Statistica Sinica **25** (2015), 1583-1598
doi:http://dx.doi.org/10.5705/ss.2013.326

MODEL AVERAGING BASED ON KULLBACK-LEIBLER DISTANCE

Xinyu Zhang[1], Guohua Zou[1,2] and Raymond J. Carroll[3]

[1]*Chinese Academy of Sciences,* [2]*Capital Normal University and* [3]*Texas A&M University*

Abstract: This paper proposes a model averaging method based on Kullback-Leibler distance under a homoscedastic normal error term. The resulting model average estimator is proved to be asymptotically optimal. When combining least squares estimators, the model average estimator is shown to have the same large sample properties as the Mallows model average (MMA) estimator developed by Hansen (2007). We show via simulations that, in terms of mean squared prediction error and mean squared parameter estimation error, the proposed model average estimator is more efficient than the MMA estimator and the estimator based on model selection using the corrected Akaike information criterion in small sample situations. A modified version of the new model average estimator is further suggested for the case of heteroscedastic random errors. The method is applied to a data set from the Hong Kong real estate market.

Key words and phrases: Akaike information, Kullback-Leibler distance, model averaging, model selection, prediction.

1. Introduction

Model averaging is an alternative to model selection for dealing with model uncertainty. By minimizing a model selection criterion, such as C_p (Mallows (1973)), AIC (Akaike (1973)), and BIC (Schwarz (1978)), one model can be chosen from a set of candidate models, but we end up "putting all our inferential eggs in one unevenly woven basket" (Longford (2005)). Model averaging often reduces the risk in regression estimation, as "betting" on multiple models provides a type of insurance against a singly selected model being poor (Leung and Barron (2006)). Additionally, it is often the case that several models fit the data equally well, but may differ substantially in terms of the variables included and may lead to different predictions (Miller (2002)). Combining these models seems to be more reasonable than choosing one of them. Averaging weights can be based on the scores of information criteria (Buckland, Burnham and Augustin (1997), Hjort and Claeskens (2003), Claeskens, Croux and van Kerckhoven (2006), Zhang and Liang (2011), Zhang, Wan, and Zhou (2012)). Other model

averaging strategies that have been developed include, for example, the adaptive regression by mixing of Yang (2001), the Mallows model averaging (MMA) of Hansen (2007) (see also Wan, Zhang, and Zou (2010)), and the optimal mean squared error averaging of Liang et al. (2011).

The C_p and AIC are both widely used criteria in model selection. The former was developed from prediction of "scaled sum of squared errors" (Mallows (1973)), and the latter was produced by an approximately unbiased estimator of the expected Kullback-Leibler (KL) distance (Akaike (1973)). In addition, GIC (Konishi and Kitagawa (1996)), KIC (Cavanaugh (1999)), and RIC (Shi and Tsai (2004)) were also developed from the KL distance. Recently, Hansen (2007) utilized the C_p criterion in model averaging (called Mallows' criterion) and presented the asymptotic optimality of the resulting MMA estimator. Motivated by these facts, proposing a novel model averaging approach from estimating the expected KL distance seems to be feasible and potentially interesting. From Shao (1997), C_p and AIC can be classified into the same class according to their asymptotic behaviors. Thus, the new approach is expected to have the same asymptotic optimality as MMA.

Hurvich and Tsai (1989) proposed a corrected version of AIC, AICc, that is an exactly unbiased estimator of the expected KL distance in linear models with normally homoscedastic error and thus has advantages over AIC and C_p under small sample situations. Following this observation, our approach is based on an unbiased estimator of the expected KL distance from the averaging model (the model with parameters estimated by model averaging) to the true data generating process, thus our approach is further expected to have advantages over MMA under small sample situations, which is verified by our simulation study. A referee mentioned that the choice of weights via a Kullback-Leibler distance was proposed in an entirely different context by Rigollet (2012), in which non-random vectors are aggregated and risk inequalities were proved.

More recently, to average estimators under a heteroscedasticity setting, Hansen and Racine (2012) proposed a jackknife model averaging (JMA) method. Liu and Okui (2013) suggested a Mallows' C_p-like criterion for a heteroscedasticity setting and referred to their method as heteroscedasticity-robust C_p model averaging. In the current paper, we further modify our approach for averaging estimators for a heteroscedasticity setting.

The remainder of this paper is organized as follows. Section 2 introduces a weight choice criterion from estimating the KL distance and proves the asymptotic optimality of the resulting model average estimator. Section 3 extends the new method to the setting with heteroscedastic errors. Section 4 investigates the finite sample performance of the proposed model average estimators through extensive simulation studies. Section 5 applies the model average estimators to

an empirical example. Section 6 has concluding remarks. Assumptions for the theoretical properties are provided in an Appendix and the proofs are reported in the Supplementary Material.

2. Weight Choice Criterion from KL Distance

Consider the data generating process

$$y = \mu + e, \tag{2.1}$$

where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is an $n \times 1$ vector of observations, $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is the mean vector of y, and $e = (e_1, \ldots, e_n)^{\mathrm{T}}$ with the e_i's independent with mean zeros and variance σ^2. We assume that e has a multivariate normal distribution when developing weight choice criteria, but the normality assumption is unnecessary when proving asymptotic optimality of the resulting model average estimators.

Assume that there are S candidate models used to approximate the data generating process given in (2.1). Write $\widehat{\mu}_{(s)}$ as the estimator of μ based on the s^{th} candidate model. Let the weight vector $w = (w_1, \ldots, w_S)^{\mathrm{T}}$, belonging to the set $\mathcal{W} = \{w \in [0,1]^S : \sum_{s=1}^S w_s = 1\}$. The model average estimator of μ is written as $\widehat{\mu}(w) = \sum_{s=1}^S w_s \widehat{\mu}_{(s)}$. Denote $\widehat{\sigma}^2$ as an estimator of σ^2.

