Economics Letters 135 (2015) 42–45

Contents lists available at ScienceDirect

Economics Letters

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

Estimating average treatment effect by model averaging

Yichen Gao[a], Wei Long[b,*], Zhengwei Wang[c]

[a] ISEM, the Capital University of Economics & Business, Beijing, 100070, China
[b] Department of Economics, Tulane University, New Orleans, LA 70118, United States
[c] PBC School of Finance, Tsinghua University, Beijing, 100083, China

HIGHLIGHTS

- We propose to use a model average method to improve the estimation of average treatment effects.
- The proposed model average estimator selects weight optimally to minimize estimation mean squared errors.
- Simulation results show that the model average estimator exhibits smaller estimation mean squared errors in post-treatment prediction than AIC or AICC methods.

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
17 July 2015
Accepted 2 August 2015
Available online 8 August 2015

JEL classification:
C5

Keywords:
Average treatment effect
Counterfactual
Model average

ABSTRACT

In this paper, we propose to use a model average method to improve the estimation performance of Hsiao et al. (2012) panel data approach for program evaluation. Instead of using the two-step model selection strategy which chooses one best model according to a criterion such as AIC or AICC, we average over a set of candidate models. Simulation results show that the model average estimator exhibits smaller estimation errors in post-treatment prediction than AIC or AICC method.

1. Introduction

Econometric models for treatment effect estimation have been extensively discussed in literature these days. To estimate the treatment effect of a program or a policy, one needs to measure the difference between the outcomes under treatment and the outcomes when treatment is absent, since we do not observe both outcomes simultaneously. Hsiao et al. (2012) propose a panel data approach to estimate the counterfactual outcomes. This method relies on the correlation between the treatment and the control units. Hsiao et al. (2012) argue that it is the presence of common factors that cause the cross-sectional dependence and drive co-movement of all the relevant cross-sectional outcome variables. Bai et al. (2014) further extend Hsiao et al. (2012) method to allow for non-stationary data. In finite sample applications, Hsiao et al. (2012) propose to select the best model through a two-step strategy, which is based on comparing the goodness-of-fit statistic and widely used model selection criterion such as Akaike information criterion (AIC, Akaike, 1970) and corrected Akaike information criterion (AICC, Hurvich and Tsai, 1989).

In this paper, we contribute to improve the estimating performance of Hsiao et al. (2012) method by using the Jackknife model average (JMA) method, which is proposed by Hansen and Racine (2012). Instead of selecting one specific "best" model based on a criterion, model average method addresses the model uncertainty problem by averaging over the set of candidate models. We direct interested readers to Buckland et al. (1997), Hansen (2007) and Wan et al. (2010) for frequentist method for model averaging, Hoeting et al. (1999) for Bayesian model averaging. Under model average framework, we select weights for each candidate model by minimizing a cross-validation criterion function. Hansen and Racine (2012) show that the computing procedure of JMA is an application of the quadratic programming technique. They further prove that JMA estimator is asymptotically optimal in the sense of achieving the lowest possible expected squared estimation loss. By

[*] Corresponding author.
E-mail address: wlong2@tulane.edu (W. Long).

http://dx.doi.org/10.1016/j.econlet.2015.08.002
0165-1765/© 2015 Elsevier B.V. All rights reserved.

replacing the two-step selection strategy with JMA method, we improve the post-treatment prediction of Hsiao et al. (2012) in terms of mean squared prediction errors (PMSE).

The rest of the paper is organized as follows. Section 2 briefly reviews both Hsiao et al. (2012) method and the JMA method. Section 3 reports simulation results to examine the finite sample performance of our proposed method. Section 4 concludes the paper.

2. Theoretical model

In this section we briefly discuss the estimation method in Hsiao et al. (2012). Suppose there is no treatment to all units up to T_1. At time $T_1 + 1$, there is only one unit that receives a treatment. Let y_t be the treatment unit's outcome at time t. Correspondingly, let $x_t = (x_{1t}, \ldots, x_{Nt})'$ be the outcomes of N control units at time t.[1] Hsiao et al. (2012) consider the case that both treatment and control units' outcomes are generated by a factor model (e.g., Bai and Ng, 2002) in the absence of treatment for $t = 1, \ldots, T_1$:

