ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

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

Model averaging by jackknife criterion in models with dependent data

Xinyu Zhang [a], Alan T.K. Wan [b,*], Guohua Zou [a]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[b] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 6 September 2011
Received in revised form
1 August 2012
Accepted 19 January 2013
Available online xxxx

JEL classification:
C51
C52

Keywords:
Asymptotic optimality
Autocorrelation
Cross-validation
Lagged dependent variables
Model averaging
Squared error

ABSTRACT

The past decade witnessed a literature on model averaging by frequentist methods. For the most part, the asymptotic optimality of various existing frequentist model averaging estimators has been established under i.i.d. errors. Recently, Hansen and Racine [Hansen, B.E., Racine, J., 2012. Jackknife model averaging. Journal of Econometrics 167, 38–46] developed a jackknife model averaging (JMA) estimator, which has an important advantage over its competitors in that it achieves the lowest possible asymptotic squared error under heteroscedastic errors. In this paper, we broaden Hansen and Racine's scope of analysis to encompass models with (i) a non-diagonal error covariance structure, and (ii) lagged dependent variables, thus allowing for dependent data. We show that under these set-ups, the JMA estimator is asymptotically optimal by a criterion equivalent to that used by Hansen and Racine. A Monte Carlo study demonstrates the finite sample performance of the JMA estimator in a variety of model settings.

1. Introduction

Model averaging is an alternative to model selection. While model selection attempts to find a single best model for the given purpose, model averaging compromises across the competing models, thus providing a kind of insurance against selecting a very poor model. Model averaging has long been a popular approach within the Bayesian paradigm. In recent years, frequentist model averaging (FMA) has also made substantial grounds. Contributions to model averaging from a fully-fledged frequentist standpoint were made by Buckland et al. (1997), Yang (2001), Hjort and Claeskens (2003, 2006), Yuan and Yang (2005), Hansen (2007, 2008), Goldenshluger (2009), Schomaker et al. (2010), Wan et al. (2010), Liang et al. (2011), Zhang and Liang (2011), Zhang et al. (2012), among others. The majority of these studies emphasize model weights determination, inference based on model averaging, and asymptotic efficiency and finite sample properties of FMA estimators under a variety of model settings. Useful surveys of this rapidly expanding body of literature are given in Claeskens and Hjort (2008) and Wang et al. (2009). There is also an emerging empirical literature that employs FMA in applied settings (Kapetanios et al., 2008a,b; Pesaran et al., 2009; Wan and Zhang, 2009).

In a recent article, Hansen and Racine (2012) (hereafter referred to as HR, 2012) developed a jackknife model averaging (JMA) estimator that selects model weights by minimizing a cross-validation criterion. One major advantage of the JMA estimator is that the asymptotic optimality theory developed for it allows for heteroscedasticity in the errors, whereas those developed for other existing FMA schemes virtually all assume i.i.d. errors. HR (2012) showed that the JMA estimator has the smallest asymptotic expected squared errors relative to a large class of linear estimators constructed from a countable set of weights, including the least squares, ridge, Nadaraya–Watson and local polynomial kernel with fixed bandwidths, spline and some other nonparametric estimators. HR's (2012) Monte Carlo results also suggest that the JMA estimator is generally preferred to several other model selection and averaging estimators; in particular, when the errors are heteroscedastic, the JMA estimator significantly outperforms the Mallows model average (MMA) estimator developed by Hansen (2007) in mean squared error (MSE) terms in a large part of the parameter space. In view of these merits of the JMA estimator, more investigations into its properties are warranted.

