American Journal of Applied Mathematics and Statistics, 2014, Vol. 2, No. 3, 150-156
Available online at http://pubs.sciepub.com/ajams/2/3/9
© Science and Education Publishing
DOI:10.12691/ajams-2-3-9

On Optimal Weighting Scheme in Model Averaging

Georges Nguefack-Tsague[*]

Department of Public Health, University of Yaounde I, Biostatistics Unit, Yaoundé, Cameroon
*Corresponding author: nguefacktsague@yahoo.fr

Received April 14, 2014; Revised May 07, 2014; Accepted May 13, 2014

Abstract Model averaging is an alternative to model selection and involves assigning weights to different models. A natural question that arises is whether there is an optimal weighting scheme. Various authors have shown their existence in others methodological frameworks. This paper investigates the derivation of optimal weights for model averaging using square error loss. It is shown that though these weights may exist in theory and depend on model parameters; once estimated they are no longer optimal. It is demonstrated using an example of linear regression that model averaging estimators with these estimated weights are unlikely to outperform post-model selection and others model averaging estimators. We provide a theoretical justification for this phenomenon.

Keywords: *model averaging, model selection, optimal weight, square error loss, model uncertainty*

Cite This Article: Georges Nguefack-Tsague, “On Optimal Weighting Scheme in Model Averaging.” *American Journal of Applied Mathematics and Statistics*, vol. 2, no. 3 (2014): 150-156. doi: 10.12691/ajams-2-3-9.

1. Introduction

In most statistical modeling applications, there are several models that are a priori plausible. It is quite common nowadays to apply some model selection procedure to select a single one. Overviews, explanations, discussion and examples of such methods may be found in the books by Linhart and Zucchini [1], McQuarrie and Tsai [2], Burnham and Anderson [3] and Claeskens and Hjort [4].

An alternative to select a single model for estimation purposes is to give weights to all plausible models and to work with the resulting weighted estimator. This leads to the class of model averaging estimators. Once decided upon the weights (these can be the result of a model selection criterion such as Akaike’s information criterion (AIC), or arising from Bayesian motivations), the problem is not so much with the construction of the estimator, as with its properties.

Since model selection corresponds to the special case of assigning weight one to the selected model and weight zero to all other considered models, the question is equally relevant for estimators obtained after model selection. We refer to these estimators as *post-model selection estimators* (PMSE). The fact that selection was data-based is often ignored in the subsequent analysis and leads to invalid inferences. Literature on this topic includes but is not limited to Bancroft [5] for pre-test estimators, Breiman [6], Hjorth [7], Chatfield [8], Draper [9], Buckland et al. [10], Zucchini [11], Candolo et al. [12], Hjort and Claeskens [13], Efron [14], Leeb and Pötscher [15], Longford [17], Claeskens and Hjort [4], Schomaker et al. [18], Zucchini et al. [19], Liu and Yang [20], Nguefack-Tsague and Zucchini [21], Nguefack-Tsague et al. [26], and Nguefack-Tsague [22,23,24,25]. Bayesian model averaging can be found in Hoeting et al. [27] and Wasserman [28]. Wang et al. [29] provide a review of frequentist model averaging estimators.

Many optimal weighting schemes have evolved recently for model averaging. Hansen [30] discusses the model averaging in least squares estimation and proposes a method that selects the weights by minimizing Mallows’ criterion. Furthermore, Hansen [31] suggests to use Mallows’ model averaging method to do forecast and shows that the Mallows’ criterion is an asymptotically unbiased estimator of both the in-sample mean squared error and the out-of-sample one-step-ahead mean squared forecast error. Hansen [32] studies least squares estimation of an autoregressive model with a root close to unity by proposing two measures to evaluate the efficiency of the estimators: the asymptotic mean squared error and forecast expected squared error. Numerical comparison of Mallows’ model averaging method with many other methods shows that Mallows’ model averaging estimator often has smaller risk. Hansen [33] applies the same idea for model averaging with autoregressions with a near unit root. Since Hansen [30] assumes that the models are nested and the weights are discrete, Wan et al. [34] relaxed these two assumptions to obtain other versions of model averaging by minimizing Mallows criterion. Their proofs are based on Li [35]. Liang et al. [36] develop a model weighting mechanism that involves minimizing the trace of an unbiased estimator of the model average estimator’s MSE. Hansen and Racine [37] propose to select the weights of least squares model averaging estimator by minimizing a deleted-1 cross-validation criterion (the jackknife model averaging (JMA)). The solutions of the above methods are obtained by quadratic programming. Zhang et al. [38] propose a model averaging scheme for linear mixed-effects models and prove their method to be asymptotically optimal under some regularity conditions.

