Economics Letters 233 (2023) 111369

Contents lists available at ScienceDirect

Economics Letters

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

Time series cross validation: A theoretical result and finite sample performance☆

Ai Deng

Charles River Associates and Johns Hopkins University, United States of America

ARTICLE INFO

Keywords:
Cross-validation
Model selection
Validation sample

ABSTRACT

We provide a theoretical result for the time series-based cross validation that sheds light on the choice of validation sample size. We also consider an alternative way to construct validation samples and demonstrate the improved performance in certain situations via simulations.

1. Introduction

As an important model selection technique, cross-validation (hereinafter CV) has a long history in the statistical literature. Most early research focused on the theory with i.i.d. observations and offered little direct guidance on how to handle data dependence, a common feature of economic time series. Racine (2000) filled this gap by proposing the hv-block CV. Since then, there have been a growing number of studies of time-series based CV. See, for example, Brownlees and Gallo (2011) and Bergmeir et al. (2018), among others.

The sample-splitting scheme of time series CV such as Racine's hv-block CV is different from the conventional leave-n-out or k-fold CV, as the validation set of time series-based CV typically consists of a sequence of moving windows of consecutive observations. As such, it can be considered a generalization of the popular pseudo out-of-sample (OOS) evaluation. This paper studies the theoretical and empirical properties of such procedures in linear models. Our theoretical result highlights an interesting implication of the size of the validation samples on model selection performance. We also consider an alternative way to construct validation samples and show via simulations that it can improve finite sample performance. We also find that the standard BIC often performs the best for the data generating process (DGP) we consider in this paper. It is worth noting that although Racine (2000) conjectured that his hv-block CV is model-selection consistent, we show in the Appendix that consistency of v-block CV (hence the hv-block) does not follow from his conjectured proofs.

It is important from the outset to emphasize that we assume that the true model is of fixed dimensions and that the dimension of the predictors does not increase with the sample size. And we evaluate the performance based on model selection accuracy, instead of predictive accuracy.[1] This is the same framework considered by Shao (1993) and Racine (2000).[2]

2. Time series CV

As noted above, a feature of time series-based CV is that the validation sample consists of consecutive observations. In Racine's terminology, for a given value v, the following would be the validation samples for a time series of size n,

$$(1, 2, \dots, v, v+1, v+2, \dots, 2v+1);$$
$$(2, 3, \dots, v+1, v+2, v+3, \dots, 2v+2);$$
$$(3, 4, \dots, v+2, v+3, v+4, \dots, 2v+3);$$
$$\dots$$
$$(n-2v, n-2v+1, \dots, n-v, n-v+1, n-v+2, \dots, n).$$

Therefore, each validation sample is of size $n_v \equiv 2v+1$. We call this the v-block CV. One could also remove h observations on both sides of the validation samples from the estimation samples to mitigate "look-ahead" bias given the serial dependence in time series data. This is the idea behind Racine's hv-block CV. Although not considered in our theory, we allow non-zero h in the simulations. For later references, we denote $n_c = n - n_v$.

☆ I thank Hal White and Jeff Racine for their comments on early versions of the article and Mike Packard for excellent research assistance. I also thank an anonymous referee for constructive suggestions. The views expressed and any errors are mine only.

E-mail address: adeng@crai.com.

[1] We report some limited simulation results to assess predictive performance in an online appendix.

[2] see Li (1987) and Shao (1997) for alternative asymptotic frameworks.

https://doi.org/10.1016/j.econlet.2023.111369
Received 14 January 2023; Received in revised form 2 August 2023; Accepted 20 September 2023
Available online 28 September 2023

3. Theory

We consider the simple DGP in standard notations $y = x'\beta + e$ of n observations.[3] Let $\alpha \in N^{p_\alpha}$ denote a subset of $\{1, \dots, p\}$ of size p_α, and x_α be the submatrix of x containing variables indexed by the integers in α. Let "model α", denoted as $\mathcal{M}_\alpha$, be given by $y = x'_\alpha \beta_\alpha + e_\alpha$, one of wihch is the DGP. Following Shao (1993), consider two categories of models $\mathcal{M}_\alpha$:

- I: At least one nonzero element of β is not in β_α
- II: β_α contains all nonzero elements of β

