Journal of Econometrics 192 (2016) 139–151

Contents lists available at ScienceDirect

Journal of Econometrics

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

Model averaging based on leave-subject-out cross-validation

Yan Gao [a,b], Xinyu Zhang [b,c,*], Shouyang Wang [b], Guohua Zou [b,d]

[a] Department of Statistics, College of Science, Minzu University of China, Beijing 100081, China
[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[c] ISEM, Capital University of Economics and Business, Beijing 100070, China
[d] School of Mathematical Science, Capital Normal University, Beijing 100037, China

ARTICLE INFO

Article history:
Received 23 August 2014
Received in revised form
16 April 2015
Accepted 20 July 2015
Available online 28 December 2015

JEL classification:
C51
C52

Keywords:
Asymptotic optimality
Leave-subject-out cross-validation
Longitudinal data
Model averaging
Time series

ABSTRACT

This paper develops a frequentist model averaging method based on the leave-subject-out cross-validation. This method is applicable not only to averaging longitudinal data models, but also to averaging time series models which can have heteroscedastic errors. The resulting model averaging estimators are proved to be asymptotically optimal in the sense of achieving the lowest possible squared errors. Both simulation study and empirical example show the superiority of the proposed estimators over their competitors.

1. Introduction

Model averaging (MA), a smoothed extension of model selection (MS), generally yields a lower risk than model selection. There are many studies on Bayesian model averaging. Hoeting et al. (1999) provided a comprehensive review in this direction. Recent years have also witnessed a booming development of frequentist model averaging methods such as weighting strategy based on the scores of information criteria (Buckland et al., 1997; Hjort and Claeskens, 2003, 2006; Zhang and Liang, 2011; Zhang et al., 2012; Xu et al., 2014), the adaptive regression by mixing by Yang (2001), Mallows model averaging (MMA) by Hansen (2007), and optimal mean squared error averaging by Liang et al. (2011). Recently, taking heteroscedasticity into consideration, Hansen and Racine (2012) proposed a jackknife model averaging (JMA) method that selects weights by minimizing a leave-one-out cross-validation criterion. The JMA estimator performs quite well in cross-sectional data. However, for longitudinal data, there generally exists within-subject correlation in error terms, thus the JMA method may not be appropriate. In the current paper, we develop a model averaging estimator called leave-subject-out model averaging (LsoMA) estimator for longitudinal data models.

There exists a rich literature on longitudinal data models. For an overview of parametric longitudinal data models, one can refer to Arellano (2003), Hsiao (2003) and Baltagi (2005). Nonparametric (Rice and Silverman, 1991; Fan and Zhang, 2000; Welsh et al., 2002; Zhu et al., 2008) and semiparametric longitudinal data models (Zeger and Diggle, 1994; Zhang et al., 1998; Lin and Ying, 2001) have also been considered. Penalized model selection methods are commonly used in nonparametric and semiparametric models. In the current paper, we use a quadratic penalty based on smoothing splines. The popular nonquadratic penalties, such as the least absolute shrinkage and selection operator (Tibshirani, 1996), hard thresholding (Antoniadis, 1997; Fan, 1997), and the smoothly clipped absolute deviation penalties (Fan and Li, 2001) can also be utilized here. For all these methods, tuning parameters need to be selected. Rice and Silverman (1991) introduced the leave-subject-out cross-validation (LsoCV) to select tuning parameters. This method has been widely used in longitudinal data model since then. For example, Xu and Huang (2012) utilized the LsoCV to select variables in the semiparametric longitudinal data model, and proved the asymptotic optimality of their approach. Further, they developed an efficient computation procedure for the LsoCV.

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: xinyu@amss.ac.cn (X. Zhang).

http://dx.doi.org/10.1016/j.jeconom.2015.07.006
0304-4076/

The current paper focuses on model averaging. By using our model averaging method, the estimators based on different covariates or different tuning parameters are asymptotically optimally combined, i.e., under some regularity conditions, our model averaging estimator minimizes predictive squared error in large sample cases. A related work dealing with longitudinal data was done by Zhang et al. (2014). They developed a model averaging method to combine forecasts from linear mixed-effects models. In the current paper, to be more general, we allow each candidate model to contain non-parametric component.