Let f and g be the true density of the distribution generating the data y, and the density of the model fitting the data, respectively. The KL distance between them is given by $I(f,g) = E_{f(y)}\{\log f(y)\} - E_{f(y)}\{\log g(y|\theta)\}$, where θ includes unknown parameters. Suppose that $\widehat{\theta}(y)$ is an estimator of θ. Then, the expected KL distance is

$$E_{f(y)}\{I(f, g_{\widehat{\theta}(y)})\} = E_{f(y^*)}\{\log f(y^*)\} - E_{f(y)}(E_{f(y^*)}[\log g\{y^*|\widehat{\theta}(y)\}]),$$

where y^* is another realization from f and independent of y. Ignoring the constant $E_{f(y^*)}\{\log f(y^*)\}$, the fit of $g\{y|\widehat{\theta}(y)\}$ can be assessed using the Akaike information (AI): $\mathrm{AI} = -2E_{f(y)}(E_{f(y^*)}[\log g\{y^*|\widehat{\theta}(y)\}])$. Here, the fitting model is assumed to be normally distributed and the unknown parameters in (2.1) are estimated by $\widehat{\theta}(y) = \{\widehat{\mu}(w), \widehat{\sigma}^2\}$. Thus, we write the Akaike information as

$$\begin{aligned}
\mathrm{AI}(w) &= -2E_{f(y)}(E_{f(y^*)}[\log g\{y^*|\widehat{\theta}(y)\}]) \\
&= E_{f(y)}\left[E_{f(y^*)}\left\{n\log 2\pi + n\log\widehat{\sigma}^2 + \|y^* - \widehat{\mu}(w)\|^2\widehat{\sigma}^{-2}\right\}\right] \\
&= E_{f(y)}\left\{n\log 2\pi + n\log\widehat{\sigma}^2 + \|\mu - \widehat{\mu}(w)\|^2 + \sigma^2 n\widehat{\sigma}^{-2}\right\}. \qquad (2.2)
\end{aligned}$$

Define

$$\begin{aligned}
\mathcal{B}(w) = {} & n\log 2\pi + n\log\widehat{\sigma}^2 + \|y - \widehat{\mu}(w)\|^2\widehat{\sigma}^{-2} + 2\sigma^2\widehat{\sigma}^{-2}\mathrm{trace}\left(\frac{\partial\widehat{\mu}(w)}{\partial y^{\mathrm{T}}}\right) \\
& + \frac{2\sigma^2}{\widehat{\sigma}^4}\{y - \widehat{\mu}(w)\}^{\mathrm{T}}\frac{\partial\widehat{\sigma}^2}{\partial y} + \frac{2\sigma^4}{\widehat{\sigma}^6}\mathrm{trace}\left(\frac{\partial\widehat{\sigma}^2}{\partial y}\frac{\partial\widehat{\sigma}^2}{\partial y^{\mathrm{T}}}\right) - \frac{\sigma^4}{\widehat{\sigma}^4}\mathrm{trace}\left(\frac{\partial^2\widehat{\sigma}^2}{\partial y\partial y^{\mathrm{T}}}\right).
\end{aligned}$$

Although the definition of $\mathcal{B}(w)$ appears complicated, the idea behind it is simple. For the purpose of selecting good weights, one should minimize $\mathrm{AI}(w)$ with $w \in \mathcal{W}$. But $\mathrm{AI}(w)$ involves unknown moments of various random variables. So, we attempt to find an unbiased estimator of $\mathrm{AI}(w)$, which is just $\mathcal{B}(w)$.

Theorem 1. *If $\widehat{\sigma}^2$ and $\partial\widehat{\sigma}^2/\partial y$ are continuous functions with piecewise continuous partial derivatives with respect to y, the expectation of $\mathcal{B}(w)$ exists, and e has a multivariate normal distribution, then for any $w \in \mathcal{W}$, $E\{\mathcal{B}(w)\} = \mathrm{AI}(w)$.*

We focus on the case that $\widehat{\mu}_{(s)}$ is linear with respect to y, $\widehat{\mu}_{(s)} = P_{(s)}y$, where the matrix $P_{(s)}$ is not related to y. This class of estimators includes least squares, ridge regression, Nadaraya-Watson and local polynomial kernel regression with fixed bandwidths, nearest neighbor estimators, series estimators, and spline estimators (Hansen and Racine (2012)). Let $P(w) = \sum_{s=1}^{S} w_s P_{(s)}$, so that $\widehat{\mu}(w) = P(w)y$.

When σ^2 is known, $\mathcal{B}(w)$ can be simplified to

$$n\log 2\pi + n\log\sigma^2 + \sigma^{-2}\|y - \widehat{\mu}(w)\|^2 + 2\mathrm{trace}\{P(w)\},$$

which, in the sense of weight choice, is equivalent to the Mallows' criterion of Hansen (2007) for the situation with known σ^2.

In practice, σ^2 is unknown. We can estimate it directly by $\widehat{\sigma}^2$, which is required to satisfy Assumptions (A.4)−(A.5) in the appendix. For simplicity, we further assume that $\widehat{\sigma}^2$ is unrelated to w, which means that $\widehat{\sigma}^2$ is not from model averaging as in the existing literature, such as Hansen (2007) and Liang et al. (2011). After removing the terms unrelated to w and multiplying by $\widehat{\sigma}^2$, $\mathcal{B}(w)$ reduces to

$$\mathcal{B}^*(w) \equiv \|y - P(w)y\|^2 + 2\widehat{\sigma}^2\mathrm{trace}\{P(w)\} - 2y^{\mathrm{T}}P^{\mathrm{T}}(w)\frac{\partial\widehat{\sigma}^2}{\partial y}, \tag{2.3}$$

which can be taken as a criterion for choosing weights. We let $w^* = \underset{w\in\mathcal{W}}{\mathrm{argmin}}\{\mathcal{B}^*(w)\}$, the resulting weights by minimizing the criterion $\mathcal{B}^*(w)$.

The predictive squared error in estimating μ is $L_n(w) = \|\widehat{\mu}(w) - \mu\|^2$. We can show the asymptotic optimality of $\widehat{\mu}(w^*)$ in the sense that $\widehat{\mu}(w^*)$ yields a squared error that is asymptotically identical to that of the infeasible optimal model average estimator. Unless otherwise stated, all limiting processes discussed are with respect to $n \to \infty$.

Theorem 2. *If Assumptions* (A.1)−(A.5) *in the Appendix are satisfied, then*

$$L_n(w^*)\{\inf_{w\in\mathcal{W}} L_n(w)\}^{-1} = 1 + o_p(1).$$

The direct use of $\widehat{\sigma}^2$ in $\mathcal{B}^*(w)$ instead of σ^2 makes $\mathcal{B}^*(w)$ not unbiased for estimating AI, up to a term unrelated to w. In what follows, we consider a situation where AI can be estimated unbiasedly using data, up to a term unrelated to w.