Theorem 1 provides a characterization of v-block CV. Let P_α denote the projection matrix, and r_α the full-sample regression residual vector, i.e., $r_\alpha = y - X_\alpha \hat{\beta}_\alpha$ for the full sample least-squares estimate $\hat{\beta}_\alpha$. We define matrix Γ whose ith diagonal element is the number of validation samples in which the ith observation appears and whose $(i,j)th$ off-diagonal element is the number of validation samples in which the $(i,j)th$ pair of observations appears and let $\widetilde{\Pi} = \frac{1}{n_c+1} diag(\Gamma)$. $\odot$ denotes the Hadamard product. $CV^{VB}_{\alpha,n}$ denotes the cross-validated mean squared errors (CVMSE) of model α. The Appendix contains the formal definition. Model selection is based on minimizing the CVMSE. Finally, we consider the same asymptotics as in Racine (2000), i.e., $n_v/n \to 1$ and $n_c = n - n_v \to \infty$.

Theorem 1. *Suppose that Assumptions 1–3 in the Appendix hold and that* $n_v/n \to 1$ *and* $n_c = n - n_v \to \infty$.

For models in Category I, then there exists $R_n \geq 0$ *such that*

$$CV^{VB}_{\alpha,n} = n_v^{-1} e' \widetilde{\Pi} e + \Delta_{\alpha,n} + o_p(1) + R_n$$

where $\Delta_{\alpha,n}$ *is defined in assumption 1 in the Appendix. It suffices to know that under Assumption 1,* $p \liminf_{n\to\infty} \Delta_{\alpha,n} > 0$ *for models in Category I.*

For models in Category II,

$$CV^{VB}_{\alpha,n} = n_v^{-1} e' \widetilde{\Pi} e + A_{\alpha 2} + o_p(A_{\alpha 2}),$$

where

$$A_{\alpha 2} = \frac{n + n_c}{n_c^2} \left[r'_\alpha \left(P_\alpha \odot \frac{\Gamma}{n_c + 1} \right) r_\alpha \right]$$

The proof of Theorem 1 can be found in the technical appendix and is based on an extension of Shao (1993). It involves analyzing CVMSEs for models in Category I and Category II separately and making careful use of the number of times a given observation is used for validation. The result decomposes CVMSE into several components, the first of which is common across all candidate models and therefore can be ignored. This decomposition reveals a practical insight: under the assumptions, all else equal, the larger the n_c in a finite sample, the smaller the probability of selecting a Category I model will be. This is because the smaller $A_{\alpha 2}$ is, the more likely that $\Delta_{\alpha,n} + o_p(1) + R_n > A_{\alpha 2} + o_p(A_{\alpha 2})$ and hence the more likely that the cross-validation function $CV^{VB}_{\alpha,n}$ takes a larger value for Category I models. Especially when the candidate models are nested as in autoregressive order selection, this means that choices of n_c can reflect asymmetric loss functions: if the loss from under-selection (Category I) is greater than the loss from over-selection, then a larger n_c may be desirable.[4] In practice, a common choice for n_c that satisfies the asymptotic conditions is $[n^\delta]$ where $0 < \delta < 1$. Simulations below confirm this theoretical insight and reveal that, perhaps unsurprisingly, the performance of the CV procedures depends on the interaction between the DGP and the choice of δ.

[3] Additional technical assumptions are found in the Appendix.

[4] Doing so optimally is outside the scope of this article.

4. Unequal validation samples

The CV scheme described above ensures that the validation samples are always of size n_v. By relaxing this condition, there are more validation samples. For example, suppose $v = 2$ (hence $n_v = 2\times 2+1 = 5$), then additional validation samples include $(1,2,3)$, $(1,2,3,4)$, $(T-3, T-2, T-1, T)$, and $(T-2, T-1, T)$. Intuitively, using these additional samples may improve the finite sample performance. Indeed, our simulations confirm such improvement in certain situations.[5]

5. Simulations

We use simulations to examine (1) how alternative validation sets affect the performance of the CV procedures and (2) how these CV methods compare with BIC. Specifically, we compare the performance of the following methods

- hv-block CV with equal validation set size (hv-block Equal, Racine's original hv-block CV);
- hv-block CV with unequal validation set sizes (hv-block Unequal);
- v-block CV with equal validation set size (v-block Equal);
- v-block CV with unequal validation set sizes (v-block Unequal);
- The standard BIC.

For comparison purposes, we follow Racine (2000) by setting $h = 0.25n$ in use with the hv-block CV. To assess the impact of the validation sample size on model selection performance that we highlighted above, we consider two choices of δ, 0.5 and 0.75.