Various above optimal weights do not use the most common straightforward square error loss function. The intention here is to use this loss to derive optimal weights. Unlike the others methods, since the risk function of the model averaging obtained depends on model parameters, comparisons should be made along the parameter space with others post-model selection and model averaging estimators. The important question to ask is whether within this framework optimal weights are really optimal? In particular in terms of risk function, is it preferable to perform model selection or model averaging? Others existing methods (not using the risk function in the parameter space) advocate model averaging over model selection. The following Section describes conceptually model averaging and PMSEs while Section 3 describes the concept of optimal weight. Section 4 illustrates the point with a simple linear regression model with derivation of optimal weights in this case, and Section 5 provides a theoretical justification for the fact that optimality does no longer hold when paramaters are estimated. The article ends with concluding remarks.

2. Model Averaging and Post-Model Selection Estimators

Let $\mathcal{M} = (M_1,\ldots,M_K)$ be a set of K plausible models to estimate μ, the quantity of interest. Denote by $\hat{\mu}_k$ the estimator of μ obtained from using model M_k. Model averaging involves finding non-negative weights, $w = (w_1,\ldots,w_K)^t$ that sum to one, and then estimating μ by

$$\hat{\mu} = \sum_{k=1}^{K} w_k \hat{\mu}_k. \qquad (1)$$

Clearly, by taking only one of the weights equal to one, and the other weights all zero, the model averaged estimator reduces to the estimator in a single model. This important sub-class of model averaging estimators is arrived at by model selection. There the weight of model M_k is set to one if and only if the model selection method selects model M_k, and the weight is zero otherwise.

Some classical model averaging weights involve penalized likelihood values. Let I_k denote an information criterion of the form

$$I_k = -2\log L_k + s_k, \qquad (2)$$

with s_k a penalty for model M_k and L_k the maximized likelihood value at model M_k. Buckland et al. [10] define Akaike-type of weights:

$$w_k = \frac{\exp(-s_k/2)L_k}{\Sigma_{l=1}^K \exp(-s_l/2)L_l} = \frac{\exp(-I_k/2)}{\Sigma_{l=1}^K \exp(-I_l/2)}. \qquad (3)$$

In particular, if we use the Akaike information criterion (AIC, Akaike [39]) with $s_k = 2q_k$, two times the number of parameters of model M_k, (3) simplifies to

$$w_{\text{aic},k} = \frac{\exp(-AIC_k/2)}{\Sigma_{l=1}^K \exp(-AIC_l/2)}. \qquad (4)$$

Extensive application of the Akaike weights can be found in Burnham and Anderson [3]. Candolo et al. [12] apply these weights to the linear regression example.

When the Bayesian information criterion (BIC, Schwarz [40]) is used, with $s_k = \log(n)q_k$ and n the sample size, the resulting weights are

$$w_{\text{bic},k} = \frac{\exp(-BIC_k/2)}{\Sigma_{l=1}^K \exp(-BIC_l/2)}. \qquad (5)$$

With equal prior weights to each of the models M_k, this may be interpreted as an approximation to the posterior probability of model M_k given the data.

In the context of regression and classification, LeBlanc and Tibshirani [41] propose to use a non-penalized likelihood value, resulting in $w_k = L_k / \Sigma_{l=1}^K L_l$. Hjort and Claeskens [13] use the smooth focused information criterion (FIC) and other model averaging schemes to study this type of model averaged, or compromise estimators, together with their limiting distributions and risk properties. Model averaging in semiparametric regression with AIC or BIC type weights is studied by Claeskens and Carroll [42]. For details discussion on model averaging and its applications, see e.g. [43-46]. Zou and Yang [47] apply model averaging for time series while, Yuan and Yang [48] explain under which conditions should one apply model averaging. Shan and Yang [49] apply model averaging for quantile estimation while Liu and Yang apply it in longitudinal data analysis. If one is able to find closed form expressions of the model selection probabilities $p_k = E(I(M_k \textit{ is selected}))$ (Note that the expectation of a Bernoulli variable is the probability of success, say π) for each model M_k then an obvious weighting scheme is to use an estimator of these probabilities.

There is also a special case of model averaging estimator where only zero/one weights apply, post-model selection estimator (PMSE, [15,16]). We use a model selection criterion to decide on a selected model $M_{\hat{k}}$, and use this model to estimate the parameter of interest by $\hat{\mu}_{\hat{k}}$, that is, the estimator of μ in the selected model. Using the notation introduced above, we may write PMSE $\hat{\mu}_{\hat{k}}$ as

$$M_{\hat{k}} = \sum_{k=1}^{K} I(\textit{model } k \textit{ is selected})M_k,$$

$$\hat{\mu}_{\hat{k}} = \sum_{k=1}^{K} I(\textit{model } k \textit{ is selected})\hat{\mu}_k.$$

It is important to stress that since the model selection method depends on the actual data, the selected model $M_{\hat{k}}$ is random as well. This implies that even when the same set of models $\mathcal{M}$ and the same selection criterion are used, different samples can lead to different models ($M_{\hat{k}}$) being selected. The selected model depends also on the selection procedure and the set of models $\mathcal{M}$.