As in such model averaging papers as Hansen (2007), Wan, Zhang, and Zou (2010), Liang et al. (2011), and Hansen and Racine (2012), we now focus on least squares estimation with $P_{(s)} = X_{(s)}(X_{(s)}^{\mathrm{T}}X_{(s)})^{-}X_{(s)}^{\mathrm{T}}$, where $X_{(s)}$ is the covariate matrix in the s^{th} candidate model and $(X_{(s)}^{\mathrm{T}}X_{(s)})^{-}$ is a generalized inverse of $X_{(s)}^{\mathrm{T}}X_{(s)}$. Let $X = \left(X_{(1)}, \ldots, X_{(S)}\right)$, $m = \operatorname{rank}(X)$, and $P = X\left(X^{\mathrm{T}}X\right)^{-}X^{\mathrm{T}}$. We adopt $\widehat{\sigma}^2(y,k) = y^{\mathrm{T}}(I_n - P)y/k$ to estimate σ^2, where k is a positive constant. Consider the situation of μ being a linear function of X, $\mu = X\beta$. Then, $\widehat{\sigma}^2(y,n)$ is the maximum likelihood estimator of σ^2 and $\widehat{\sigma}^2(y,n-m)$ is an unbiased estimator of σ^2. Substitute $\widehat{\sigma}^2(y,k)$ for $\widehat{\sigma}^2$ in (2.2) and denote the resulting $\mathrm{AI}(w)$ as $\mathrm{AI}_k(w)$. Define

$$\begin{aligned}\mathcal{C}(w) \equiv{}& n\log 2\pi + n\log\widehat{\sigma}^2(y,k) + 2k(n-m-2)^{-1}\operatorname{trace}\{P(w)\}\\ &+\|y-\widehat{\mu}(w)\|^2\widehat{\sigma}^{-2}(y,k) + 4\sigma^2\widehat{\sigma}^{-2}(y,k) - 2k^{-1}(n-m-4)\sigma^4\widehat{\sigma}^{-4}(y,k).\end{aligned}$$

Because $\mathrm{AI}_k(w)$ involves unknown moments of various random variables, in a manner similar to that leading to Theorem 1, we derive its unbiased estimator, which is just $\mathcal{C}(w)$.

Theorem 3. *Suppose e has a multivariate normal distribution and μ is a linear function of X. For any $k > 0$, if the expectation of $\mathcal{C}(w)$ exists, then $E\{\mathcal{C}(w)\} = \mathrm{AI}_k(w)$.*

By removing the terms unrelated to w and multiplying by $\widehat{\sigma}^2(y,k)$, $\mathcal{C}(w)$ simplifies to

$$\mathcal{C}^*(w) \equiv \|y-\widehat{\mu}(w)\|^2 + 2y^{\mathrm{T}}(I_n - P)y(n-m-2)^{-1}\operatorname{trace}\{P(w)\},$$

which we refer to as the KL model averaging (KLMA) criterion. Let $\widehat{w} = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\{\mathcal{C}^*(w)\}$. The resulting model average estimator is called the KLMA estimator.

Remark 1. By comparing the criterion $\mathcal{C}^*(w)$ and the Mallows' criterion of Hansen (2007), the only difference is that $n-m-2$ is used here, while $n-m$ is used in Mallows' criterion. The quantity $n-m-2$ is from calculating the mean of the inverse Chi-squared distribution; see (S3.1) of the Supplementary Material. So the KLMA estimator will have the same large sample properties as the MMA estimator, and thus the asymptotic optimality of the MMA estimator presented by Hansen (2007) and Wan, Zhang, and Zou (2010) also holds for the

KLMA estimator. In particular, our Assumptions (A.1) and (A.4) are sufficient for the asymptotic optimality of the KLMA estimator and Assumptions (A.2), (A.3), and (A.5) are not necessary.

Remark 2. Let $c(w) = e'\{I_n - P(w)\}\mu + \sigma^2 \text{trace}\{P(w)\} - e'P(w)e$. Obviously, $|E\{c(w)\}| = 0$, but our weight vector $\widehat{w}$ is determined by data, so that $|E\{c(\widehat{w})\}|$ may not be zero. We show in the Supplementary Material that

$$E\{L_n(\widehat{w})\} \leq \inf_{w\in\mathcal{W}} E\{L_n(w)\} + |E\{c(\widehat{w})\}|, \tag{2.4}$$

which means that the expected predictive squared error by using $\widehat{w}$ is upper-bounded by the minimum expected error of model averaging estimators plus the term $|E\{c(\widehat{w})\}|$. This result holds for finite sample sizes. Similar results have been developed by Yang (2001) and Zhang, Lu and Zou (2013). If $\inf_{w\in\mathcal{W}} E\{L_n(w)\} \to \infty$, then the term $|c(\widehat{w})|$ is of order lower than $\inf_{w\in\mathcal{W}} E\{L_n(w)\}$ under some regularity conditions (Wan, Zhang, and Zou (2010)).

3. The KLMA Estimator under a Heteroscedastic Error Setting

When the covariance matrix of e, Ω, is a general diagonal matrix, it follows from (2.2) that the Akaike information is

$$\begin{aligned}\text{AI}_{hetero} &= E_{f(y)}\left(E_{f(y^*)}\left[n\log 2\pi + \log|\widehat{\Omega}| + \{y^* - \widehat{\mu}(w)\}^{\mathrm{T}}\widehat{\Omega}^{-1}\{y^* - \widehat{\mu}(w)\}\right]\right)\\ &= E_{f(y)}\left[n\log 2\pi + \log|\widehat{\Omega}| + \{\mu - \widehat{\mu}(w)\}^{\mathrm{T}}\widehat{\Omega}^{-1}\{\mu - \widehat{\mu}(w)\} + \text{trace}(\widehat{\Omega}^{-1}\Omega)\right],\end{aligned}$$

where $\widehat{\Omega}$ is an estimator of Ω and is also diagonal. Using similar conditions to those of Theorem 1 and the same argument as in the proof of Theorem 1, we see that

$$\begin{aligned}\mathcal{D}(w) \equiv\ & n\log 2\pi + \log|\widehat{\Omega}| + \{y - \widehat{\mu}(w)\}^{\mathrm{T}}\widehat{\Omega}^{-1}\{y - \widehat{\mu}(w)\}\\ & + 2\text{trace}\left\{\Omega\widehat{\Omega}^{-1}\frac{\partial\widehat{\mu}(w)}{\partial y^{\mathrm{T}}}\right\} + 2\{y - \widehat{\mu}(w)\}^{\mathrm{T}}\Omega\widehat{\Omega}^{-2}\widehat{a} + \widehat{\delta}\end{aligned} \tag{3.1}$$

has expectation AI_{hetero}, where $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_n)^{\mathrm{T}}$, $\widehat{a}_i = \partial\widehat{\Omega}_{ii}/\partial y_i$, $\widehat{\Omega}_{ii}$ is the i^{th} diagonal element of $\widehat{\Omega}$, and $\widehat{\delta}$ is a scale related to $\partial\widehat{\Omega}_{ii}/\partial y_i$ and $\partial^2\widehat{\Omega}_{ii}/\partial y_i^2$, but unrelated to w.

