and $\tau' = \operatorname{tr}(P_Q)$, where Q corresponds to a "large" approximating model. Denote $\widehat{\sigma}^2_Q = (n-\tau')^{-1}(Y - \widehat{\mu}_Q)^T(Y - \widehat{\mu}_Q)$. The coming theorem will show that $\widehat{\sigma}^2_Q$ is a consistent estimate of σ^2.

Theorem 10. *If $\tau'/n \to 0$ when $\tau' \to \infty$ and $n \to \infty$, we have $\widehat{\sigma}^2_Q \xrightarrow{p} \sigma^2$ as $n \to \infty$.*

Proof. Writing $\Lambda = D + XR^{-1}D^{\star}$, one obtains

$$\frac{\left(Y - \widehat{\mu}_Q\right)^T\left(Y - \widehat{\mu}_Q\right)}{n - \tau'} = \frac{e^T M_Q e}{n - \tau'} + \frac{\|M_Q\Lambda\|^2}{n - \tau'} + \frac{2e^T M_Q\Lambda}{n - \tau'}, \tag{56}$$

where $M_Q = I - P_Q$.

Since $E(e^T M_Q e) = \sigma^2(n - \tau')$, it leads to

$$\begin{aligned}E\left|e^T M_Q e - \sigma^2(n - \tau')\right|^2 &= E\left|e^T M_Q e - E(e^T M_Q e)\right|^2\\ &= \operatorname{Var}\left(e^T M_Q e\right) = 2\sigma^4\operatorname{tr}\left(M_Q\right).\end{aligned} \tag{57}$$

For any $\varepsilon > 0$, it follows from (57) that

$$\begin{aligned}\Pr\left\{\left|\frac{e^T M_Q e}{n - \tau'} - \sigma^2\right| > \varepsilon\right\} &\le \frac{E\left|e^T M_Q e - (n - \tau')\sigma^2\right|^2}{\varepsilon^2(n - \tau')^2}\\ &\le \frac{2\sigma^4}{\varepsilon^2\left(n - \tau'\right)} \longrightarrow 0.\end{aligned} \tag{58}$$

Equation (58) implies that $(n - \tau')^{-1}(e^T M_Q e) \xrightarrow{p} \sigma^2$. Let $\mathfrak{J}_{Q,t}$ be the tth diagonal element of the "hat" matrix P_Q. Then, $0 \le 1 - \mathfrak{J}_{Q,t} < 1$ is the tth diagonal element of M_Q and satisfies $\sum_{t=1}^{n}(1 - \mathfrak{J}_{Q,t}) = n - \tau'$.

Because $\phi(z_{i,l}, \psi_l)$ and $\phi(a_{s,k}, \psi_k)$ converge to mean square, we have $E(\Lambda^2_{Q,t}) \to 0$ as $\tau' \to \infty$, where $\Lambda_{Q,t}$ is the tth element of Λ, $t = 1, \dots, n$. Notice that

$$\begin{aligned}\frac{E\left(\Lambda^T M_Q\Lambda\right)}{(n - \tau')} &= \frac{1}{n - \tau'}\sum_{t=1}^{n} E\left(\Lambda^2_{Q,t} M_{Q,t}\right)\\ &= \frac{1}{n - \tau'}\sum_{t=1}^{n} E\left(\Lambda^2_{Q,t}\right)\left[1 - \mathfrak{J}_{Q,t}\right]\\ &\le \max_t\left\{E\left(\Lambda^2_{Q,t}\right)\right\}\frac{1}{n - \tau'}\sum_{t=1}^{n}\left[1 - \mathfrak{J}_{Q,t}\right]\\ &= \max_t\left\{E\left(\Lambda^2_{Q,t}\right)\right\} \longrightarrow 0.\end{aligned} \tag{59}$$

The above expression implies that the second term on the right hand of (56) approaches to zero. By the similar proof of (59), we obtain that the final term on the right-hand of (56) also tends to zero. The proof of Theorem 10 is complete. □

In the case of independently and identically distributed errors, if we replace σ^2 by $\widehat{\sigma}_Q^2$ in the k-GIC, the k-GIC can be simplified to

$$\overline{C}_n(w) = \frac{\|Y - \widehat{\mu}(w)\|_k^2}{n} + \frac{2\hbar\widehat{\sigma}_Q^2 \operatorname{tr}(P(w))}{n}. \tag{60}$$

Here, we intend to illustrate that the model selection procedure of minimizing $\overline{C}_n(w)$ is also asymptotically optimal.