3. Optimal Model Averaging Estimator

A question that arises is whether one can select the weights so as to optimize the performance of this averaged estimator, in terms of some specified measure, say a loss function L. Finding the optimal weights involves solving the following optimization problem over $\pi = (\pi_1, \ldots, \pi_K)$:

$$\min_{\pi} E_{true} L(\sum_{k=1}^{K} \pi_k \hat{\mu}_k, \quad \mu), \qquad \textit{such that } \pi_k \geq 0, \forall k \textit{ and } \sum_{k=1}^{K} \pi_k = 1, \tag{6}$$

where the expectation is taken with respect to the true model M_{true}, which may, or may not, be in the set of competing models, $\mathcal{M}$. If the true model is known then, at least in theory, it is possible to find the optimal weights, if they exist. However, the optimal weights, obtained by minimizing (6), depend on the parameters of M_{true}, which are unknown and thus have to be estimated. Using estimates for the π_k may lead to weights that are no longer optimal.

Since the optimal weight (if exists) depends on the parameters, Hjort and Claeskens [13] suggest to minimize the estimated risk. A closed-form solution for optimal weights in general is unlikely to exist when the models are complexes, but the intention here is to investigate the lung run properties of model averaging estimator when a close-form solution exists.

4. Ilustration with Simple Linear Regression

4.1. Problem Set-Up

Consider a simple linear regression model in which x is a covariate and Y is the response variable, given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n, \tag{7}$$

where the $\varepsilon_i : N(0, \sigma^2)$, σ known (for simplicity).

The OLS estimators are given by

$$\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2};$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and Cov } (\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}.$$

For simplicity of computations, suppose $\bar{x} = 0$, without loss of generality, since linear regression model (7) can be parametrised as

$$Y_i = \lambda_0 + \lambda_1 (x_i - \bar{x}) + \varepsilon_i, i = 1, \ldots, n, \tag{8}$$

where $\lambda_0 = \beta_0 + \beta_1 \bar{x}$ and $\lambda_1 = \beta_1$.

Thus, under $\bar{x} = 0$, $\hat{\beta}_0 = \bar{y}$ and $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$, also $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed, therefore $\hat{\beta}_0$ and $\hat{\beta}_1$ are independant.

Let x_+ be a future value of the covariate. The aim is to estimate the mean $\mu = E(Y \mid x_+)$.

Consider two models

$M_0 : \mu = \beta_0$ and

$M_1 : \mu = \beta_0 + \beta_1 x_+$.

$v_0 = \frac{\sigma^2}{n}$; v_1 =Var $(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma_{i=1}^{n} x_i^2}$. Let $Z_0 = \frac{\hat{\beta}_0 - \beta_0}{v_0^{1/2}}$, then $Z_0 : N(0,1)$ and $\hat{\beta}_0 = v_1^{1/2}(Z_0 + b_0)$, where $b_0 = \frac{\beta_0}{v_0^{1/2}}$ (standardized intercept). Let $Z_1 = \frac{\hat{\beta}_1 - \beta_1}{v_1^{1/2}}$, then $Z_1 : N(0,1)$, $\hat{\beta}_1 = v_1^{1/2}(Z_1 + b_1)$, where $b_1 = \frac{\beta_1}{v_1^{1/2}}$ (standardized slope).

4.2. Post-model Selection Estimators

Consider a selection criterion of the form (2) where $s_k = hq_k$, $q_2 = 2$, $q_1 = 1$, $\{h \in R : h > 0\}$. M_1 is chosen if $| Z_1 + b_1 | \geq h^{1/2}$.

PMSE estimator can be written as

$$\begin{aligned} \hat{\mu}_{\hat{k},h} &= \hat{\beta}_0 I_0 (| Z_1 + b_1 | < h^{1/2}) \\ &+ (\hat{\beta}_0 + \hat{\beta}_1 x_+) I_1 (| Z_1 + b_1 | \geq h^{1/2}), \end{aligned} \tag{9}$$

where I_0 and I_1 are, respectively, indicator functions under M_0 and M_1 with $I_0 + I_1 = 1$.