$$\tilde{y}_t = a + Bf_t + u_t, \tag{1}$$

where $\tilde{y}_t = (y_t, x_{1t}, \ldots, x_{Nt})'$, $a = (a_1, \ldots, a_{N+1})'$, f_t is a $K \times 1$ vector of common factors (they may be unobservable) that affect outcomes, B is a $(N + 1) \times K$ matrix of factor loading, $u_t = \left(u_{1t}, \ldots, u_{(N+1)t}\right)'$ is a vector of idiosyncratic error. Let y_t^1 and y_t^0 denote the outcomes of the treated unit with and without the policy intervention, respectively. Given that there is a treatment at time $T_1 + 1$, we are interested in estimating the average treatment effects $\Delta_1 = E(y_t^1 - y_t^0)$. The difficulty is that we cannot observe y_t^0 for $t \geq T_1 + 1$. Hsiao et al. (2012) suggest using control units' outcomes x_t to estimate y_t^0 when $t \geq T_1 + 1$. This can be done by replacing f_t by x_t in the treatment unit's equation $y_t = a_1 + b_1' f_t + u_t$ to obtain

$$y_t = \gamma_0 + x_t'\gamma + v_t, \tag{2}$$

for $t = 1, \ldots, T_1$, where γ_0 is intercept, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_N)'$, v_t satisfies that $E(v_t) = 0$, $E(v_t x_t) = 0$ and $var(v_t)$ is finite. Let $\hat{\gamma}_0$ and $\hat{\gamma}$ denote the least square estimators of γ_0 and γ based on (2), then we estimate the counterfactual outcome of y_t^0 by

$$\hat{y}_t^0 = \hat{\gamma}_0 + x_t'\hat{\gamma}, \tag{3}$$

for $t = T_1 + 1, \ldots, T$. Let $T_2 = T - T_1$, then the average treatment effect is estimated by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left(y_t - \hat{y}_t^0\right). \tag{4}$$

In application, N may not be small relative to T_1. Thus, it is advantageous to use only a subset of the N control units rather than all of them to predict the counterfactuals. For N control units, there are 2^N different models, and the most appropriate model should balance the within-sample fit with the out-sample prediction error. Hsiao et al. (2012) propose a two-step model selection procedure to find out which model is the most appropriate. Specifically, in the first step, they use R^2 to select the best predictor for y_t^0 using k control units out of N control units, denoted by $M(k)^*$, for $k = 1, \ldots, N$. Then, in the second step, from $M(1)^*, M(2)^*, \ldots, M(N)^*$, they pin down one best model from the N candidate models in terms of model selection criterion such as Akaike information criterion (AIC) and corrected Akaike information criterion (AICC).