We focus on the case with $\widehat{\mu}(w) = P(w)y$. After removing some terms unrelated to w and estimating Ω by $\widehat{\Omega}$ in (3.1), $\mathcal{D}(w)$ reduces to

$$\mathcal{D}^*(w) \equiv \{y - \widehat{\mu}(w)\}^{\mathrm{T}}\widehat{\Omega}^{-1}\{y - \widehat{\mu}(w)\} + 2\text{trace}\{P(w)\} - 2y^{\mathrm{T}}P^{\mathrm{T}}(w)\widehat{\Omega}^{-1}\widehat{a}.$$

It is straightforward to show that when $\widehat{\Omega} = \widehat{\sigma}^2 I_n$, $\mathcal{D}^*(w)$ simplifies to $\mathcal{B}^*(w)$. Let $\widehat{w}_{\text{hetero}} = \underset{w\in\mathcal{W}}{\text{argmin}}\{\mathcal{D}^*(w)\}$, the resulting weights by minimizing $\mathcal{D}^*(w)$.

Under the heteroscedastic error setting, we define the predictive squared error in estimating μ as $L_{\text{hetero},n}(w) = \{\widehat{\mu}(w) - \mu\}^{\mathrm{T}}\Omega^{-1}\{\widehat{\mu}(w) - \mu\}$. A result is the asymptotic optimality of $\widehat{\mu}(\widehat{w}_{\text{hetero}})$ in the sense of minimizing $L_{\text{hetero},n}(w)$.

Theorem 4. *If Assumptions* (A.2) *and* (A.3)*, and Assumptions* (B.2)−(B.5) *in the Appendix are satisfied, then*

$$L_{\text{hetero},n}(\widehat{w}_{\text{hetero}})\{\inf_{w\in\mathcal{W}} L_{\text{hetero},n}(w)\}^{-1} = 1 + o_p(1). \tag{3.2}$$

When the structure of Ω is known and it is related to an unknown parameter vector η, $\Omega = \Omega(\eta)$, we can estimate Ω by the maximum likelihood (ML) approach based on the model with the largest number of covariates. Let $\widehat{\eta}$ be the ML estimator of η. Then $\widehat{a}_i = \partial\widehat{\Omega}_{ii}/\partial y_i = \partial\widehat{\eta}^{\mathrm{T}}/\partial y_i(\partial\widehat{\Omega}_{ii}/\partial\widehat{\eta})$. The Supplementary Material provides some formulas for calculating $\partial\widehat{\eta}^{\mathrm{T}}/\partial y$. The resulting estimator is referenced as version 1 modified KLMA (mKLMA$_1$) estimator.

When the structure of Ω is unknown, we use residuals from model averaging to estimate Ω. Specifically, we use a two-stage procedure to get the weights.

Stage 1. Estimate μ using the methods developed in Sections 2, then use the residual vector $y - \widehat{\mu}(w^*)$ for the estimation of Ω, where w^* is the weight vector minimizing $\mathcal{B}^*(w)$. Specifically, let $\widehat{\Omega}_{ii} = \{y_i - \widehat{\mu}(w^*)_i\}^2$, where y_i and $\widehat{\mu}(w^*)_i$ are the i^{th} elements of y and $\widehat{\mu}(w^*)$, respectively. Ignoring the randomness of w^*, we have $\widehat{a}_i = \partial\widehat{\Omega}_{ii}/\partial y_i = 2\{y_i - \widehat{\mu}(w^*)_i\}\{1 - P(w^*)_{ii}\}$, where $P(w^*)_{ii}$ is the i^{th} diagonal element of $P(w^*)$. When focusing on least squares model averaging, we utilize $\widehat{w}$ instead of w^*.

Stage 2. To obtain the weights, minimize

$$\begin{aligned}\mathcal{E}(w) \equiv{}& \{y - \widehat{\mu}(w)\}^{\mathrm{T}}\widehat{\Omega}^{-1}\{y - \widehat{\mu}(w)\} + 2\text{trace}\{P(w)\} - 4y^{\mathrm{T}}P^{\mathrm{T}}(w)\widehat{\Omega}^{-1}\\ &\times[\{y_1 - \widehat{\mu}(w^*)_1\}\{1 - P(w^*)_{11}\}, \ldots, \{y_n - \widehat{\mu}(w^*)_n\}\{1 - P(w^*)_{nn}\}]^{\mathrm{T}}.\end{aligned}$$

The resulting estimator is termed the version 2 modified KLMA (mKLMA$_2$) estimator.

4. Simulations

4.1. Homoscedastic error setting

We conducted simulation experiments to compare the small sample performance of the KLMA estimator and the MMA estimator under the homoscedastic error setting. The results from the estimator selected by AICc, a method that has been shown to perform better than C_p, AIC and BIC in model selection in small sample situations (see, for example, Hurvich and Tsai (1989) and Hurvich, Simonoff and Tsai (2002)), are also presented. In the first example, the number of covariates was fixed, while in the second example, it increased with the sample size n.

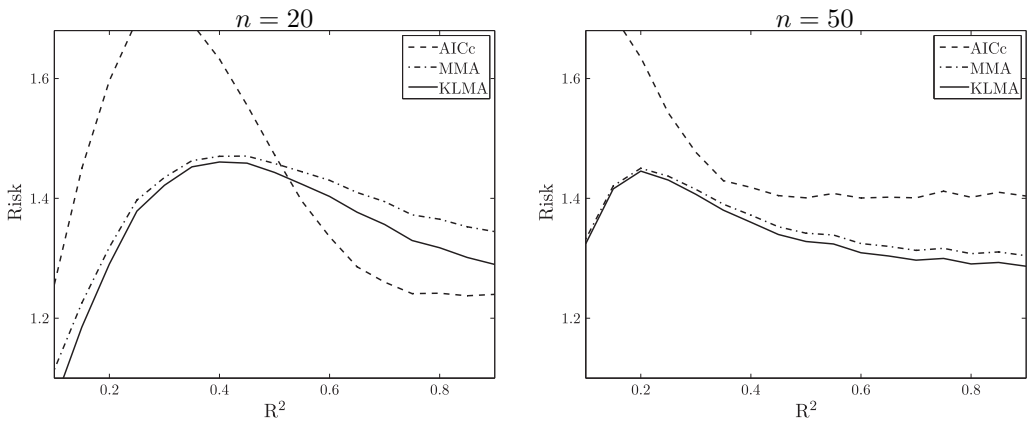

Figure 1. Results for Example 1: risk comparisons under L_μ as a function of R^2.