It follows that

$$\begin{aligned} \hat{\mu}_{\hat{k},h} &= \hat{\beta}_0 + \hat{\beta}_1 x_+ I_1 (| Z_1 + b_1 | \geq h^{1/2}) \\ &= \hat{\beta}_0 + x_+ \tilde{\beta}_1 h, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \tilde{\beta}_1 h &= \hat{\beta}_1 I_1 (| Z_1 + b_1 | \geq h^{1/2}) \\ &= v_1^{1/2} (Z_1 + b_1) I_1 (| Z_1 + b_1 | \geq h^{1/2}) = v_1^{1/2} A_h \end{aligned} \tag{11}$$

with $A_h = (Z_1 + b_1) I_1 (| Z_1 + b_1 | \geq h^{1/2})$.

The expression (11) is equivalent to

$$\tilde{\beta}_{1h} = \begin{cases} 0 & | Z_1 + b_1 | < h^{1/2} \\ \hat{\beta}_1 & | Z_1 + b_1 | \geq h^{1/2}. \end{cases} \tag{12}$$

We used simulated data to investigate the properties of different PMSEs, namely 10^5 samples of size $n = 20$, with $\sigma^2 = 1$, $x_i = \frac{2i}{n+1} - 1$, $i = 1, 2, \ldots, n$, $\beta_0 = 0$, and $x_+ = 0.5$. The reported results are not sensitive to the choice of these selected values and data; in particular, increasing the sample size, has minimal impact on the results. All expectations here were taken with respect to the full model, M_1. As the risk functions are symmetric around zero we only display the graphs for $b_1 > 0$. All computations are performed with the software R [50].

For some classical selection procedures, values of h are given by

$$h = \begin{cases} z^2_{1-\frac{\alpha}{2}} & \textit{for Hypothesis Testing} \\ 2 & \textit{for AIC} \\ \log(n) & \textit{for BIC} \\ \log(\log(n)) & \textit{for HQ.} \end{cases}$$

$$\begin{aligned} \hat{\mu}_{\hat{k},h} &= v_0^{1/2}(Z_0 + b_0) \\ &\quad + x_+ v_1^{1/2}(Z_1 + b_1) I_1(|Z_1 + b_1| \geq h^{1/2}). \end{aligned} \tag{13}$$

Nguefack-Tsague and Zucchini [21] show that for Mallows’Cp [51], PMSE is given by

$$\begin{aligned} \hat{\mu}_{\hat{k}\ ,Cp} &= v_0^{1/2}(Z_0 + b_0) \\ &\quad + x_+ v_1^{1/2}(Z_1 + b_1) I_1(F(1, n-2, b_1^2) > 2). \end{aligned} \tag{14}$$

with F as the non central Fisher distribution, namely $F(1, n-2, b_1^2)$.

4.3. Model Averaging

For any weighting scheme w_0 and w_1 for model M_0 and model M_1 , $w_0 + w_1 = 1$, the model averaging estimator is

$$\begin{aligned} \hat{\mu}_{MA} &= w_0 \hat{\mu}_0 + w_1 \hat{\mu}_1 \\ &= w_0 \hat{\beta}_0 + w_1(\hat{\beta}_0 + x_+ \hat{\beta}_1) = \hat{\beta}_0 + (1-w)\hat{\beta}_1 x_+, \end{aligned} \tag{15}$$

where $w_0 = w$.

Using formulae of Akaike weights and likelihood weights given in (4), (3) and (5), we have $AIC_1 - AIC_0 = 2(\log L_0 - \log L_1) + 2$ and $w_{aic,1} = \dfrac{e^{(AIC_0 - AIC_1)}}{1 + e^{(AIC_0 - AIC_1)}}$. Akaike weights are then given by

$$w_{\text{aic},1} = \frac{e^{\frac{1}{2}(Z_1+b_1)^2 - 1}}{1 + e^{\frac{1}{2}(Z_1+b_1)^2 - 1}}, \tag{16}$$

and BIC weights by

$$w_{\text{bic},1} = \frac{e^{\frac{1}{2}(Z_1+b_1)^2 - \frac{\log(n)}{2}}}{1 + e^{\frac{1}{2}(Z_1+b_1)^2 - \frac{\log(n)}{2}}}. \tag{17}$$

More generally here, weights using penalized information criterion of the from (2) where $s_k = hq_k$ are given by

$$w_{\text{h},1} = \frac{e^{\frac{1}{2}(Z_1+b_1)^2 - \frac{h}{2}}}{1 + e^{\frac{1}{2}(Z_1+b_1)^2 - \frac{h}{2}}}. \tag{18}$$

Thus for non-penalized information criterion, likelihood weights are given by

$$w_{\text{h},1} = \frac{e^{\frac{1}{2}(Z_1+b_1)^2}}{1 + e^{\frac{1}{2}(Z_1+b_1)^2}}. \tag{19}$$

4.4. Optimal Weight for Simple Linear Regression

Consider the loss here to be the square error, therefore the measure for the optimal weight is the mean square error.