On the other hand, series dependence always exists in time series models. A natural idea is to treat those with high correlation as a subject, and thus we can use the LsoMA method to combine time series models. In this paper, we will show that the resulting LsoMA estimator is also asymptotically optimal. In Cheng and Hansen (2015), the LsoCV criterion was suggested to average forecasts for factor-augmented regressions, but the corresponding asymptotic optimality was not achieved in that work.

We do a Monte Carlo study to compare the finite sample performance of the proposed LsoMA method with others including model selection methods by AIC, BIC and LsoCV, and model averaging methods by smoothed AIC and smoothed BIC (Buckland et al., 1997), and JMA in both longitudinal data model and time series model. Simulation results indicate that the LsoMA estimator performs better than its competitors in most cases. We also conduct an empirical study on the Chinese consumer price index, which shows that our method has better forecasting performance than the commonly used model selection and averaging methods.

The remainder of this paper is organized as follows. Section 2 proposes the LsoMA estimator for longitudinal data model and develops its asymptotic optimality theory. Section 3 studies the LsoMA method for time series models. Section 4 numerically compares our LsoMA estimators with those obtained from some commonly used model selection and model averaging methods. Section 5 conducts an empirical study. Section 6 concludes. The proofs are relegated to Appendix.

2. Leave-subject-out model averaging for longitudinal data models

2.1. Model framework

Suppose that $(y_{ij}, \mathbf{x}_{ij}), j = 1, \ldots, T_i$, are observations for subject i, $i = 1, \ldots, n$. Let $\mathbf{Y}_i = (y_{i1}, \ldots, y_{iT_i})'$, $\widetilde{\mathbf{X}}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i})'$ and $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1', \ldots, \widetilde{\mathbf{X}}_n')'$. We consider the following semiparametric model for longitudinal data

$$\mathbf{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$
$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT_i})',$$
$$\mu_{ij} = E(y_{ij}|\widetilde{\mathbf{X}}_i) = \mathbf{x}_{ij,0}'\boldsymbol{\beta}_0 + \sum_{l=1}^{L} f_l(x_{ij,l}), \quad j = 1, \ldots, T_i,$$

where $\mathbf{x}_{ij,0}$ contains variables of parametric component, $x_{ij,1}, \ldots, x_{ij,L}$ are variables of nonparametric component, $\boldsymbol{\beta}_0$ is the coefficient vector of the linear component, $f_1, \ldots, f_L$ are smooth functions, and $\boldsymbol{\varepsilon}_i$'s are independent disturbances with $E(\boldsymbol{\varepsilon}_i|\widetilde{\mathbf{X}}_i) = 0$ and $E(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\widetilde{\mathbf{X}}_i) = \boldsymbol{\Sigma}_i$. We can use a basis expansion to approximate each f_l. Then, there exist a design matrix $\mathbf{X}_i$ and an unknown parameter vector $\boldsymbol{\beta}$ such that $\boldsymbol{\mu}_i \approx \mathbf{X}_i\boldsymbol{\beta}$. Specifically, $\mathbf{X}_i = (\mathbf{X}_{1,i}^*, \mathbf{X}_{2,i}^*)$ consists of two parts: the linear regression variables matrix $\mathbf{X}_{1,i}^*$, and the basis matrix $\mathbf{X}_{2,i}^*$ used to approximate the nonparametric component.

We estimate $\boldsymbol{\beta}$ by minimizing the following penalized weighted least squares (Xu and Huang, 2012)

$$\mathrm{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{l=1}^{L}\lambda_l\boldsymbol{\beta}'\mathbf{F}_l\boldsymbol{\beta},$$

where $\mathbf{V}_i$'s are working covariance matrices, $\mathbf{F}_l$ is a positive semi-definite matrix such that $\boldsymbol{\beta}'\mathbf{F}_l\boldsymbol{\beta}$ serves as a roughness penalty for f_l, and $\lambda_1, \ldots, \lambda_L$ are tuning parameters. In the current paper, penalties are put only on the nonlinear parts, so $\mathbf{F}_l$ is a block diagonal matrix with the block corresponding to the linear part being $\mathbf{0}$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)'$, $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n')'$, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_n')'$, and $\mathbf{V} = \mathrm{diag}\{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$. The estimator of $\boldsymbol{\beta}$ can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \sum_{l=1}^{L}\lambda_l\mathbf{F}_l\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