Example 1 (the fixed number of covariates)**.** This example is based on the setting of Hurvich and Tsai (1989): the model (2.1) with

$$\mu = X\beta, \quad \beta = (1, 2, 3, 0, 0, 0, 0)^{\mathrm{T}}, \quad \text{and} \quad X_j \sim \text{Normal}(0, I_n), j = 1, \ldots, 7,$$

where X_j is the j^{th} column of X. Seven candidate models were considered with $X_{(s)} = (X_1, \ldots, X_s)$, $s = 1, \ldots, 7$, respectively. Let $\mathrm{R}^2 = \text{Var}(\mu_i)/\text{Var}(y_i) = \text{Var}(\mu_i)/\{\text{Var}(\mu_i) + \sigma^2\} = 14/(14 + \sigma^2)$, controlled by σ^2. We varied σ^2 such that R^2 varied in the range $[0.1, 0.9]$. The estimator $\widehat{\mu}$ was evaluated in terms of its risk under the loss function $L_\mu = \|\widehat{\mu} - \mu\|^2$, the predictive loss of $\widehat{\mu}$. We did this by computing the average across 1,000 replications. The sample size n was 20 and 50.

The simulation results are shown in Figure 1. For clearer comparison, we normalized the risk by dividing by the risk of the infeasible optimal least squares estimator. It is encouraging that the KLMA has a lower risk than the MMA in the entire range of R^2 we considered, and the superiority is more obvious for $n = 20$. When $n = 50$, the two model average estimators have similar performance, which is expected as they have the same large sample properties. In most situations, the model averaging outperforms model selection by the AICc.

The estimators were also evaluated in terms of risk under the loss function $L_\beta = \|\widehat{\beta} - \beta\|^2$. The simulation results are presented in Section S8 of the Supplementary Material. The comparison results are analogous to those under L_μ and support our proposed KLMA.

Example 2 (an increasing number of covariates)**.** This example is based on the setting in Hansen (2007): $y_i = \mu_i + e_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + e_i$, $x_{1i} = 1$, all other x_{ji} are Normal$(0, 1)$, e_i is Normal$(0, 1)$, independent of x_{ji}, all x_{ji} are mutually

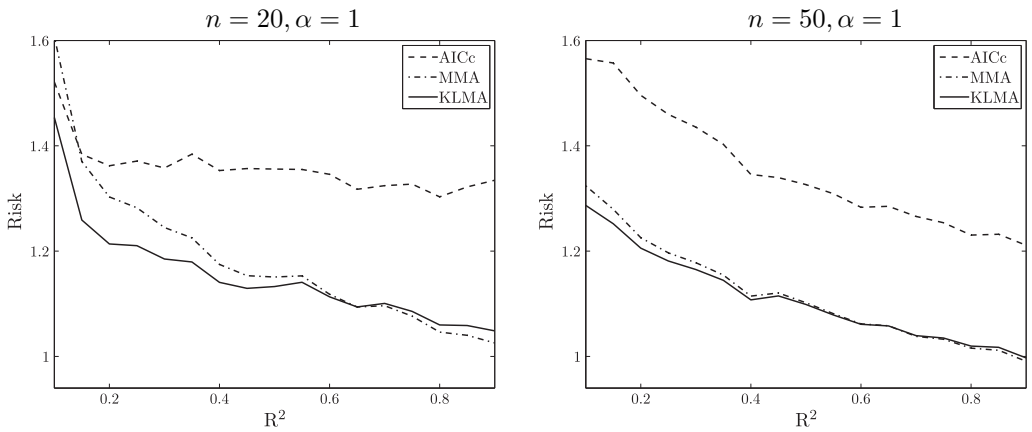

Figure 2. Results for Example 2: risk comparisons under L_μ as a function of R^2.

independent, $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$, $R^2 = c^2/(1+c^2) \in [0.1, 0.9]$, controlled by c, and α is set to 0.5, 1.0, and 1.5. Like Hansen (2007), we considered $S = [3n^{1/3}]$ nested approximating models with the s^{th} model comprising the first s regressors, where $[3n^{1/3}]$ returns the nearest integer from $3n^{1/3}$. As in Example 1, we focused on the small sample cases, with $n = 20$ and 50. Following Hansen (2007), our evaluation was based on the predictive loss function L_μ with 1,000 replications.

The simulation results with $\alpha = 1$ are depicted in Figure 2 and all simulation results are shown in Section S9 of the Supplementary Material. It is seen that the MMA estimator typically yields better estimates than the model selection estimator, which is in accordance with what was observed by Hansen (2007). The KLMA estimator is found to be superior to the MMA estimator in a large region of the parameter space, and this superiority is most marked when R^2 is small and α is large. This performance is particularly encouraging in view of the fact that this experiment is performed under the setting of Hansen (2007), where it has been shown that the MMA estimator performs better than many commonly used model selection and averaging methods. When R^2 is large, MMA can be slightly better than KLMA. When n increases, they perform more similarly.

4.2. Heteroscedastic error setting

We conducted simulation experiments with heteroscedastic errors to compare the mKLMA_1 and mKLMA_2 estimators with the JMA estimator in Hansen and Racine (2012). The weight vector of the JMA estimator was obtained by minimizing a jackknife criterion.