Proposition 1. Under Equations (7) and (15) the optimal weighting scheme is $w^* = w_0^* = \dfrac{v_1}{v_1 + \beta_1^2}$ and $w_1^* = 1 - w^* = \dfrac{\beta_1^2}{\beta_1^2 + v_1} = \dfrac{b_1^2}{1 + b_1^2}$.

Proof: $$\begin{aligned} (\hat{\mu}_{MA} - \hat{\mu})^2 &= (\hat{\beta}_0 + (1-w)\hat{\beta}_1 x_+ - \beta_0 - \beta_1 x_+)^2 \\ &= [(\hat{\beta}_0 - \beta_0) + x_+(\hat{\beta}_1 - \beta_1) - w\hat{\beta}_1 x_+]^2 \end{aligned}.$$

$$MSE(\hat{\mu}_{MA}) = E(\hat{\mu}_{MA} - \mu)^2$$

$$= E\begin{bmatrix} (\hat{\beta}_0 - \beta_0) \\ +x_+(\hat{\beta}_1 - \beta_1) \end{bmatrix}^2 + w^2 x_+^2 E(\hat{\beta}_1^2) - 2x_+ wE\begin{bmatrix} (\hat{\beta}_0 - \beta_0) \\ +x_+(\hat{\beta}_1 - \beta_1)\hat{\beta}_1 \end{bmatrix}$$

$$= A + w^2 x_+^2 E(\hat{\beta}_1^2) - 2x_+^2 w(E(\hat{\beta}_1^2) - \beta_1^2)$$

where $A = E[(\hat{\beta}_0 - \beta_0) + x_+(\hat{\beta}_1 - \beta_1)]^2$ is constant i.e does not depend on w .

$$\frac{\partial(MSE(\hat{\mu}_{MA}))}{\partial w} = 2wx_+^2 E(\hat{\beta}_1^2) - 2x_+^2(E(\hat{\beta}_1^2) - \beta_1^2).$$

$$\frac{\partial(MSE(\hat{\mu}_{MA}))}{\partial w} = 0$$

$$\Leftrightarrow w^* = \frac{E(\hat{\beta}_1^2) - \beta_1^2}{E(\hat{\beta}_1^2)} = \frac{V(\hat{\beta}_1)}{V(\hat{\beta}_1) + \beta_1^2} = \frac{v_1}{v_1 + \beta_1^2}.$$

$\dfrac{\partial^2(MSE(\hat{\mu}_{MA}))}{\partial w^2} = 2x_+^2 E(\hat{\beta}_1^2) \geq 0$, therefore w^* is a minimum.

Corollary 1. The model averaging estimator based on estimates of the optimal weights is

$$\begin{aligned} \hat{\mu}^*_{optimal} &= \hat{w}^* \hat{\mu}_0 + (1 - \hat{w}^*)\hat{\mu}_1 \\ &= \hat{w}^* \hat{\beta}_0 + (1 - \hat{w}^*)(\hat{\beta}_0 + x_+ \hat{\beta}_1) \end{aligned}$$

where $\hat{w}^* = \dfrac{1}{1 + (Z_1 + b_1)^2}$.

Proof. From Proposition 1, w^* depends on the unknown parameter β_1 , need to be estimated. $w^* = \dfrac{v_1}{\beta_1^2 + v_1} = \dfrac{1}{1 + \beta_1^2 / v_1}$. Thus $\hat{w}^* = \dfrac{1}{1 + \hat{\beta}_1^2 / v_1} = \dfrac{1}{1 + (Z_1 + b_1)^2}$. Replacing the weights in Equation (15) yields the result.

Figure 1 shows for each PMSE, its risk and the risks of various weighting scheme. It shows that *none* of the weighting schemes (including the optimal weights) is better than any PMSE over the whole range of b_1 . Optimal weight scheme is even worse for larger b_1 .