Theorem 11. *Assume that the random error e is i.i.d with mean zero and variance* σ^2*. Under the conditions* (35) *and* (36)*, the* $\overline{C}_n(w)$ *is still asymptotically valid.*

Proof. Using a similar technique of deriving (39), we obtain

$$\begin{aligned} n\overline{C}_n(w) &= \|Y - \widehat{\mu}(w)\|_k^2 + 2\hbar\widehat{\sigma}_Q^2 \operatorname{tr}(P(w)) \\ &= nL_n(w) + 2\left\langle e^T, M(w)\mu - \eta(w)\right\rangle_k \\ &\quad + 2\hbar\widehat{\sigma}_Q^2 \operatorname{tr}(P(w)) - 2\left\langle e^T, P(w)e\right\rangle_k + \|e\|_k^2 \\ &= 2\hbar\left(\widehat{\sigma}_Q^2 - \sigma^2\right)\operatorname{tr}(P(w)) + 2\left\langle e^T, M(w)\mu - \eta(w)\right\rangle_k \\ &\quad + nL_n(w) + 2\hbar\sigma^2 \operatorname{tr}(P(w)) \\ &\quad - 2\left\langle e^T, P(w)e\right\rangle_k + \|e\|_k^2. \end{aligned} \tag{61}$$

With an appropriate modification of the proofs (41)–(43), one only needs to verify

$$\sup_{w\in\mathscr{W}} \frac{\hbar\left|\sigma^2 - \widehat{\sigma}_Q^2\right| \operatorname{tr}(P(w))}{nR_n(w)} \xrightarrow{p} 0, \tag{62}$$

which is equivalent to showing

$$\sup_{w\in\mathscr{W}} \hbar^2 \frac{\left((1/n)\left|\sigma^2 - \widehat{\sigma}_Q^2\right| \operatorname{tr}(P(w))\right)^2}{R_n^2(w)} \xrightarrow{p} 0. \tag{63}$$

From the proof of Theorem 10, we have

$$\begin{aligned} &\Pr\left\{\left|\frac{\hbar e^T M_Q e}{n - \tau'} - \hbar\sigma^2\right| > \varepsilon' R_n^{-1/2}(w)\right\} \\ &\quad \le \hbar^2 \frac{E\left|e^T M_Q e - (n-\tau')\sigma^2\right|^2}{\varepsilon'^2 (n-\tau')^2 R_n(w)} \\ &\quad \le \frac{2\hbar^2\sigma^4}{\varepsilon'^2 (n-\tau') R_n(w)}, \end{aligned} \tag{64}$$

where $\varepsilon' > 0$.

By assumption (35), one knows that the above equation is close to zero. Thus, we obtain $\{R_n(w)\}^{-1}|\hbar\sigma^2 - \hbar\widehat{\sigma}_Q^2|^2 \xrightarrow{p} 0$. It follows from Lemma 3 and the definition of $R_n(w)$ that $(n^{-1}\operatorname{tr}(P(w)))^2$ is no greater than $R_n(w)$. Then, the formula (63) is confirmed. Therefore, we conclude that the minimizing of $\overline{C}_n(w)$ is also asymptotically optimal. □

5. Monte Carlo Simulation

In this section, Monte Carlo simulations are performed to investigate the finite sample properties of the proposed restricted linear model selection. This data generating process is

$$\begin{aligned} y_i &= x_i\theta + \sum_{l=1}^{1000} z_{i,l}\psi_l + e_i, \\ \text{subject to } \theta &= -\sum_{l=1}^{1000} \psi_l + 1, \quad i = 1, \ldots, n, \end{aligned} \tag{65}$$

where x_i follows $t(3)$ distribution, $z_{i,1} = 1$, and $z_{i,l}$, $l = 2, \ldots, 1000$ is independently and identically distributed $N(0,1)$. The parameter ψ_l, $l = 1, \ldots, 1000$ is determined by the rule $\psi_l = \zeta l^{-1}$, where ζ is a parameter which is selected to control the population $\widetilde{R}^2 = \zeta^2/(1+\zeta^2)$. The error e_i is independent of x_i and $z_{i,l}$. We consider two cases of the error distribution.

Case 1. e_i is independently and identically distributed $N(0,1)$.

Case 2. $e_1, \ldots, e_n$ obey multivariate normal distribution $N(0, \Omega_n)$, where Ω_n is a $n \times n$ dimensional covariance matrix. The jth diagonal element of Ω_n is σ_j^2 generated from uniform distribution on $(0,1)$. The jlth $(j \ne l)$ nondiagonal element of Ω_n is $\Omega_n^{jl} = \exp(-0.5|x_{ij} - x_{il}|^2)$, where x_{ij} and x_{il} denote the jth and lth elements of x_i, respectively.