In this paper, we avoid choosing one "best" model out of the 2^N models as it could be time consuming when N is large. On the contrary, we apply the model average method. First, we regress y_t on the outcomes of each individual control unit $x_{1t}, \ldots, x_{Nt}$, $t = 1, 2, \ldots, T_1$, and then obtain their respective goodness-of-fit statistics $R_1^2, R_2^2, \ldots, R_N^2$. We order these goodness-of-fit statistics so that $R_{(1)}^2 \geq R_{(2)}^2 \geq \cdots \geq R_{(N-1)}^2 \geq R_{(N)}^2$, where $R_{(\cdot)}^2$ is the order statistic, and $x_{(1)}, x_{(2)}, \ldots, x_{(N-1)}, x_{(N)}$ correspond to the regressor in each model, i.e., $x_{(j)}$ is the regressor that has a goodness-of-fit $R_{(j)}^2$, $j = 1, \ldots, N$. Second, we form our first candidate model by regressing y_t on $(1, x_{(1)})$ and then denote the first model's linear fit by a $T_1 \times 1$ vector $\hat{\mu}^1 = \left(\hat{\mu}_1^1, \hat{\mu}_2^1, \ldots, \hat{\mu}_{T_1}^1\right)'$, where $\hat{\mu}_t^1 = (1, x_{(1),t})\hat{\boldsymbol{\gamma}}^1$ for $t = 1, 2, .., T_1$ and $x_{(1),t}$ represents the observation for $x_{(1)}$ at time t. $\hat{\boldsymbol{\gamma}}^1 = \left(\hat{\gamma}_0^1, \hat{\gamma}_{(1)}^1\right)'$ denotes the OLS estimator based on regressing y_t on $(1, x_{(1),t})$ with $t = 1, \ldots, T_1$. The second candidate model regresses y_t on $(1, x_{(1),t}, x_{(2),t})$ with $t = 1, \ldots, T_1$, and the corresponding linear estimate is $\hat{\mu}^2 = \left(\hat{\mu}_1^2, \hat{\mu}_2^2, \ldots, \hat{\mu}_{T_1}^2\right)'$, where $\hat{\mu}_t^2 = (1, x_{(1),t}, x_{(2),t})\hat{\boldsymbol{\gamma}}^2$ for $t = 1, 2, .., T_1$ and $\hat{\boldsymbol{\gamma}}^2 = \left(\hat{\gamma}_0^2, \hat{\gamma}_{(1)}^2, \hat{\gamma}_{(2)}^2\right)'$ denotes the OLS estimator obtained from regressing y_t on $(1, x_{(1),t}, x_{(2),t})$ with $t = 1, \ldots, T_1$. We continue this procedure until all N control units are included, i.e., the last one is to regress y_t on $\left(1, x_{(1),t}, x_{(2),t}, \ldots, x_{(N),t}\right)$ and we denote the fit as $\hat{\mu}^N = \left(\hat{\mu}_1^N, \hat{\mu}_2^N, \ldots, \hat{\mu}_{T_1}^N\right)'$, where $\hat{\mu}_t^N = (1, x_{(1),t}, x_{(2),t}, \ldots, x_{(N),t})\hat{\boldsymbol{\gamma}}^N$ for $t = 1, 2, .., T_1$, $\hat{\boldsymbol{\gamma}}^N = (\hat{\gamma}_0^N, \hat{\gamma}_{(1)}^N, \hat{\gamma}_{(2)}^N, \ldots, \hat{\gamma}_{(N)}^N)'$ denotes the OLS estimator obtained from regressing y_t on $(1, x_{(1),t}, x_{(2),t}, \ldots, x_{(N),t})$ with $t = 1, \ldots, T_1$. Thus, we obtain N candidate models' fits with the corresponding set of estimates denoted as $\left\{\hat{\mu}^1, \hat{\mu}^2, \ldots, \hat{\mu}^N\right\}$. Let $\mathbf{w} = \left(w^1, w^2, \ldots, w^N\right)'$ be a set of weights which are non-negative and $\sum_{i=1}^N w^i = 1$. Given $\mathbf{w}$, a model averaging estimator could be formulated as

$$\hat{\mu}(\mathbf{w}) = \sum_{i=1}^{N} w^i \hat{\mu}^i = \hat{\mu}\,\mathbf{w}, \tag{5}$$

where $\hat{\mu} = (\hat{\mu}^1, \hat{\mu}^2, \ldots, \hat{\mu}^N)$ is a $T_1 \times N$ matrix.

Hansen and Racine (2012) propose to select the empirical weights $\mathbf{w}$ in the Jackknife manner (leave-one-out cross-validation). For the ith model's estimate $\hat{\mu}^i$, we denote the Jackknife estimator as a $T_1 \times 1$ vector $\tilde{\mu}^i = \left(\tilde{\mu}_{1,-1}^i, \tilde{\mu}_{2,-2}^i, \ldots, \tilde{\mu}_{T_1,-T_1}^i\right)'$, where $\tilde{\mu}_{k,-k}^i$ represents the estimator of $\hat{\mu}_k^i$ with the kth observation deleted. Thus, the Jackknife residual for the ith model, written in a $T_1 \times 1$ vector form, is $\tilde{\mathbf{e}}^i = \mathbf{y} - \tilde{\mu}^i$, where $\mathbf{y} = \left(y_1, y_2, \ldots, y_{T_1}\right)'$. We further define the Jackknife version of the averaging estimator as

$$\tilde{\mu}(\mathbf{w}) = \sum_{i=1}^{N} w^i \tilde{\mu}^i = \tilde{\mu}\,\mathbf{w},$$

where $\tilde{\mu} = \left(\tilde{\mu}^1, \tilde{\mu}^2, \ldots, \tilde{\mu}^N\right)$ is a $T_1 \times N$ matrix, and the Jackknife residuals as

