Economics Letters 131 (2015) 5–8

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Prediction model averaging estimator

Tian Xie *

Wuhan University, Economics and Management School (EMS), Wuhan, Hubei, 430072, China

HIGHLIGHTS

- Propose a new prediction model averaging (PMA) estimator.
- Prove that the PMA estimator is asymptotically optimal.
- Show that the PMA estimator has good performance in simulation.
- Demonstrate that PMA can lead to large gains in box office prediction accuracy.

ARTICLE INFO

Article history:
Received 26 January 2015
Received in revised form
19 March 2015
Accepted 19 March 2015
Available online 26 March 2015

JEL classification:
C52
C53
D03

Keywords:
Model averaging
Convex optimization
Social media big data

ABSTRACT

This paper proposes a new estimator for least squares model averaging. We propose computing the model weights by minimizing a prediction model averaging (PMA) criterion. We prove that the PMA estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error. In simulation experiments the PMA estimator is shown to have good finite sample performance. As an empirical illustration, we demonstrate that using PMA to account for model uncertainty can lead to large gains in box office prediction accuracy.

1. Introduction

Hansen (2007) proposed the Mallows model average (MMA) method that computes model weights by minimizing a Mallows criterion. MMA can be interpreted as the model averaging version of the Mallows' C_p by Mallows (1973). Like its predecessor, MMA requires a preliminary estimate of the variance of error term. Hansen and Racine (2012) proposed a jackknife model averaging (JMA) that is free of preliminary estimate and is capable of dealing with heteroskedasticity. However, when the error term is homoskedastic, JMA can be less efficient than MMA, especially when sample size is small.[1] In a recent paper, Zhang et al. (in press) proposed a modified MMA based on Kullback–Leibler distance (KLMA). KLMA is shown to be more efficient than MMA in small sample size and an extended version is suggested for heteroskedasticity.

In this paper, we propose a new model average estimator with empirical weights computed through numerical minimization of a prediction model averaging (PMA) criterion. Our criterion can be seen as a model averaging version of the original prediction criterion (PC) proposed by Amemiya (1980). Amemiya (1980) demonstrated that PC does not rely on any preliminary estimates and has a better prediction efficiency than Mallows' C_p.

We prove that the PMA estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error. In simulation experiments the PMA estimator is shown to have significant efficiency gains over other methods. As an empirical illustration, we consider the case of using social media big data to predict subsequent box office revenue of movies. We demonstrate that using PMA to account for model uncertainty can lead to large gains in box office prediction accuracy.

This paper continues with an introduction of the framework of PMA in Section 2. Section 3 proves the asymptotic optimality of the PMA estimator. In Section 4, we conduct simulation experiments. Section 5 presents the empirical application.

* Tel.: +86 27 68752337.
E-mail address: xietian@whu.edu.cn.

[1] This has been demonstrated in the simulation section of Hansen and Racine (2012).

http://dx.doi.org/10.1016/j.econlet.2015.03.027
0165-1765/

2. Prediction model averaging estimator

Let $(y_i, \boldsymbol{x_i}) : i = 1, \ldots, n$ be a random sample, where y_i and $\boldsymbol{x_i} = [x_{i1}, x_{i2}, \ldots]$ are real-valued. Assume the data generating process is $y_i = \mu_i + u_i$, where $\mu_i = \sum_{j=1}^{\infty} \beta_j x_{ij}$, $\mathbb{E}(u_i|\boldsymbol{x_i}) = 0$ and $\mathbb{E}(u_i^2|\boldsymbol{x_i}) = \sigma^2$. The DGP in matrix form is $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{u}$, where all three vectors are $n \times 1$.