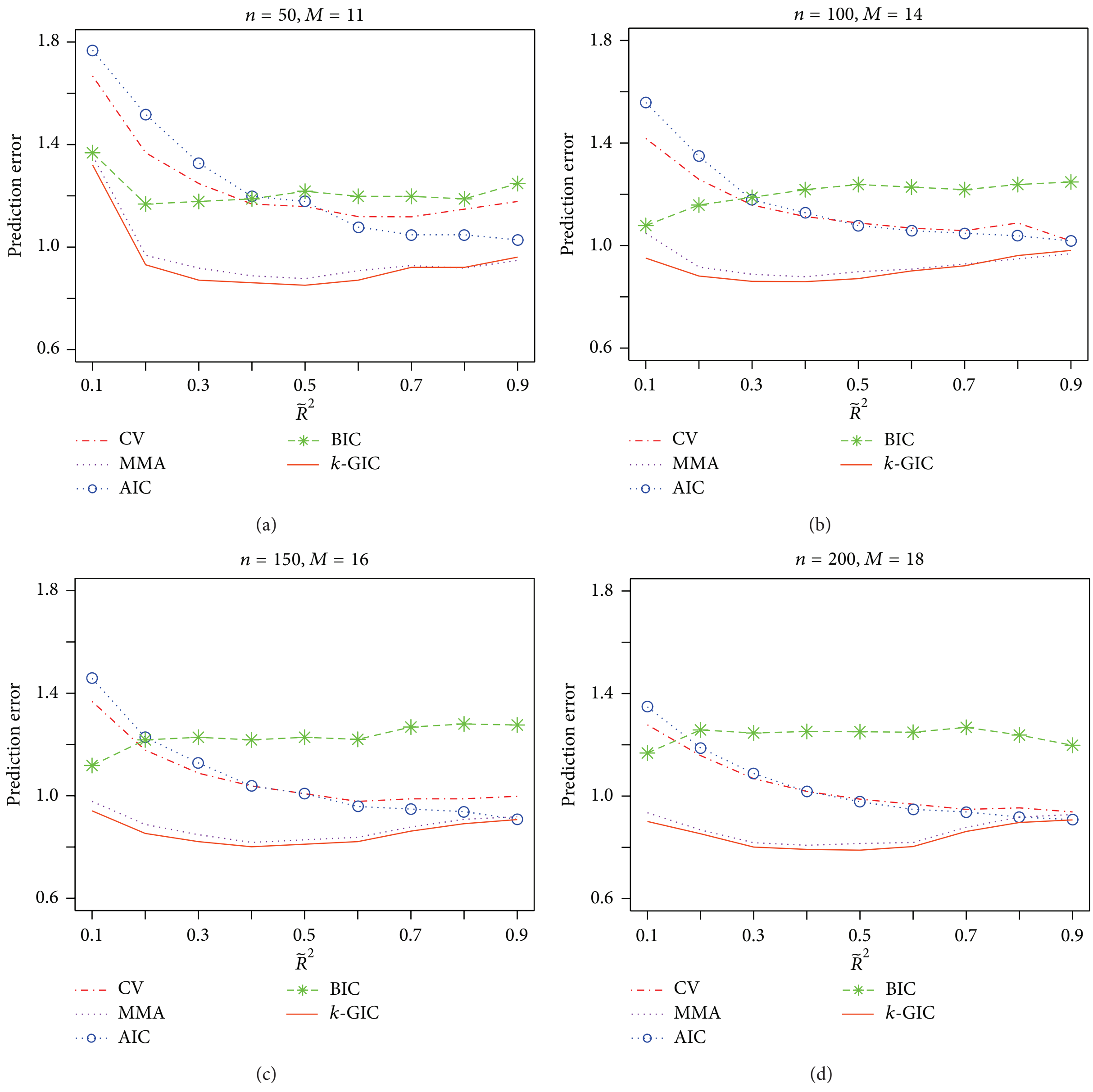
The sample size is varied between $n = 50, 100, 150$ and 200. The number of models is determined by $M = \lfloor 3n^{1/3} \rfloor$, where $\lfloor s \rfloor$ stands for the integer part of s. We set ζ so that $\widetilde{R}^2$ varies on a grid between 0.1 and 0.9. The number of simulation trials is $\Pi = 500$. For the k-GIC, the value of k takes one and $\hbar$ adopts the effective number of parameters.