$$\begin{aligned}\tilde{\mathbf{e}}(\mathbf{w}) &= \mathbf{y} - \tilde{\mu}(\mathbf{w}) = \sum_{i=1}^{N} w^i \mathbf{y} - \sum_{i=1}^{N} w^i \tilde{\mu}^i = \sum_{i=1}^{N} w^i \left(\mathbf{y} - \tilde{\mu}^i\right) \\ &= \sum_{i=1}^{N} w^i \tilde{\mathbf{e}}^i = \tilde{\mathbf{e}}\,\mathbf{w},\end{aligned}$$

where $\tilde{\mathbf{e}} = \left(\tilde{\mathbf{e}}^1, \tilde{\mathbf{e}}^2, \ldots, \tilde{\mathbf{e}}^N\right)$ is a $T_1 \times N$ matrix. Then, the least square cross-validation criterion function could be written as

$$CV_{T_1}(\mathbf{w}) = \frac{1}{T_1}\,\tilde{\mathbf{e}}(\mathbf{w})'\,\tilde{\mathbf{e}}(\mathbf{w}) = \mathbf{w}'\,\mathbf{S}_{T_1}\,\mathbf{w}, \tag{6}$$

[1] Under Hsiao et al. (2012), there are $N+1$ units $y_1, y_2, \ldots, y_{N+1}$. The first identity is assumed to be the treatment unit and the remained N are control units. So under our notations, $y = y_1, x_1 = y_2, x_2 = y_3, \ldots, x_N = y_{N+1}$.

Table 1
Comparison of PMSE between model average method (MA) and HCW method (AICC and AIC): 1 factor.

	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
	MA	AICC	AIC	MA	AICC	AIC	MA	AICC	AIC
$T_1 = 25, T = 35$									
Avg. #	–	3.56	10.92	–	3.63	11.14	–	3.66	10.95
PMSE	1.8404	2.0967	6.0158	0.9199	1.0540	3.1426	0.1838	0.2101	0.7461
$T_1 = 40, T = 50$									
Avg. #	–	3.71	5.91	–	3.75	5.94	–	3.76	5.97
PMSE	1.3841	1.7310	1.8935	0.6919	0.8635	0.9643	0.1383	0.1728	0.1915
$T_1 = 60, T = 70$									
Avg. #	–	4.14	5.60	–	4.20	5.37	–	4.19	5.39
PMSE	1.2553	1.4180	1.4722	0.6282	0.7258	0.7473	0.1258	0.1453	0.1496

Table 2
Comparison of PMSE between model average method (MA) and HCW method (AICC and AIC): 2 factors.

	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
	MA	AICC	AIC	MA	AICC	AIC	MA	AICC	AIC
$T_1 = 25, T = 35$									
Avg. #	–	4.16	11.09	–	4.11	10.97	–	4.24	11.15
PMSE	2.0749	2.4358	6.2314	1.0626	1.1658	2.7615	0.2195	0.2372	0.6337
$T_1 = 40, T = 50$									
Avg. #	–	3.94	6.25	–	4.05	6.28	–	3.98	6.35
PMSE	1.6196	1.8450	1.9634	0.8160	0.9380	1.0153	0.1639	0.1865	0.1970
$T_1 = 60, T = 70$									
Avg. #	–	4.35	5.58	–	4.50	5.78	–	4.54	5.83
PMSE	1.3829	1.6194	1.6502	0.6962	0.7904	0.7986	0.1400	0.1578	0.1605

where $\mathbf{S}_{T_1} = \frac{1}{T_1} \tilde{\mathbf{e}}' \tilde{\mathbf{e}}$ is a $N \times N$ matrix. The Jackknife weight $\hat{\mathbf{w}}$ is the value that minimizes (6) under the restrictions that each weight is between 0 and 1 and their summation equals to 1. Since Eq. (6) is quadratic in $\mathbf{w}$, we could get $\hat{\mathbf{w}}$ by applying the standard quadratic programming technique which requires short computing time. With the selected weight $\hat{\mathbf{w}}$ above, the Jackknife model average (JMA) estimator of μ could be written as $\hat{\mu}(\hat{\mathbf{w}}) = \hat{\mu}\hat{\mathbf{w}}$, and our model averaging estimation of the treatment effect defined in (4) could be calculated through $\hat{y}_t^0 = \hat{\mu}(\hat{\mathbf{w}}) = \sum_{i=1}^N \hat{w}^i \ \hat{\mu}^i$ for $t = T_1 + 1, \ldots, T$.