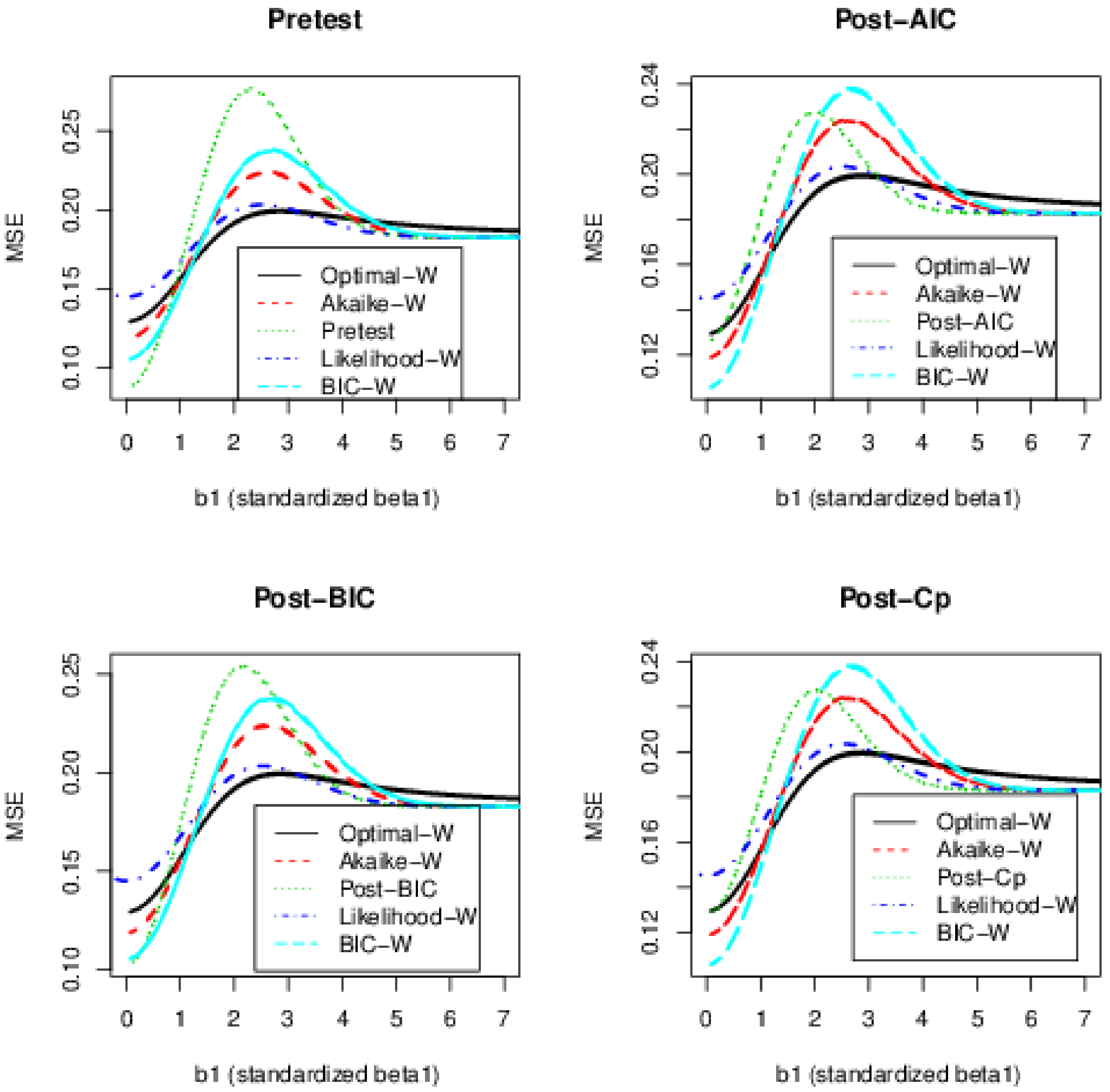

Figure 1. Risks functions for classical PMSEs together with model averaging estimators, optimal weight, Akaike weight, BIC-weight, and likelihood weight; based on 100 000 samples of size $n = 20$, with $\sigma^2 = 1$, $x_i = \frac{2i}{n+1} - 1$, $i = 1,2,\dots,n$, $\beta_0 = 0$, and $x_+ = 0.5$

5. Why Optimal Is Not Optimal?

Consider the regression model

$$Y_i = f(x_i) + \varepsilon_i \quad (i = 1,2,\dots,n), \tag{20}$$

where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the value of a p-dimensional design variable at the i^th^ observation, Y_i is the response, f is the true regression function and the random errors ε_i are assumed to be independent and normally distributed with mean 0 and variance σ^2.

For simplicity, let assume that the model is parametric and can be written in vector form as

$$Y = f_k(x, \theta_k) + \varepsilon, \tag{21}$$

where for each model M_k, the family $\mathcal{F}_{M_k} = \{ f_k(x, \theta_k), \ \theta_k \in \Theta_k \}$ is a linear family of regression functions with θ_k of finite dimension m_k.

Assumption 1 (Yang [48], pp.941). There exist two models M_1 and $M_2 \in M$ such that:

(a) $\mathcal{F}_{M_1} = \{ f_1(x, \theta_1), \ \theta_1 \in \Theta_1 \}$ is a linear subspace of $\mathcal{F}_{M_2} = \{ f_2(x, \theta_2), \ \theta_2 \in \Theta_2 \}$;

(b) there exists a function $\phi(x) \in \mathcal{F}_{M_2}$ orthogonal to $\mathcal{F}_{M_1}$ at the design points, with $n^{-1} \sum_{i=1}^{n} \phi^2(x)$ bounded between two positive constants, at least for large enough n;

(c) there exists a function $f_0 \in \mathcal{F}_{M_1}$ such that f_0 is not in any family $\mathcal{F}_{M_k}$, for $M_k \in M$, that does not contain $\mathcal{F}_{M_1}$.