Consider a sequence of linear approximation models $m = 1, 2, \ldots, M$. The concept of approximation model can be vague. In this paper, an approximation model m uses $k^{(m)}$ regressors belonging to $\boldsymbol{x_i}$ such that $\boldsymbol{y} = \boldsymbol{X}^{(m)}\boldsymbol{\beta}^{(m)} + \boldsymbol{u}^{(m)}$, where $\boldsymbol{\beta}^{(m)}$ is a $k^{(m)} \times 1$ coefficient in model m. Let $\boldsymbol{P}^{(m)} = \boldsymbol{X}^{(m)}(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)})^{-1}\boldsymbol{X}^{(m)\top}$ be the projection matrix. The least squares estimate of $\boldsymbol{\mu}$ from model m is $\hat{\boldsymbol{\mu}}^{(m)} = \boldsymbol{P}^{(m)}\boldsymbol{y}$.

Let $\boldsymbol{w} = \left[w^{(1)}, \ldots, w^{(M)}\right]^\top$ be a weight vector in the unit simplex in $\mathbb{R}^M$,

$$\boldsymbol{H}_M \equiv \left\{\boldsymbol{w} \in [0, 1]^M : \sum_{m=1}^{M} w^{(m)} = 1\right\}.$$

Define the model average estimator of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^{(m)}\hat{\boldsymbol{\mu}}^{(m)} = \sum_{m=1}^{M} w^{(m)}\boldsymbol{P}^{(m)}\boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y},$$

where $\boldsymbol{P}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^{(m)}\boldsymbol{P}^{(m)}$ is the weighted average projection matrix.

We propose the prediction model averaging (PMA) criterion:

$$\mathrm{PMA}_n(\boldsymbol{w}) = \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)^\top\big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)\left(\frac{n + k(\boldsymbol{w})}{n - k(\boldsymbol{w})}\right), \tag{1}$$

where $k(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^{(m)}k^{(m)}$ is the effective number of parameters. PMA can be understood as the model averaging version of the prediction criterion by Amemiya (1980). Like most model selection criteria and model averaging criteria, PMA balances between the fit and the complexity of a model. Criterion (1) can be used to calculate the empirical weight vector $\hat{\boldsymbol{w}}$, in which

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\boldsymbol{H}_M}{\arg\min}\ \mathrm{PMA}_n(\boldsymbol{w}).$$

The Mallows' model average (MMA) criterion proposed by Hansen (2007) is

$$\mathrm{MMA}_n(\boldsymbol{w}) = \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)^\top\big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big) + 2\sigma^2 k(\boldsymbol{w}). \tag{2}$$

The empirical weights $\hat{\boldsymbol{w}}$ can be selected by minimizing (2) subject to $\boldsymbol{w} \in \boldsymbol{H}_M$. Note that the penalty term includes an unknown σ^2 that must be replaced by a sample estimate (usually provided by the largest model).

The MMA estimator is a two-step estimator since a sample estimate of σ^2 must be provided prior to estimation. In contrast, the PMA estimator is a continuous updating estimator that requires only one step of calculation. Estimating $\hat{\boldsymbol{w}}$ from the MMA criterion with constraints is a classic quadratic programming problem, while estimating $\hat{\boldsymbol{w}}$ by the PMA estimator is a convex optimization problem.[2]

Jackknife model averaging (JMA) (Hansen and Racine, 2012) is also known as leave-one-out cross-validation model averaging. As its name indicates, JMA requires the use of the jackknife residuals for the average estimator. The jackknife residual vector for model m can be conveniently written as $\hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(m)} = \boldsymbol{D}^{(m)}\hat{\boldsymbol{u}}^{(m)}$, where $\hat{\boldsymbol{u}}^{(m)}$ is the least squares residual vector and $\boldsymbol{D}^{(m)}$ is the $n \times n$ diagonal matrix with the ith diagonal element equal to $(1 - h_i^{(m)})^{-1}$. The term $h_i^{(m)}$ is the ith diagonal element of the projection matrix $\boldsymbol{P}^{(m)}$. Define an $n \times M$ matrix that collects all the jackknife residuals, in which $\hat{\boldsymbol{U}}_{\boldsymbol{J}} = \left[\hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(1)}, \ldots, \hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(M)}\right]$. The least squares cross-validation criterion for JMA is simply

$$\mathrm{CV}_n(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{w}^\top\hat{\boldsymbol{U}}_{\boldsymbol{J}}^\top\hat{\boldsymbol{U}}_{\boldsymbol{J}}\boldsymbol{w} \quad \text{with } \hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\boldsymbol{H}_M}{\operatorname{argmin}}\,\mathrm{CV}_n(\boldsymbol{w}).$$