Let $\mu_t = E(y_t | x_t)$ denotes the conditional mean of y_t given x_t, $t = 1, 2, \ldots, T_1$, so that $\mu = (\mu_1, \mu_2, \ldots, \mu_{T_1})'$ is a $T_1 \times 1$ vector. Define $L_{T_1}(\mathbf{w}) = \frac{1}{T_1} \left(\mu - \hat{\mu}(\mathbf{w})\right)' \left(\mu - \hat{\mu}(\mathbf{w})\right)$ as a measure for the fit of the pre-treatment period, and $R_{T_2}(\mathbf{w}) = E\left(\frac{1}{T_2} \left(\mathbf{y}^0 - \hat{\mu}^0 \, \mathbf{w}\right)' \left(\mathbf{y}^0 - \hat{\mu}^0 \, \mathbf{w}\right)\right)$ as a measure for the prediction errors of the counterfactual outcomes for the post-treatment period, where $\mathbf{y}^0 = (y_{T_1+1}^0, y_{T_1+2}^0, \ldots, y_T^0)'$, $\hat{\mu}^0 = \left(\hat{\mu}^{0,1}, \hat{\mu}^{0,2}, \ldots, \hat{\mu}^{0,N}\right)$ is a $T_2 \times N$ matrix with $\hat{\mu}^{0,j} = \left(\hat{\mu}_{T_1+1}^{0,j}, \hat{\mu}_{T_1+2}^{0,j}, \ldots, \hat{\mu}_T^{0,j}\right)'$ as a $T_2 \times 1$ vector, where for $j = 1, \ldots, N$, $\hat{\mu}_t^{0,j} = \left(1, x_{(1),t}, \ldots, x_{(j),t}\right) \hat{\boldsymbol{\gamma}}^j$ (for $t = T_1 + 1, T_1 + 2, \ldots, T$) with $\hat{\boldsymbol{\gamma}}^j$ being the OLS estimate of the coefficient vector from regressing y_t on $\left(1, x_{(1),t}, \ldots, x_{(j),t}\right)$ using the pre-treatment period data $t = 1, \ldots, T_1$. Thus, $\hat{\mu}^0$ represents the out-sample predictions of the counterfactual outcome from the N candidate models. Therefore, $\hat{\mu}^0\mathbf{w}$ is (a $T_2 \times 1$ vector) the model averaging estimate of the counterfactual outcome $\mathbf{y}^0 = (y_{T_1+1}^0, y_{T_1+2}^0, \ldots, y_T^0)'$. Under mild assumptions, Hansen and Racine (2012) prove the asymptotic optimality of the Jackknife selected weights by showing the following:

$$\frac{L_{T_1}(\hat{\mathbf{w}})}{\inf\limits_{w \in D_w} L_{T_1}(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T_1 \to \infty, \tag{7}$$

where $D_w = \{w \in \mathcal{R}_+^N \mid \sum_{j=1}^N w_j = 1\}$. Eq. (7) indicates that the mean of squared estimation loss of the Jackknife model average (JMA) estimator is asymptotically identical to that of the infeasible best possible model averaging estimator. Let $\mu^0 = (\mu_{T_1+1}, \mu_{T_1+2}, \ldots, \mu_{T_1+T_2})'$ so $R_{T_2}(\mathbf{w}) = E\left(\frac{1}{T_2} \left(\mu^0 - \hat{\mu}^0 \, \mathbf{w}\right)' \left(\mu^0 - \hat{\mu}^0 \, \mathbf{w}\right)\right) + \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} E(v_t^2)$, where the second term is unrelated to $\mathbf{w}$. Hansen (2008) proves $E\left(\frac{1}{T_2} \left(\mu^0 - \hat{\mu}^0 \, \mathbf{w}\right)' \left(\mu^0 - \hat{\mu}^0 \, \mathbf{w}\right)\right) = E\left(L_{T_1}(\mathbf{w})\right) (1 + o_p(1)) \approx E\left(L_{T_1}(\mathbf{w})\right)$, which together with the asymptotic optimality shown by (7) prompts us to use the JMA estimator for out-sample prediction in the current paper.