Consider a model averaging method τ. Let π_k be the resulting data-dependent weight for model M_k satisfying $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. The regression estimator is thus

$\hat{f}(x) = \sum_{k=1}^{K} \pi_k f_k(x, \hat{\theta}_k)$. Let $R(f, \tau, n) = n^{-1} \sum_{i=1}^{n} E(f(x_i) - \hat{f}(x_i))^2$.

Definition 1. A model averaging method τ is said to be consistent in weighting if, when the true model $M_{k^*} \in M$, we have that $\pi_{k^*} \to 1$ as $n \to \infty$.

Theorem 1 (Theorem 2 of Yang [48], pp.943). Under Assumption 1, if any model averaging method τ is consistent in weighting, then we must have

$$n \sup_{f \in \mathcal{F}_{M_2}} R(f, \tau, n) \to \infty. \qquad (22)$$

Theorem 1 clearly explains why within this framework, none of the weight can be expected to dominate all the others in terms of risk function.

6. Concluding Remarks

The aim of this paper has been to show that tough many optimal model averaging schemes have evolved recently, they may fail to exist under square error loss when different estimators are compared in the parameter space using the risk function. We show this by deriving the optimal weighting scheme and demonstrated that these weights are no longer optimal when the parameters are estimated. In particular within this framework model averaging is not preferable to model selection. The example used is very simple but is enough to illustrate the problem.

Acknowledgement

I wish to thank Professor Walter Zucchini for valuable comments, which led to an improvement of this paper.

References

[1] Linhart, H. and Zucchini, W. *Model selection*, John Wiley and Sons, New York, 1986.

[2] McQuarrie, A. D. R. and Tsai, C. L. *Regression and time series model selection*, World Scientific, Singapore, 1998.

[3] Burnham, P. K. and Anderson, D. R. *Model selection and multimodel inference, a practical information-theoretic approach*, Second Edition, Springer-Verlag, New York, 2002.

[4] Claeskens, G. and Hjort, N. L. *Model selection and model averaging*, Cambridge University Press, Cambridge, 2008.

[5] Bancroft, T. A. *On bias in estimation due to the use of preliminary tests of significance*, Annals of Mathematical Statistics 15 1944, 190-204.

[6] Breiman, L. *The little bootstrap and other methods for dimensionality selection in regression: X-Fixed predictor error*, Journal of the American Statistical Association 87 1992, 738-754.

[7] Hjorth, J. *Computer intensive statistical methods:Validation, model selection, and bootstrap*, Chapman and Hall, London, 1994.

[8] Chatfied, C. *Model Uncertainty, data mining and statistical inference (with discussion)*, Journal of the Royal Statistical Society, series B 158 1995, 419-466.

[9] Draper, D. *Assessment and propagation of model uncertainty (with discussion)*, Journal of the Royal Statistical Society, series B 57 1995 45-97.

[10] Buckland, S. T., Burnham, K. P. and Augustin, N. H. *Model selection: An integral part of inference*, Biometrics 53 1997, 603–618.

[11] Zucchini, W. *An introduction to model selection*, Journal of Mathematical Psychology 44 2000, 41–61.

[12] Candolo, C., Davison, A. C. and Demétrio, C. G. B. *A note on model uncertainty in linear regression*, The Statistician 158 2003, 165-177.

[13] Hjort, N. L. and Claeskens, G. *Frequentist model average estimators*, Journal of the American Statistical Association 98 2003, 879-899.

[14] Efron, B. *The estimation of prediction error: covariance penalties and cross-validation*, Journal of the American Statistical Association 99 2004, 619-642.

[15] Leeb, H. and Pötscher, B. M. *Model selection and inference: Fact and fiction*, Econometric Theory 21 2005, 21-59.

[16] Berk, R., Brown, L. D., Buja, A., Zhang, K. and Zhao, L. *Valid post-selection inference*, Submitted to Annals of Statistics, 2012.

[17] Longford, N. T. *Editorial: Model selection and efficiency-is 'which model ...?' the right question?*, Journal of Royal Statistical Society-A 168, Part 3 2005, 469-472.

[18] Schomaker, M., Wan, A. T. K. and Heumann, C. *Frequentist model averaging with missing observations*, Computational Statistics and Data Analysis 54 (12) 2010, 3336-3347.

[19] Zucchini, W., Claeskens, G. and Nguefack-Tsague, G. *Model selection*, In International Encyclopedia of Statistical Sciences, Editor: M. Lovric, Springer. Part 13, 830-833, 2011.