We report the findings of an autoregressive order selection exercise.[6] This was also studied by Racine (2000) and more recently Bergmeir et al. (2018). Specifically, the DGP is an $AR(3)$ given by $y_i = \beta_1 y_{i-1} + \beta_2 y_{i-2} + \beta_3 y_{i-3} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.5$. We consider two sets of coefficients.

- $(\beta_1, \beta_2, \beta_3) = (0.7, 0.15, 0.1)$.
- $(\beta_1, \beta_2, \beta_3) = (0.4, 0.3, 0.25)$.

Both DGPs are nearly integrated with a characteristic root around 0.97. Note that in the first set of coefficients, the first lag accounts for the most serial dependence, in which case a parsimonious model may be a reasonable approximation. The numbers are more evenly defined in the second set of coefficients. We consider five candidate models, AR(1) through AR(5), and a range of sample sizes with 2000 replications. Albeit the simple setup, such a comparison is not available in the literature to the author's knowledge and allows us to examine effect of validation sample and n_c in a simple framework.

5.1. Finite sample performance

Table 1a corresponds to $(\beta_1, \beta_2, \beta_3) = (0.7, 0.15, 0.1)$ and $\delta = 0.5$ and shows that the use of unequal validation samples alone can improve the performance. In fact, it makes the performance of v-block CV comparable and even slightly better than the original hv-block CV. BIC does not perform as well in smaller samples but becomes better as the sample size increases.

In Table 1b ($\delta = 0.75$), some of the observations made about Table 1a are still true. The most notable difference is that using $\delta = 0.75$ significantly improves the performance of all CV procedures which further outperform BIC by a large margin as long as $n < 1000$. This can be explained by two observations: first is the theoretical insight that the larger δ makes the probability of under-selection smaller and second, the fact that given the specific DGP, BIC turns out to be too conservative.

[5] There are other ways to split the sample. The scheme considered here is not based on any theoretical optimality argument.

[6] We also examined the performance in a static model mimicking (Racine, 2000). Results are available in an extended Appendix.

Table 1a
Autoregressive order selection (0.7, 0.15, 0.1). $\delta = 0.5$.

n	AR(1)	AR(2)	**AR(3)**	AR(4)	AR(5)
v-block	Unequal				
100	0.868	0.121	0.010	0.001	0.001
200	0.491	0.299	0.171	0.029	0.011
500	0.100	0.395	0.405	0.072	0.029
1000	0.007	0.298	0.578	0.080	0.038
hv-block	Unequal				
100	0.837	0.144	0.016	0.003	0.001
200	0.453	0.313	0.193	0.032	0.010
500	0.084	0.389	0.413	0.080	0.035
1000	0.005	0.282	0.582	0.092	0.041
v-block	Equal				
100	0.866	0.107	0.017	0.007	0.005
200	0.708	0.182	0.083	0.020	0.008
500	0.432	0.322	0.170	0.056	0.022
1000	0.231	0.435	0.234	0.064	0.037
hv-block	Equal				
100	0.856	0.127	0.014	0.002	0.001
200	0.502	0.287	0.171	0.031	0.010
500	0.131	0.383	0.375	0.077	0.035
1000	0.016	0.333	0.520	0.086	0.046
BIC					
100	0.588	0.345	0.059	0.008	0.002
200	0.240	0.615	0.134	0.011	0.001
500	0.003	0.599	0.393	0.005	0.001
1000	0.000	0.313	0.681	0.006	0.001

Table 1b
Autoregressive order selection (0.7, 0.15, 0.1). $\delta = 0.75$.

n	AR(1)	AR(2)	**AR(3)**	AR(4)	AR(5)
v-block	Unequal				
100	0.261	0.403	0.249	0.058	0.030
200	0.072	0.427	0.362	0.090	0.050
500	0.003	0.280	0.575	0.103	0.040
1000	0.000	0.121	0.759	0.083	0.038
hv-block	Unequal				
100	0.219	0.359	0.280	0.086	0.057
200	0.072	0.375	0.365	0.109	0.080
500	0.004	0.268	0.547	0.121	0.061
1000	0.000	0.131	0.679	0.122	0.069
v-block	Equal				
100	0.343	0.400	0.191	0.044	0.024
200	0.118	0.476	0.294	0.076	0.038
500	0.011	0.364	0.494	0.089	0.043
1000	0.000	0.206	0.678	0.084	0.033
hv-block	Equal				
100	0.233	0.355	0.271	0.086	0.057
200	0.085	0.381	0.345	0.110	0.080
500	0.005	0.281	0.529	0.124	0.062
1000	0.000	0.151	0.670	0.123	0.057
BIC					
100	0.588	0.345	0.059	0.008	0.002
200	0.240	0.615	0.134	0.011	0.001
500	0.003	0.599	0.393	0.005	0.001
1000	0.000	0.313	0.681	0.006	0.001