3. Numerical studies

We compare the performance of the Jackknife model average method with the method in Hsiao et al. (2012) through Monte Carlo simulations. For comparing purposes, we generally follow the settings of their work: we generate the model with 1 treatment unit and 20 control units ($N = 20$), and the pre-treatment period $T_1 = 25, 40$ and 60. The post-treatment period includes $T - T_1 = 10$ observations. As in Hsiao et al. (2012), we consider 1-factor, 2-factor and 3-factor structures as the following:

1-factor:

$$f_{1t} = 0.95 f_{1t-1} + \epsilon_{1t},$$

2-factor:

$$f_{1t} = 0.3 f_{1t-1} + \epsilon_{1t},$$
$$f_{2t} = 0.6 f_{2t-1} + \epsilon_{2t},$$

3-factor:

$$f_{1t} = 0.8 f_{1t-1} + \epsilon_{1t},$$
$$f_{2t} = -0.6 f_{2t-1} + \epsilon_{2t} + 0.8\epsilon_{2t-1},$$
$$f_{3t} = \epsilon_{3t} + 0.9\epsilon_{3t-1} + 0.4\epsilon_{3t-2},$$

where ϵ_{it} is distributed to $N(0, 1)$, $i = 1, 2, 3$. u_{it} in (1) is generated from $N(0, \sigma^2)$ where $\sigma^2 = 1$, 0.5 and 0.1, and B is generated from $N(1, 1)$. We examine the performance of our model average

Table 3
Comparison of PMSE between model average method (MA) and HCW method (AICC and AIC): 3 factors.

	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
	MA	AICC	AIC	MA	AICC	AIC	MA	AICC	AIC
$T_1 = 25, T = 35$									
Avg. #	–	4.40	11.18	–	4.54	11.70	–	4.74	11.72
PMSE	2.4758	2.8493	6.5231	1.2956	1.4640	3.2716	0.2732	0.3051	0.7024
$T_1 = 40, T = 50$									
Avg. #	–	4.90	7.10	–	5.08	7.22	–	5.04	7.24
PMSE	1.7900	1.9705	2.1136	0.9051	0.9716	1.0804	0.1829	0.1987	0.2095
$T_1 = 60, T = 70$									
Avg. #	–	5.12	6.34	–	5.18	6.36	–	5.24	6.42
PMSE	1.5205	1.6437	1.6842	0.7659	0.8302	0.8516	0.1542	0.1677	0.1685

method with Hsiao et al. (2012) by comparing the post-treatment mean squared prediction errors (PMSE), which is defined as

$$PMSE = \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} (y_t^0 - \hat{y}_t^0)^2,$$

where $\hat{y}_t^0$ is the estimated counterfactual outcome by using the AIC method, or the AICC method, or our proposed Jackknife model average method.

We repeat each of the structures 1000 times. The results are displayed in Tables 1–3. The Avg. # is the average number of control units selected by the AIC or the AICC methods. Simulation results show that our method has smaller PMSE in all cases, indicating improved predicting performance by replacing the two-step model selection strategy with the JMA method.

4. Conclusion

In this paper we suggest to replace the two-step model selection strategy in Hsiao et al. (2012) with the Jackknife model average (JMA) method to estimate average treatment effect of a program or a policy. By applying the JMA, we show that the post-treatment predicting performance improves in terms of prediction mean squared error.

Acknowledgment

Zhengwei Wang's work is partially supported by the National Science Foundation of China (Project No. 71472100).

References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22 (1), 203–217.
Bai, C., Li, Q., Ouyang, M., 2014. Property taxes and home prices: A tale of two cities. J. Econometrics 180 (1), 1–15.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.
Buckland, S., Burnham, K., Augustin, N., 1997. Model selection: an integral part of inference. Biometrics 53 (2), 603–618.
Hansen, B., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.
Hansen, B., 2008. Least-squares forecast averaging. J. Econometrics 146 (2), 342–350.
Hansen, B., Racine, J., 2012. Jackknife model averaging. J. Econometrics 167 (1), 38–46.
Hoeting, J., Madigan, D., Raftery, A., Volinsky, C., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 14 (4), 382–417.
Hsiao, C., Ching, H., Wan, K., 2012. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. J. Appl. Econometrics 27 (5), 705–740.
Hurvich, C., Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.
Wan, A.T.K., Zhang, X., Zou, G., 2010. Least squares model average by Mallows criterion. J. Econometrics 156 (2), 277–283.