Although HR's (2012) model set-up admits heteroscedastic errors, it rules out serial correlations in the errors. Their set-up also assumes complete exogeneity of regressors. An interesting question is whether the JMA estimator remains meritorious under other settings, particularly in models that admit dependent data. The

* Corresponding author. Tel.: +852 34427146; fax: +852 34420189.
E-mail addresses: msawan@cityu.edu.hk, Alan.Wan@cityu.edu.hk (A.T.K. Wan).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2013.01.004

current paper takes steps in this direction by enlarging HR's scope of analysis to include two other commonly encountered model settings, both involving dependent data. The first setting retains the regressor exogeneity assumption as in HR but admits a non-diagonal covariance structure in the errors, thus allowing for error processes such as ARMA and GARCH in addition to pure heteroscedastic and i.i.d. processes; it also nests the model of HR as a special case. The second setting allows for lagged dependent variables but assumes that the errors are i.i.d. Although neither of these two model settings is more general than the other, they both allow dependent data, a property not shared by the model examined in HR. We prove that the JMA estimator when applied to these model settings achieves an asymptotic optimality criterion equivalent to that under the model set-up of HR. Our theoretical analysis follows the approach of Wan et al. (2010) by allowing the model weights to be continuous. This is unlike the method of HR which follows that of Hansen (2007) by restricting the weights to a discrete set. We consider the extension from discrete to continuous weighting an advance as the latter has obvious appeal. It is instructive to point out that the conditions required for optimality by our method are neither stronger nor weaker than those required by HR's method. Like the latter method, our method also allows for an infinite number of models. Detailed comparisons of the technical conditions that underpin our theoretical results and those of HR are provided in Sections 2.2 and 2.3. For our second model setting that involves lagged dependent variables, we prove the asymptotic optimality of the JMA estimator using results in Ing and Wei (2003).

In a Monte Carlo study we also compare the finite sample performance of the JMA estimator with several other estimators, including the MMA, leave-one-out cross-validation, and AIC and BIC model selection estimators under the two model set-ups considered. Our Monte Carlo results suggest that under strictly exogenous regressors and ARMA and GARCH-type errors, the JMA estimator is frequently preferred to these alternative estimators. On the other hand, when the regressors are not strictly exogenous but contain lagged dependent variables, the JMA estimator has comparable efficiency to the MMA estimator. The latter estimator is known to exhibit performance superiority in many settings (Hansen, 2007, 2008).

The plan of this paper is as follows. In Section 2 we examine the JMA criterion and present results on the asymptotic optimality of the JMA estimator under a setting that assumes exogeneity of regressors but allows for both serial correlation and heteroscedasticity. While the main theorem in this section is applicable to general linear estimators, a special focus of discussion will be given to least squares estimation in the linear regression model. Section 3 examines the case of an infinite order linear autoregressive (AR) data generating process, and JMA being performed across models containing lagged dependent variables and possibly other regressors. Section 4 reports results of the Monte Carlo study. Section 5 concludes, and proofs of theorems are contained in Appendix.

2. Jackknife model averaging under a non-diagonal error covariance structure

2.1. Model framework and the jackknife criterion

We follow HR's (2012) notations as much as possible for readers' convenience. Wherever appropriate we point out the differences in the two set-ups. Consider the data generating process (DGP)

$$y_i = \mu_i + e_i = f(x_i) + e_i, \quad i = 1, \dots, n, \tag{1}$$

with $x_i = (x_{i1}, x_{i2}, \dots)$ being countably infinite, and $f(\cdot)$ a function with respect to x_i. Write $y = (y_1, \dots, y_n)'$, $X = (x'_1, \dots, x'_n)'$, $\mu = (\mu_1, \dots, \mu_n)'$, and $e = (e_1, \dots, e_n)'$. Further, assume that $E(e|X) = 0$ so that $\mu = E(y|X)$, and denote $Var(e|X) = \Omega$, where Ω is a positive definite symmetric matrix.