Example 3. This example is based on the same setting as in Example 1 except that n varied in $\{20, 50, 150, 400\}$, and $e \sim \text{Normal}[0, \text{diag}\{\exp(\eta X_{2,1}), \ldots,$

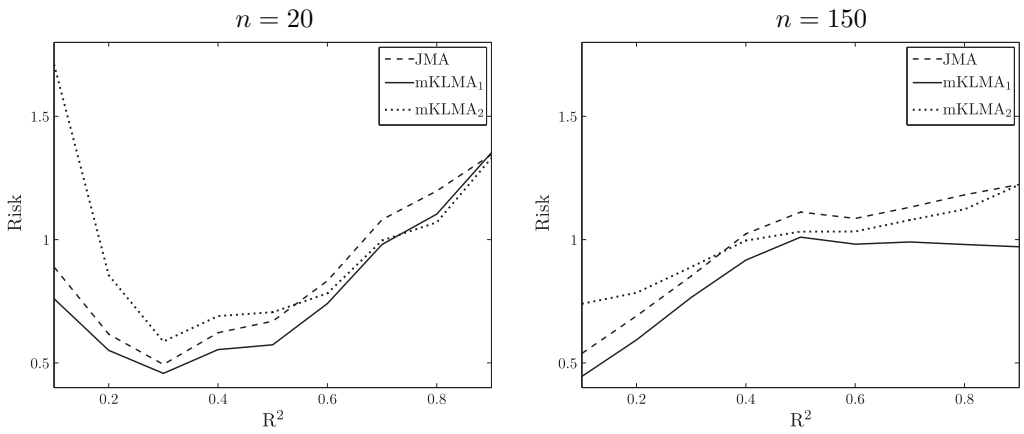

Figure 3. Results for Example 3: risk comparisons under L_μ as a function of R^2.

$\exp(\eta X_{2,n})\}]$, where $X_{2,i}$ is the i^{th} element of X_2 and $\eta > 0$. We changed the value of η such that $R^2 = \text{Var}(\mu_i)/\text{Var}(y_i) \approx \text{Var}(\mu_i)/[\text{Var}(\mu_i) + E\{\exp(\eta X_{2,i})\}] = 14/\{14 + \exp(\eta^2/2)\}$ varied in the range $[0.1, 0.9]$.

The risk comparison results of mKLMA$_1$, mKLMA$_2$, and JMA estimators under L_μ loss are presented in Figure 3 with $n = 20$ and 150 (the results with $n = 50$ and 400 are shown in Figure S.3 of the Supplementary Material). It is clear that mKLMA$_1$ generally leads to the lowest risk. The mKLMA$_2$ and JMA methods perform comparably; the latter has been shown to have advantages over the MMA estimator and other estimators selected by AIC, BIC, and cross-validation (Hansen and Racine (2012)). When R^2 is small, JMA produces a lower risk than mKLMA$_2$, while mKLMA$_2$ is superior to JMA when R^2 is large. The risk comparison under L_β loss is presented in Figure S.4 of the Supplementary Material. As in Example 1, the patterns under L_μ and L_β are almost the same.

We also evaluated estimators in terms of risk under the loss function $L_{\text{hetero},\mu} = (\widehat{\mu} - \mu)\Omega^{-1}(\widehat{\mu} - \mu)$. Figure 4 shows risk comparison results with $n = 20$ and 150 (other results are shown in Figure S.5 of the Supplementary Material), from which, we see that mKLMA$_2$ and JMA are still comparable, and that mKLMA$_1$ performs much better.

In Sections S11-S13 of the Supplementary Material, for a robustness check, we provide some more simulation examples. It is seen that our method is still superior to the other methods when the errors are not normally distributed or the coefficients depend on the sample size.

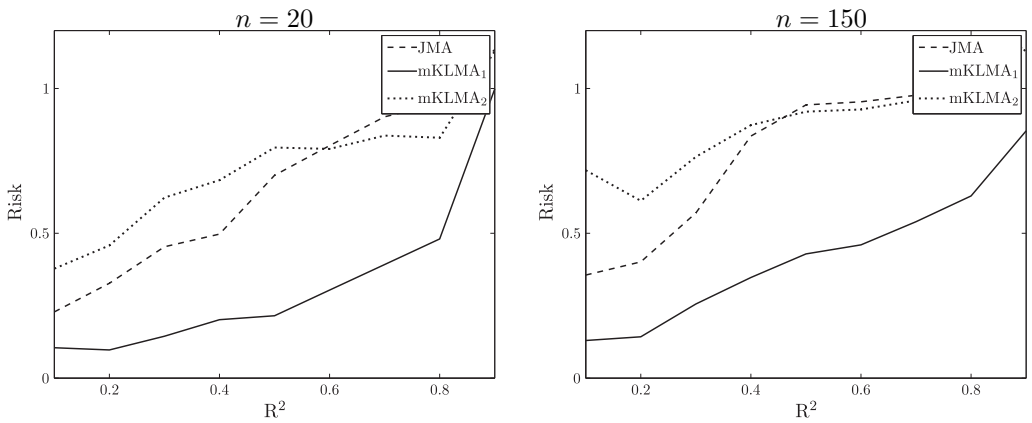

Figure 4. Results for Example 3: risk comparisons under $L_{\text{hetero},\mu}$ as a function of R^2.

5. Empirical Example

We applied our methods to a data set from the Hong Kong residential property market. The data set consists of 560 transactions of the housing estate 'South Horizon' located in the South of Hong Kong, recorded by Centaline Property Agency Ltd. from January 2004 to October 2007. The model from Magnus, Wan and Zhang (2011) is adopted to analyze this data set:

$$\begin{aligned} LPRICE_t = {} & \beta_1 + \beta_2 LAREA_t + \beta_3 LFLOOR_t + \beta_4 GARV_t + \beta_5 INDV_t \\ & + \beta_6 SEAVF_t + \beta_7 SEAVS_t + \beta_8 SEAVM_t + \beta_9 MONV_t \\ & + \beta_{10} STRI_t + \beta_{11} STRN_t + \beta_{12} UNLUCK_t + e_t \end{aligned} \tag{5.1}$$

for $t = 1, \ldots, 560$, where $LPRICE$ is the natural logarithm of the sales price per square foot, and the twelve regressors, including the constant term, are shown in Table 1. As in Magnus, Wan and Zhang (2011), we treated the first six variables as focus regressors and the other six variables as auxiliary regressors, and so we combine $2^6 = 64$ models.

We used indices of the six auxiliary regressors to indicate these candidate models. For example, $(7, 8)$ indicates the model including $SEAVS$ and $SEAVM$. We examined the predictive power of the six model selection and averaging methods used in the simulation study: AICc, MMA, KLMA, JMA, mKLMA_1, and mKLMA_2, the last three of which are developed for the heteroscedastic setting. Magnus, Wan and Zhang (2011) has found that the heteroscedasticity structure of this data set is

$$\Omega = \text{diag}\{\exp(\eta STRN_1), \ldots, \exp(\eta STRN_n)\},$$

so we also used this structure when implementing mKLMA_1.