Table 2a
Autoregressive order selection (0.4, 0.3, 0.25). $\delta = 0.5$.

n	AR(1)	AR(2)	**AR(3)**	AR(4)	AR(5)
v-block	Unequal				
100	0.720	0.224	0.054	0.003	0.000
200	0.132	0.285	0.492	0.071	0.021
500	0.003	0.064	0.736	0.130	0.068
1000	0.001	0.004	0.784	0.149	0.063
hv-block	Unequal				
100	0.663	0.255	0.076	0.007	0.001
200	0.121	0.268	0.509	0.078	0.025
500	0.003	0.057	0.734	0.131	0.077
1000	0.000	0.005	0.779	0.151	0.066
v-block	Equal				
100	0.852	0.118	0.021	0.006	0.004
200	0.519	0.238	0.184	0.045	0.015
500	0.095	0.333	0.441	0.086	0.046
1000	0.017	0.179	0.618	0.129	0.058
hv-block	Equal				
100	0.727	0.212	0.057	0.005	0.001
200	0.171	0.281	0.456	0.071	0.023
500	0.005	0.083	0.702	0.134	0.077
1000	0.001	0.006	0.762	0.155	0.077
BIC					
100	0.027	0.416	0.526	0.029	0.003
200	0.000	0.145	0.826	0.027	0.003
500	0.000	0.002	0.989	0.008	0.001
1000	0.000	0.000	0.991	0.009	0.000

Table 2b
Autoregressive order selection (0.4, 0.3, 0.25). $\delta = 0.75$.

n	AR(1)	AR(2)	**AR(3)**	AR(4)	AR(5)
v-block	Unequal				
100	0.013	0.233	0.583	0.107	0.066
200	0.001	0.069	0.737	0.126	0.068
500	0.000	0.001	0.810	0.132	0.058
1000	0.000	0.000	0.861	0.102	0.037
hv-block	Unequal				
100	0.014	0.200	0.563	0.133	0.091
200	0.001	0.070	0.671	0.167	0.093
500	0.000	0.006	0.733	0.170	0.092
1000	0.000	0.000	0.772	0.153	0.076
v-block	Equal				
100	0.036	0.329	0.503	0.084	0.049
200	0.001	0.119	0.707	0.119	0.055
500	0.000	0.007	0.806	0.125	0.062
1000	0.000	0.001	0.841	0.119	0.040
hv-block	Equal				
100	0.019	0.217	0.538	0.131	0.095
200	0.001	0.081	0.658	0.168	0.093
500	0.000	0.008	0.725	0.178	0.090
1000	0.000	0.000	0.781	0.149	0.071
BIC					
100	0.027	0.416	0.526	0.029	0.003
200	0.000	0.145	0.826	0.027	0.003
500	0.000	0.002	0.989	0.008	0.001
1000	0.000	0.000	0.991	0.009	0.000

The DGP for Tables 2a ($\delta = 0.5$) and 2b ($\delta = 0.75$) correspond to the second set of coefficients. The results in Table 2a shows a similar pattern where the use of unequal validation samples improves the performance of both hv-block CV and v-block CV. However, for this DGP, BIC selects the correct model much more frequently than any of the CV variants. Using $\delta = 0.75$ again improves the performance of the CV procedures especially in smaller samples, making it comparable to BIC. In larger samples ($n = 500$ or 1000), BIC continues to outperform.

6. Conclusions

We show that the use of unequal validation sets can indeed perform better than the equal validation sets originally proposed and the choice of estimation sample size n_c has a predictable effect on the model selection performance in finite sample. We also find that BIC often outperforms these time series CV methods, especially in relatively large samples.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2023.111369.

References

Li, K.-C., 1987. Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set. Ann. Statist. 15, 958–975.

Shao, J., 1997. An asymptotic theory for linear model selection (with discussions). Statist. Sinica 7, 221–264.

Racine, J., 2000. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. J. Econometrics 99, 39–61.

Brownlees, C.T., Gallo, G., 2011. Shrinkage estimation of semiparametric multiplicative error models. Int. J. Forecast. 27, 365–378.

Bergmeir, C., Hyndman, R.J., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput. Statist. Data Anal. 120, 70–83.

Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88 (422), 86–494.