To assess the performance of k-GIC, we consider five estimators which are (1) AIC model selection estimators (AIC), (2) BIC model selection estimators (BIC), (3) leave-one-out cross-validated model selection estimator (CV, [17]), (4) Mallows model averaging estimators (MMA, [11]), and (5) k-GIC model selection estimator (k-GIC). Following Machado [30], the AIC and BIC are defined, respectively, as

$$\begin{aligned} \mathrm{AIC}_m &= 2n\ln\left\{\frac{1}{n}\left\|Y - \widehat{\mu}_m\right\|^2\right\} + 2\mathscr{K}_m, \\ \mathrm{BIC}_m &= 2n\ln\left\{\frac{1}{n}\left\|Y - \widehat{\mu}_m\right\|^2\right\} + \ln(n)\,\mathscr{K}_m. \end{aligned} \tag{66}$$

We employ the out-of-sample prediction error to evaluate each estimator. For each replication, $\{y_\ell, x_\ell, z_\ell\}_{\ell=1}^{100}$ are generated as out-of-sample observations. In the πth simulation, the prediction error is

$$\mathrm{PE}(\pi) = \frac{1}{100}\sum_{\ell=1}^{100}\left(y_\ell - \widehat{\mu}_\ell(\widehat{w})\right)^2, \tag{67}$$



(a)

(b)

(c)

(d)

FIGURE 1: Results for Monte Carlo design.

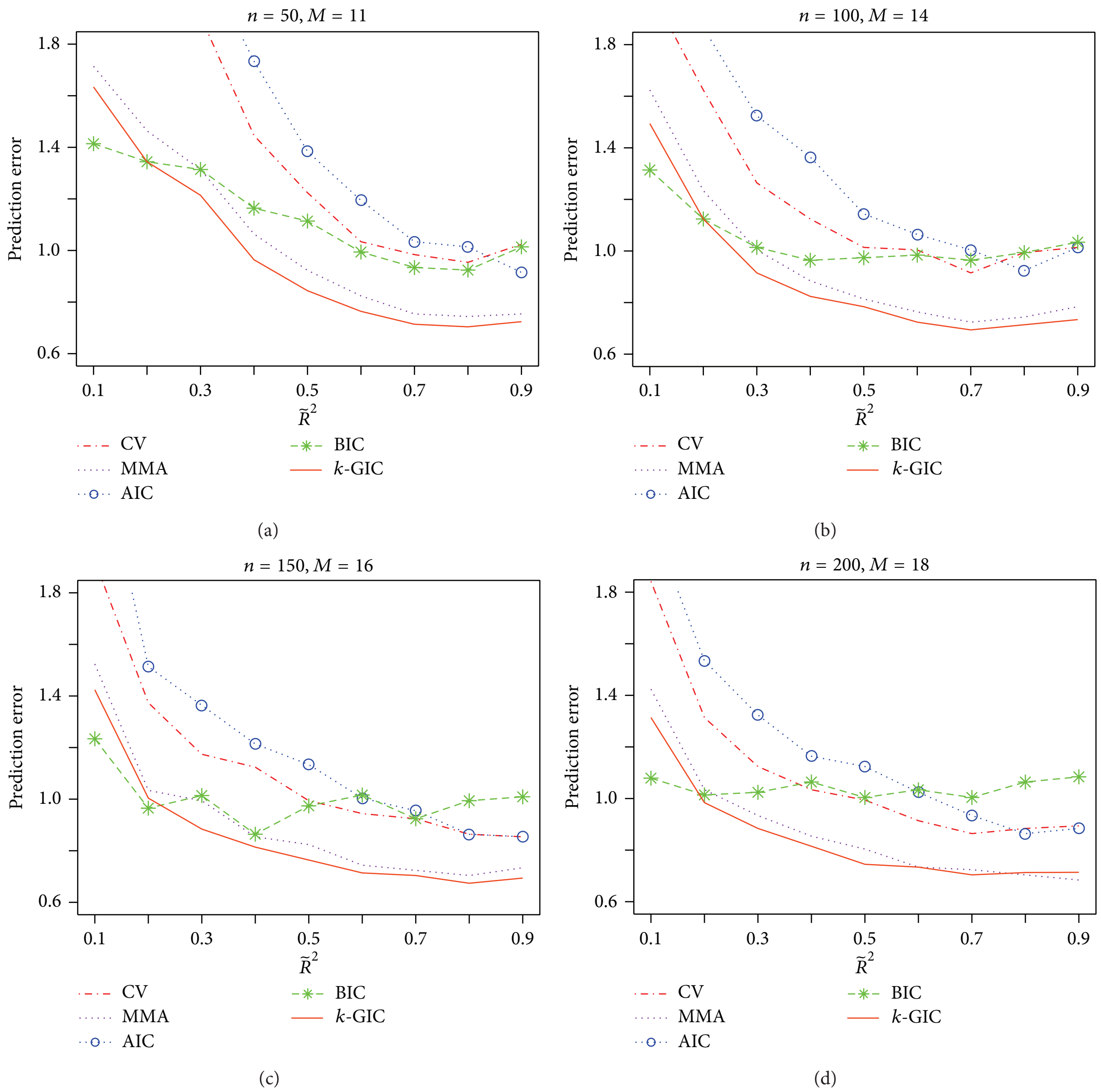
where $\widehat{w}$ is selected by one of the five methods. Then, the out-of-sample prediction error is calculated by