Table 1. Regressors in application. See Magnus, Wan and Zhang (2011) for a detailed description of these variables.

Index	Regressor	Explanation
1	*INTER.*	Constant term
2	*LAREA*	Size of dwelling in square feet (natural logarithm)
3	*LFLOOR*	Floor level of dwelling (natural logarithm)
4	*GARV*	1 if garden view; 0 otherwise
5	*INDV*	1 if industry view; 0 otherwise
6	*SEAVF*	1 if full sea view; 0 otherwise
7	*SEAVS*	1 if semi sea view; 0 otherwise
8	*SEAVM*	1 if minor sea view; 0 otherwise
9	*MONV*	1 if mountain view; 0 otherwise
10	*STRI*	1 if internal street view; 0 otherwise
11	*STRN*	1 if no street view; 0 otherwise
12	*UNLUCK*	1 if located on floors 4, 14, 24, 34 or in block 4; 0 otherwise.

Table 2. Weights estimated by model averaging methods.

Model	MMA	KLMA	JMA	mKLMA_2	mKLMA_1
(7)	0.06	0.06	0.01	0.18	0.52*
(8)	0.00	0.00	0.00	0.00	0.14
(7,8)	0.00	0.00	0.16	0.00	0.08
(7, 10)	0.22	0.22	0.16	0.16	0.00
(8, 9)	0.11	0.11	0.15	0.02	0.00
(8, 10)	0.00	0.00	0.00	0.00	0.16
(7, 8, 12)	0.21	0.20	0.08	0.31*	0.00
(7, 10, 12)	0.09	0.10	0.00	0.04	0.00
(8, 10, 12)	0.25*	0.25*	0.18	0.27	0.11
(7, 10, 11, 12)	0.06	0.06	0.25*	0.01	0.00

Table 2 shows weights for all model averaging methods. We list only the models whose largest weights for all model averaging methods are not smaller than 0.01. In each column, the largest weight is indicated by an asterisk. It is seen that MMA and KLMA perform very closely and both put the largest weights on model $(8, 10, 12)$. JMA, mKLMA_2, and mKLMA_1 put the largest weights on models $(7, 10, 11, 12)$, $(7, 8, 12)$ and (7), respectively. The model selected by AICc is $(7, 8, 10, 12)$.

In many applications, it is often the case that a prediction may be sensitive to the sample that is used to estimate the forecasting model. Too early observations may not be useful or even lead to worse results in prediction, so we used a moving window of samples for estimation. We let $n = 50$ and 400. For each n, we did $560 - n$ one-step-ahead predictions.

To make comparison results easily detected, in each prediction, we subtracted minimum squared prediction error (SPE) of the six methods, from all SPEs.

Table 3. MSPEDs by model averaging and selection methods and their standard errors in forecasting Hong Kong estate price ($\times 10^{-3}$).

n		AICc	MMA	KLMA	JMA	mKLMA$_2$	mKLMA$_1$
50	MSPED	1.522	1.175	1.164	1.276	1.309	1.081
	s.e.	0.152	0.092	0.090	0.115	0.156	0.092
400	MSPED	0.771	0.690	0.690	0.684	0.682	0.654
	s.e.	0.099	0.063	0.063	0.065	0.081	0.083

The corresponding values are called SPE distances. Table 3 displays mean SPE distances (MSPEDs) and their standard errors based on $560 - n$ predictions. Again, it is seen that KLMA performs better than MMA for relatively small sample size situation and they have very similar performance for the large sample sizes. We also find that mKLMA$_1$ performs best, and JMA and mKLMA$_2$ are comparable.

6. Concluding Remarks

We have developed a novel weight choice criterion based on the KL distance. Like the well-known MMA estimator, the resulting KLMA estimator is asymptotically optimal. More importantly, for finite sample situation, the KLMA estimator has been observed to be generally superior to the MMA estimator. We have further extended the KLMA estimator to the setting with heteroscedasticity and proved the corresponding asymptotic optimality. The simulation study and application have shown the promise of the proposed model average estimators.

For the purpose of statistical inference, it is necessary to obtain the limiting distribution of a model average estimator. Under the commonly used models with the local misspecification assumption, the limiting distribution theory of model average estimator using weights with an explicit form has been established in the literature such as Hjort and Claeskens (2003). Deriving the limiting distributions of our model average estimators, whose weight vectors have no explicit expressions, warrants further investigation.

Lastly, we remark that unbiasedness built in Theorems 1 and 3 are based on the normality assumption of e. Although a robustness check in the simulation study shows that our method still outperforms its competitors when e follows a uniform or Chi-squared distribution, we cannot conclude that our approach can be generally applied to other error distribution cases. Developing specific weight choice criteria for other distributions is an interesting open question for future studies.

Acknowledgements

The authors are grateful to Co-Editor Naisyin Wang, an associate editor and two referees for their constructive comments. Zhang's research was par-

tially supported by National Natural Science Foundation of China (Grant nos. 71101141 and 11471324). Zou's research was partially supported by National Natural Science Foundation of China (Grant nos. 11331011 and 11271355) and a grant from the Hundred Talents Program of the Chinese Academy of Sciences. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030). This work occurred when the first author visited Texas A&M University.

Supplementary Material SuppMat.pdf contains the technical proofs and provides figures for the outcomes of the numerical studies.

Appendix: Assumptions

Let $\lambda_{\max}(A)$ denote the maximum singular value for a matrix A, $R_n(w) = E\{L_n(w)\}$, $\xi_n = \inf_{w\in\mathcal{W}} R_n(w)$, w_s^0 be an $S\times 1$ vector in which the s^{th} element is one and the others are zeros, and $\widehat{T}$ be a matrix such that $\partial\widehat{\sigma}^2/\partial y = \widehat{T}y$.

Assumption A.1. For a constant κ_1 and some fixed integer $1 \le G < \infty$, $E(e_i^{4G}) \le \kappa_1 < \infty,\ i = 1,\ldots,n$, and $S\xi_n^{-2G}\sum_{s=1}^{S} R_n^G(w_s^0) = o(1)$.

Assumption A.2. $\max_{s\in\{1,\ldots,S\}} \lambda_{\max}(P_{(s)}) = O(1)$.

Assumption A.3. $\|\mu\|^2 n^{-1} = O(1)$.