Let M_n be the number of candidate models in the model average, and $\{\widehat{\mu}^1, \dots, \widehat{\mu}^{M_n}\}$ be a set of linear estimators of μ such that the mth estimator in the set, i.e., the estimator of μ in the mth model, may be written as $\widehat{\mu}^m = \mathbf{P}_m y$, where $\mathbf{P}_m$ is dependent on X but not on y. Many well-known estimators including the least squares, ridge, nearest neighbors, and spline are members of this class. Now, let $w = (w^1, \dots, w^{M_n})'$ be a weight vector in the continuous set:

$$\mathcal{H}_n = \left\{ w \in [0,1]^{M_n} : \sum_{m=1}^{M_n} w^m = 1 \right\}.$$

The model averaging estimator of μ is obtained by compromising across the linear estimators $\{\widehat{\mu}^1, \dots, \widehat{\mu}^{M_n}\}$ in the model space. It has the form

$$\widehat{\mu}(w) = \sum_{m=1}^{M_n} w^m \widehat{\mu}^m = \sum_{m=1}^{M_n} w^m \mathbf{P}_m y \equiv \mathbf{P}(w) y. \tag{2}$$

The above set-up is the same as that of HR (2012) except for the following aspects. First, HR (2012) restricted Ω to be a diagonal matrix, but we permit Ω to be non-diagonal, thus allowing the errors to be both autocorrelated and heteroscedastic. This also allows y to be dependent when the design matrix X is assumed fixed. Second, although HR (2012) defined $\widehat{\mu}(w)$ as in (2), when proving the asymptotic optimality of the JMA estimator, they restricted $\mathcal{H}_n$ to the subset $\mathcal{H}_n^*$, which consists of the discrete weights w^m from the set $\{0, 1/N, 2/N, \dots, 1\}$ for some positive integer N. We do not impose the same restriction in our analysis.

Denote $\tilde{\mu}^m$ as the estimator of μ when jackknife estimation based on the delete-one cross-validation is used. Write $\tilde{\mu}^m = (\tilde{\mu}_1^m, \dots, \tilde{\mu}_n^m)'$, where $\tilde{\mu}_i^m$ is the estimator of μ_i obtained with the ith observation (y_i, x_i) removed from the sample. Thus, we can write $\tilde{\mu}^m = \tilde{\mathbf{P}}_m y$, where $\tilde{\mathbf{P}}_m$ has zeros on the diagonal and depends only on X. The model averaging estimator that smooths across the M_n jackknife estimators is thus

$$\tilde{\mu}(w) = \sum_{m=1}^{M_n} w^m \tilde{\mu}^m = \sum_{m=1}^{M_n} w^m \tilde{\mathbf{P}}_m y \equiv \tilde{\mathbf{P}}(w) y. \tag{3}$$

HR (2012) adopted the following squared error loss criterion for choosing the weight vector w:

$$CV_n(w) = \|y - \tilde{\mu}(w)\|^2, \tag{4}$$

where $\|a\|^2 = a'a$. Now, let $\widehat{w} = \arg\min_{w \in \mathcal{H}_n} CV_n(w)$ be the weight vector that minimizes $CV_n(w)$. The JMA estimator of μ is $\widehat{\mu}(\widehat{w})$. It is obtained by substituting $\widehat{w}$ for w in (2). Thus, the JMA estimator is a weighted average of the linear estimators $\widehat{\mu}^m$'s using $\widehat{w}$ as weight. It is different from the estimator $\tilde{\mu}(w)$ which combines the jackknife estimators $\tilde{\mu}^m$'s.

Denote $\tilde{e}^m = y - \tilde{\mu}^m$ and $\tilde{e} = (\tilde{e}^1, \dots, \tilde{e}^{M_n})$. Then we can write

$$CV_n(w) = w'\tilde{e}'\tilde{e}w, \tag{5}$$

a quadratic function of w. Thus, the minimization of $CV_n(w)$ with respect to w is a quadratic programming problem. Numerous software packages are available for obtaining a solution to this problem (e.g., Matlab and R), and they generally work effectively and efficiently even when M_n is large; for example, when $n = 200$ and $M_n = 100$, it takes only 0.15 s to obtain the solution to (5) by Matlab.

A referee pointed out that one could consider block cross-validation as an alternative to delete-one cross-validation. Although traditionally the choice of block lengths has been an issue, recent advances in automated methods (e.g., Politis and White, 2004; Patton et al., 2009) have made the selection of optimal block length practically feasible. Racine (1997) also showed that the amount of calculations needed for deleting a block can