$$\mathrm{PE} = \frac{1}{\Pi} \sum_{\pi=1}^{\Pi} \mathrm{PE}(\pi), \tag{68}$$

where $\Pi = 500$ is the number of replication. Obviously, the smaller PE implies the better method of model estimator. We consider PE under homoscedastic errors at first. The prediction error calculations are summarized in Figure 1. The four panels in each graph depict results for a variety of sample sizes. In each panel, PE is displayed on the y-axis and $\widetilde{R}^2$ is displayed on the x-axis.

We find that the k-GIC estimators are almost the best estimators among those considered. When $\widetilde{R}^2$ is very large, the MMA estimators can sometimes be marginally preferred to the k-GIC estimators. In each panel, the AIC and CV have quite similar prediction errors. For a smaller $\widetilde{R}^2$, the AIC obtains a higher prediction error than CV. However, the AIC estimators yield smaller PEs than the CV estimators, when $\widetilde{R}^2$ is increasing. In many situations, the PEs of BIC estimator with a large $\widetilde{R}^2$ are quite poor relative to the other methods.

Next, we discuss PE under correlative errors and the PE calculation is summarized in Figure 2. Broadly speaking, the conclusions are similar to those found in homoscedastic cases. The k-GIC estimator frequently yields the most accurate estimators followed by the MMA estimator, and both average estimators enjoy significantly smaller PEs than the other three estimators over a large portion of the $\widetilde{R}^2$ space.



(a)

(b)

(c)

(d)

Figure 2: Results for Monte Carlo design.

When $\tilde{R}^2 \leq 0.2$, the BIC estimator outperforms the k-GIC estimator. Again, the AIC estimator is habitually the worst performing estimator with the CV being a close second in a large region of the $\tilde{R}^2$ space. Besides, their relative efficiency relies closely on sample size with the BIC estimator revealing increasing PE and the remaining four estimators showing decreasing PE, as n increases.

6. Conclusions

In risk investment, an important subject is to find an optimal portfolio. The commonly used techniques are the optimization methods based on the scheme of mean-variance. However, those methods are cumbersome in computing and cannot obtain the closed solutions for some complex problems. To make up the defects of mean-variance, an alternative methodology for obtaining an optimal portfolio is to use model selection. This paper attempts to develop a statistical program to consider the selection problem of optimal tracking portfolio. We build the theoretical models of tracking portfolios by constrained linear models. Then, the selection problems of optimal portfolio boil down to choosing an optimal constrained linear model.

In the setting of unrestricted models or homoscedastic models, a large number of works investigate the problems of model selection. In distinction, we discuss the model selection for constrained models with dependent errors. The restricted models are estimated by the method of weighted average least squares. Thus, the selection of an optimal constrained model is equivalent to finding a series of optimal weights. We select the weights by minimizing a k-class generalized information criterion (k-GIC), which is an estimate of the average squared error from the model average fit. The procedure of selecting weights is proved to be asymptotically optimal. Through Monte Carlo simulation, the performance of k-GIC is compared against that of four other methods. It is found that the k-GIC gives the best performance in most cases.

There are two limitations of our results which are open for further research. First, what is the asymptotic distribution of the parametric estimators? Second, can the theory be generalized to allow for continuous weights? These questions remain to be answered by future research. In this work, we mainly adopt the method of regression analysis to solve the selection problem. In fact, some alternative mathematical tools can also be employed to explore the theoretical properties of model selection. For example, the optimal model can be selected by the methods of linear optimization or quadratic programming and we can apply the techniques of linear functional analysis and stochastic control to consider the inferences of parametric estimator. Besides, we mention that the applications of this study can also be extended to some other fields including risk management, ruin theory, and factor analysis.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the Shandong Institute of Business and Technology (SIBT Grant 521014306203).