There are many methods for constructing the basis matrix $\mathbf{X}_2^* \equiv (\mathbf{X}_{2,1}^{*\prime}, \ldots, \mathbf{X}_{2,n}^{*\prime})'$ and the penalty matrix $\mathbf{F}_l$. For example, one can use the spline basis to generate $\mathbf{X}_2^*$. Linear spline is the simplest method but has a sharp corner disadvantage. Quadratic spline basis can remedy this disadvantage as it has a continuous first derivative. The truncated power basis of degree higher than two provides more complex regression functions. However, it may lead to numerical instability due to its nonorthogonality, which can be overcome by the B-spline basis. For the penalty matrix $\mathbf{F}_l$, we can utilize the squared second-difference penalty, the squared second derivative penalty or the thin-plate splines penalty. More details on the basis and penalty matrices can be found in Green and Silverman (1994) and Ruppert et al. (2003).

Following Claeskens et al. (2009), we take penalized B-spline as an example to show how to construct the basis and penalty matrices. For simplicity, we consider the model with only one nonparametric covariate, so that there is only one tuning parameter λ and $L = 1$. Let r be the order of B-splines. Define a sequence of knots on the interval $[I_{low}, I_{up}]$: $I_{low} = m_{-(r-1)} = \cdots = m_0 < m_1 < \cdots < m_{K_n} < m_{K_n+1} = \cdots = m_{K_n+r} = I_{up}$. Basis functions can be expressed as

$$B_{j,1}(x) = \begin{cases} 1, & m_j \le x < m_{j+1}, \\ 0, & \text{otherwise}, \end{cases}$$

$$B_{j,r}(x) = \frac{x - m_j}{m_{j+r-1} - m_j}B_{j,r-1}(x) + \frac{m_{j+r} - x}{m_{j+r} - m_{j+1}}B_{j+1,r-1}(x),$$

for $j = -(r-1), \ldots, K_n$. Then, the kth row of $\mathbf{X}_{2,i}^*$ is $(B_{-(r-1),r}(x_{ik,1}), \ldots, B_{K_n,r}(x_{ik,1}))$. The penalty term can be written as $\lambda\boldsymbol{\beta}_2'\boldsymbol{\Delta}_q'\mathbf{R}\boldsymbol{\Delta}_q\boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_2$ is the coefficient vector of $\mathbf{X}_2^*$, $\mathbf{R}$ is a $(K_n + r - q) \times (K_n + r - q)$ matrix with its ij element $R_{ij} = \int_{I_{low}}^{I_{up}} B_{j,r-q}(x)B_{i,r-q}(x)dx$, and $\boldsymbol{\Delta}_q$ is a matrix of qth order difference operator. If the knots are equidistant, i.e., $m_j - m_{j-1} = \delta$ for $j = 1, \ldots, K_n + 1$, then $\boldsymbol{\Delta}_q$ can be expressed in terms of the qth backward difference operator ∇_q, i.e., $\boldsymbol{\Delta}_q = \delta^{-q}\nabla_q$. Each element of the matrix ∇_q is defined recursively via $\nabla_q = \nabla_1(\nabla_{q-1})$ and $\nabla_1\beta_j = \beta_j - \beta_{j-1}$. If we take $q = r - 1$, then $\mathbf{R}$ reduces to a diagonal matrix with the diagonal element δ.

The working covariance matrices $\mathbf{V}_i$'s are generally estimated based on the working correlation structures of $\boldsymbol{\varepsilon}_i$'s. In practice, compound symmetry and autoregressive structures are commonly used working correlation structures. As commented by Liang and Zeger (1986), a possibly misspecified correlation structure also has a potential to improve the estimation efficiency over a method that completely ignores the within-subject correlation. Diggle et al. (2002) provided details on the choice of working correlation structure for longitudinal data. In the current paper, following Xu and Huang (2012), we set $\mathbf{V}_i$'s to be the identity matrices at first, based on which the model is estimated to get residuals, and then $\mathbf{V}_i$'s are estimated using these residuals.

2.2. Model averaging criterion

Assume that candidate estimators differ from each other in regressors and/or tuning parameters. Let $\mathcal{X} = \{\mathbf{X}^{(1)}, \ldots,$