The Kullback–Leibler distance based model averaging (KLMA) method proposed by Zhang et al. (in press) for the homoskedasticity case can be expressed as

$$\mathrm{KLMA}_n(\boldsymbol{w}) = \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)^\top\big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big) + 2\hat{\sigma}_L^2\left(\frac{n - k_L}{n - k_L - 2}\right)k(\boldsymbol{w}),$$

where $\hat{\sigma}_L^2$ is the estimated σ^2 by the largest model and k_L is its number of parameters. We can see that KLMA is a modified MMA with more penalty on complexity. As Zhang et al. (in press) demonstrated, KLMA is more efficient than MMA in small sample size and is asymptotically optimal like MMA in a similar fashion. Zhang et al. (in press) also include an extended KLMA for heteroskedasticity.

3. Asymptotic properties

In this section, we demonstrate the asymptotic optimality of the PMA estimator by showing that it achieves the lowest possible mean squared error as $n \to \infty$.

Define the average mean squared error as $L_n(\boldsymbol{w}) \equiv \big(\boldsymbol{\mu}(\boldsymbol{w}) - \boldsymbol{\mu}\big)^\top\big(\boldsymbol{\mu}(\boldsymbol{w}) - \boldsymbol{\mu}\big)$ and the conditional average mean squared error as $R_n(\boldsymbol{w}) \equiv \mathbb{E}\big(L_n(\boldsymbol{w})|\boldsymbol{X}\big)$. We assume the following:

Assumption 1. For some fixed integer $1 \le G < \infty$, we have $\mathbb{E}(|u_i|^{4G}|\boldsymbol{x}_i) \le \kappa < \infty$.

Assumption 2. As $n \to \infty$, $\xi_n^{-2G}M\sum_{m=1}^{M}\big(R_n(\boldsymbol{w}_m^0)\big)^G \to 0$, where $\xi_n = \inf_{\boldsymbol{w}\in\boldsymbol{H}_M} R_n(\boldsymbol{w})$ and $\boldsymbol{w}_m^0$ is an $M \times 1$ vector of which the mth element is one and the others are zeros.

Assumption 3. As $n \to \infty$, $k^{(m)} \to \infty$ and $k^{(m)}/n \to 0$ for all m.

Assumption 1 is a bound condition on the conditional moments of the error term. Assumption 2 is the convergence condition. Assumption 3 states that as n goes to infinity, $k^{(m)}$ goes to infinity at a slower rate for $m = 1, \ldots, M$. The following theorem demonstrates the asymptotic optimality of the PMA estimator.

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Then, as $n \to \infty$*

$$\frac{L_n(\hat{\boldsymbol{w}})}{L_n(\boldsymbol{w}_{opt})} \xrightarrow{p} 1,$$

where

$$\boldsymbol{w}_{opt} = \underset{\boldsymbol{w}\in\boldsymbol{H}_M}{\arg\inf}\ L_n(\boldsymbol{w}).$$

Theorem 1 states that by using the empirical weight vector $\hat{\boldsymbol{w}}$, the mean squared error is asymptotically equivalent to the lowest possible mean squared error. This implies that the PMA estimator is asymptotically optimal in the class of model average estimators where the weight vector belongs to the set $\boldsymbol{H}_M$.

Remark. The proof of Theorem 1 is a straightforward application of the same technique demonstrated in Li (1987) and Wan et al. (2010). See the appendix file for a detailed proof (see Appendix A).

[2] Note that convex optimization usually requires slightly more computation time than quadratic programming.