References

[1] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *The Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[2] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.

[3] H. Zou and M. Yuan, "Composite quantile regression and the oracle model selection theory," *The Annals of Statistics*, vol. 36, no. 3, pp. 1108–1126, 2008.

[4] M. Caner, "Lasso-type GMM estimator," *Econometric Theory*, vol. 25, no. 1, pp. 270–290, 2009.

[5] C. Y. Tang and C. Leng, "Penalized high-dimensional empirical likelihood," *Biometrika*, vol. 97, no. 4, pp. 905–919, 2010.

[6] L. C. Jennifer, A. D. Jurgen, and F. H. David, *Model Selection in Equations With Many Small Effects*, Oxford Bulletin of Economics and Statistics, 2013.

[7] E. E. Leamer, *Specification Searches*, John Wiley & Sons, New York, NY, USA, 1978.

[8] P. J. Brown, M. Vannucci, and T. Fearn, "Bayes model averaging with selection of regressors," *Journal of the Royal Statistical Society B. Statistical Methodology*, vol. 64, no. 3, pp. 519–536, 2002.

[9] W. S. Rodney and K. D. Herman, *Bayesian Model Selection With An Uninformative Prior*, Oxford Bulletin of Economics and Statistics, 2003.

[10] N. L. Hjort and G. Claeskens, "Frequentist model average estimators," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[11] B. E. Hansen, "Least squares model averaging," *Econometrica. Journal of the Econometric Society*, vol. 75, no. 4, pp. 1175–1189, 2007.

[12] H. Liang, G. Zou, A. T. K. Wan, and X. Zhang, "Optimal weight choice for frequentist model average estimators," *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 1053–1066, 2011.

[13] X. Zhang and H. Liang, "Focused information criterion and model averaging for generalized additive partial linear models," *The Annals of Statistics*, vol. 39, no. 1, pp. 194–200, 2011.

[14] B. E. Hansen and J. S. Racine, "Jackknife model averaging," *Journal of Econometrics*, vol. 167, no. 1, pp. 38–46, 2012.

[15] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, B. N. Petrov and F. Csake, Eds., pp. 267–281, Akademiai Kiado, Budapest, Hungary, 1973.

[16] C. L. Mallows, "Some comments on CP," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[17] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society B. Methodological*, vol. 36, pp. 111–147, 1974.

[18] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[19] P. Craven and G. Wahba, "Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1979.

[20] D. W. K. Andrews, "Consistent moment selection procedures for generalized method of moments estimation," *Econometrica. Journal of the Econometric Society*, vol. 67, no. 3, pp. 543–564, 1999.

[21] G. Claeskens and N. L. Hjort, "The focused information criterion," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 900–945, 2003, With discussions and a rejoinder by the authors.

[22] Y. Zhang, R. Li, and C.-L. Tsai, "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 312–323, 2010.

[23] G. Xu and J. Z. Huang, "Asymptotic optimality and efficient computation of the leave-subject-out cross-validation," *The Annals of Statistics*, vol. 40, no. 6, pp. 3003–3030, 2012.

[24] M. K. P. So and T. Ando, "Generalized predictive information criteria for the analysis of feature events," *Electronic Journal of Statistics*, vol. 7, pp. 742–762, 2013.

[25] J. J. J. Groen and G. Kapetanios, "Model selection criteria for factor-augmented regressions," *Oxford Bulletin of Economics and Statistics*, vol. 75, no. 1, pp. 37–63, 2013.

[26] S. Lai and Q. Xie, "A selection problem for a constrained linear regression model," *Journal of Industrial and Management Optimization*, vol. 4, no. 4, pp. 757–766, 2008.

[27] J. Shao, "An asymptotic theory for linear model selection," *Statistica Sinica*, vol. 7, no. 2, pp. 221–264, 1997, With comments and a rejoinder by the author.

[28] D. W. K. Andrews, "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, vol. 47, no. 2-3, pp. 359–377, 1991.

[29] K.-C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set," *The Annals of Statistics*, vol. 15, no. 3, pp. 958–975, 1987.

[30] J. A. F. Machado, "Robust model selection and M-estimation," *Econometric Theory*, vol. 9, no. 3, pp. 478–493, 1993.