[20] Liu, W. and Yang, Y. *Parametric or nonparametric? a parametricness index for model selection*, Annals of Statistics 39 (4) 2011, 2074-2102.

[21] Nguefack-Tsague, G. and Zucchini, W. *Post-model selection inference and model averaging*, Pakistan Journal of Statistics and Operation Research 7(2-Sp) 2011, 347-361.

[22] Nguefack-Tsague, G. *An alternative derivation of some commons distributions functions: A post-model selection approach*, International Journal of Applied Mathematics and Statistics 42(12) 2013, 138-147.

[23] Nguefack-Tsague, G. *On bootstrap and post-model selection inference*, International Journal of Mathematics and Computation 21(4) 2013, 51-64.

[24] Nguefack-Tsague, G. *Bayesian estimation of a multivariate mean under model uncertainty*, International Journal of Mathematics and Statistics 13(1) 2013, 83-92.

[25] Nguefack-Tsague, G. *Estimation of a multivariate mean under model selection uncertainty*, Pakistan Journal of Statistics and Operation Research, 2014, forthcoming.

[26] Nguefack-Tsague, G., Zucchini, W. and Fotso, S. *On correcting the effects of model selection on inference in linear regression*, Syllabus Review 2(3) 2011, 122-140.

[27] Hoeting J., Madigan D., Raftery A. and Volinsky C. *Bayesian model averaging: A tutorial*, Statistical Science 4 1999, 382-417.

[28] Wasserman, L. *Bayesian model selection and model averaging*, Journal of Mathematical Psychology 44 2000, 92-107.

[29] Wan, A. T. K., Zhang, X. and Zou, G. *Frequentist model averaging estimation: a review*, Journal of Systems Science and Complexity 22 (4) 2009, 732-748.

[30] Hansen, B. E. *Least squares model averaging*, Econometrica 75 2007, 1175-1189.

[31] Hansen, B. E. *Least squares forecast averaging*, Journal of Econometrics 146 2008, 342-350.

[32] Hansen, B. E. *Averaging estimators for regressions with a possible structural break*, Econometric Theory 35 2009, 1498-1514.

[33] Hansen, B. E. *Averaging estimators for autoregressions with a near unit root*, Journal of Econometrics 158 2010, 142-155.

[34] Wan, A. T. K., Zhang, X. and Zou, G. *Least squares model averaging by Mallows criterion*, Journal of Econometrics 156(2) 2010, 277-283.

[35] Li, K. C. *Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set*, Annals of Statistics 15 1987, 958-975.

[36] Liang, H., Zou, G., Wan, A. T. K. and Zhang, X. *Optimal weight choice for frequentist model average estimators*, Journal of the American Statistical Association 106 2011, 1053-1066.

[37] Hansen, B. E. and Racine, J. S. *Jackknife model averaging*, Journal of Econometrics 167 (1) 2012, 38-46.

[38] Zhang X., Zou G. and Liang H. *Model averaging and weight choice in linear mixed-effects models*, Biometrika 101 (1) 2014, 205-218.

[39] Akaike, H. *Information theory and an extension of the maximum likelihood principle*. In Second International Symposium on Information Theory, eds. B. Petrov and F. Csáki, Budapest: Akadémiai Kiadó 1973, 267-281.

[40] Schwarz, G. *Estimating the dimension of a model*, The Annals of Statistics 6 1978, 461-464.

[41] Leblanc, M. and Tibshirani, R. Combining estimates in regression and classification. *Journal of the American Statistical Association* 91 1996, 1641-1650.

[42] Claeskens, G. and Carroll, R. J. *An asymptotic theory for model selection inference in general semiparametric problems*, Biometrika 94 2007, 249-265.

[43] Yan, Y. *Combining different procedures for adaptive regression*, Journal of Multivariate Analysis 74 2000, 135-161.

[44] Yan, Y. *Regression with multiple candidate models: selecting or mixing?*, Statistica Sinica 13 2003, 783-809.

[45] Yan, Y. *Combining forecasting procedures: some theoretical results*, Econometric Theory 20 2004, 176-222.

[46] Yan, Y. *Aggregating regression procedures to improve performance*, Bernoulli 10 2004, 25-47.

[47] Zou, H. and Yang, Y. *Combining time series models for forecasting*, International Journal of Forecasting 20 2004, 69-84.

[48] Yuan, Z. and Yang, Y. *Combining linear regression models: when and how?*, Journal of the American Statistical Association 100 2005, 1202-1204.

[49] Shan, K. and Yang, Y. *Combining regression quantile estimators*, Statistica Sinica 19 2009, 1171-1191.

[50] R Development Core Team. *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Austria, 2011.

[51] Mallows, C. L. *Some comments on Cp*, Technometrics 15 1973, 661-675.