Assumption A.4. $\sup_{w\in\mathcal{W}}[|(\widehat{\sigma}^2 - \sigma^2)\text{trace}\{P(w)\}|R_n^{-1}(w)] = o_p(1)$.

Assumption A.5. $n\lambda_{\max}(\widehat{T})\xi_n^{-1} = o_p(1)$.

Assumptions (A.1)−(A.3) are commonly used in such literature on model selection and model averaging as Li (1987), Andrews (1991), Shao (1997), Hansen (2007), and Wan, Zhang, and Zou (2010). The normality of e required in Theorem 1 is not necessary for asymptotic optimality. In Section S7 of the Supplementary Material, we present a discussion on Assumption (A.1) and its relationship with the normality of e.

Assumption (A.4) restricts the estimator $\widehat{\sigma}^2$. In Hansen (2007) and Wan, Zhang, and Zou (2010), the model with the largest rank of regressor matrix, denoted as r, is used to estimate σ^2. In this case, Assumption (A.4) is implied by Assumptions (A.1)−(A.3) and $r^2n^{-1} = O(1)$. See the proof of Theorem 2 in Wan, Zhang, and Zou (2010) for the derivation.

Assumption (A.5) places a constraint on the robustness of the estimator $\widehat{\sigma}^2$. Under any candidate model s, a natural estimator of σ^2 is $\widehat{\sigma}^2 = \|y - \widehat{\mu}_{(s)}\|^2/n = y^{\mathrm{T}}(I_n - P_{(s)})^{\mathrm{T}}(I_n - P_{(s)})y/n$, and then Assumption (A.5) is obviously implied by Assumptions (A.1)−(A.2).

Let $R_{\text{hetero},n}(w) = E\{L_{\text{hetero},n}(w)\}$, $\xi_{\text{hetero},n} = \inf_{w\in\mathcal{W}} R_{\text{hetero},n}(w)$, $\widehat{A}$ be a matrix such that $\widehat{a} = \widehat{A}y$, and $\widetilde{P}(w) = \Omega^{-1/2}P(w)\Omega^{1/2}$.

Assumption B.1. For a constant κ_2 and some fixed integer $1 \leq G_1 < \infty$, $E(e_i^{4G_1}) \leq \kappa_2 < \infty,\ i = 1, \ldots, n$, and $S\xi_{\text{hetero},n}^{-2G_1} \sum_{s=1}^{S} R_{\text{hetero},n}^{G_1}(w_s^0) = o(1)$.

Assumption B.2. There exist two constants c_1 and c_2 such that $0 < c_1 \leq \min_{i\in\{1,\ldots,n\}} \Omega_{ii} \leq \max_{i\in\{1,\ldots,n\}} \Omega_{ii} \leq c_2 < \infty$.

Assumption B.3. $(\max_{i\in\{1,\ldots,n\}} |\widehat{\Omega}_{ii} - \Omega_{ii}|)^2 n\xi_{\text{hetero},n}^{-1} = o_p(1)$.

Assumption B.4. $\max_{i\in\{1,\ldots,n\}} |\widehat{\Omega}_{ii} - \Omega_{ii}| \sup_{w\in\mathcal{W}}[R_{\text{hetero},n}^{-1}(w)\text{trace}\{\widetilde{P}(w)\widetilde{P}^{\mathrm{T}}(w)\}] = o_p(1)$.

Assumption B.5. $n\lambda_{\max}(\widehat{A})\xi_{\text{hetero},n}^{-1} = o_p(1)$.

Assumptions (B.1) and (B.5) are similar to Assumptions (A.1) and (A.5), respectively. Assumptions (B.3)−(B.4) restrict the estimator $\widehat{\Omega}$. When the structure of Ω is known and it is related to a parameter vector η, $\Omega = \Omega(\eta)$, we generally have $\|\widehat{\eta} - \eta\| = O_p(n^{-1/2})$ and $\max_{i\in\{1,\ldots,n\}} |\widehat{\Omega}_{ii} - \Omega_{ii}| = O_p(n^{-1/2})$ under some regularity conditions and, in this case, Assumptions (B.3)−(B.4) are implied by Assumption (B.1) and formula (S5.4) in the Supplementary Material, respectively.

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255-265.

Andrews, D. W. K. (1991). Asymptotic optimality of generalized c_l, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *J. Econometrics* **47**, 359-377.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603-618.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statist. Probab. Lett.* **42**, 333-343.

Claeskens, G., Croux, C. and van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62**, 972-979.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica* **75**, 1175-1189.

Hansen, B. E. and Racine, J. (2012). Jacknife model averaging. *J. Econometrics* **167**, 38-46.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**, 879-899.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (2002). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. Ser. B* **60**, 271-293.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875-890.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52**, 3396-3410.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958-975.

Liang, H., Zou, G., Wan, A. T. K. and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *J. Amer. Statist. Assoc.* **106**, 1053-1066.

Liu, Q. and Okui, R. (2013). Heteroskedasticity-robust C_p model averaging. *The Econometrics J.* **16**, 463-472.

Longford, N. T. (2005). Editorial: Model selection and efficiency-is 'which model?' the right question? *J. Roy. Statist. Soc. Ser. A* **168**, 469-472.

Magnus, J. R., Wan, A. T. K. and Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Comput. Statist. Data Anal.* **55**, 1331–1341.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

Miller, A. J. (2002). *Subset Selection in Regression.* 2nd edition. Chapman and Hall, London.

Rigollet, R. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40**, 639-665.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221-242.

Shi, P. and Tsai, C. L. (2004). A joint regression variable and autoregressive order selection criterion. *J. Time Series Anal.* **25**, 923-941.

Wan, A. T. K., Zhang, X. and Zou, G. (2010). Least squares model averaging by Mallows criterion. *J. Econometrics* **156**, 277-283.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96**, 574-588.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.* **39**, 174-200.

Zhang, X., Lu, Z. and Zou, G. (2013). Adaptively combined forecasting for discrete response time series. *J. Econometrics* **176**, 80-91.

Zhang, X., Wan, A. T. K. and Zhou, S. Z. (2012). Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *J. Bus. Econom. Statist.* **30**, 132-142.

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

E-mail: xinyu@amss.ac.cn

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and School of Mathematical Science, Capital Normal University, Beijing 100037, China.

E-mail: ghzou@amss.ac.cn

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.

E-mail: carroll@stat.tamu.edu

(Received October 2013; accepted October 2014)