Computational Statistics and Data Analysis 120 (2018) 70–83

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A note on the validity of cross-validation for evaluating autoregressive time series prediction

Christoph Bergmeir [a,*], Rob J. Hyndman [b], Bonsoo Koo [b]

[a] Faculty of Information Technology, Monash University, Melbourne, Australia
[b] Department of Econometrics & Business Statistics, Monash University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 10 June 2016
Received in revised form 30 October 2017
Accepted 3 November 2017
Available online 22 November 2017

Keywords:
Cross-validation
Time series
Autoregression

ABSTRACT

One of the most widely used standard procedures for model evaluation in classification and regression is *K*-fold cross-validation (CV). However, when it comes to time series forecasting, because of the inherent serial correlation and potential non-stationarity of the data, its application is not straightforward and often replaced by practitioners in favour of an out-of-sample (OOS) evaluation. It is shown that for purely autoregressive models, the use of standard *K*-fold CV is possible provided the models considered have uncorrelated errors. Such a setup occurs, for example, when the models nest a more appropriate model. This is very common when Machine Learning methods are used for prediction, and where CV can control for overfitting the data. Theoretical insights supporting these arguments are presented, along with a simulation study and a real-world example. It is shown empirically that *K*-fold CV performs favourably compared to both OOS evaluation and other time-series-specific techniques such as non-dependent cross-validation.

1. Introduction

Cross-validation (CV) (Stone, 1974; Arlot and Celisse, 2010) is one of the most widely used methods to assess the generalizability of algorithms in classification and regression (Hastie et al., 2009), and is subject to ongoing active research (e.g., Budka and Gabrys, 2013; Borra and Di Ciaccio, 2010; Bergmeir et al., 2014; Moreno-Torres et al., 2012). However, when it comes to time series prediction, practitioners are often unsure of the best way to evaluate their models. There is often a feeling that we should not be using future data to predict the past. In addition, the serial correlation in the data, along with possible non-stationarities, make the use of CV appear problematic as it does not account for these issues (Bergmeir and Benítez, 2012). Practitioners usually resort to out-of-sample (OOS) evaluation instead, where a section from the end of the series is withheld for evaluation. In this way, only one evaluation on a test set is considered, whereas with the use of cross-validation, various such evaluations are performed. So, by using OOS, the benefits of CV, especially for small datasets, cannot be exploited. One important part of the problem is that in the traditional forecasting literature, OOS evaluation is the standard evaluation procedure, partly because fitting of standard models such as exponential smoothing (Hyndman et al., 2008) or ARIMA models are fully iterative in the sense that they start estimation at the beginning of the series. In addition, some research has demonstrated cases where standard CV fails in a time series context. For example, Opsomer et al. (2001) show that standard CV underestimates bandwidths in a kernel estimator regression framework if autocorrelation of the error is high, so that the method overfits the data. As a result, several CV techniques have been developed especially for the

* Correspondence to: Faculty of Information Technology, P.O. Box 63 Monash University, Victoria 3800, Australia.
E-mail address: christoph.bergmeir@monash.edu (C. Bergmeir).

https://doi.org/10.1016/j.csda.2017.11.003
0167-9473/

dependent case (Györfi et al., 1989; Burman and Nolan, 1992; Burman et al., 1994; McQuarrie and Tsai, 1998; Racine, 2000; Kunst, 2008).

This study addresses the problem in the following way. When purely (non-linear, non-parametric) autoregressive methods are applied to forecasting problems, as is often the case (e.g., when using Machine Learning methods), the aforementioned problems of CV are largely irrelevant, and CV can and should be used without modification, as in the independent case. To the best of our knowledge, this is the first paper to justify the use of the standard K-fold CV in the dependent setting without modification. Our paper draws the line of applicability of the procedure between models that have uncorrelated errors, and models that are heavily misspecified and thereby do not produce uncorrelated errors. In practice, this means that the method can be used without problems to detect overfitting, as overfitted models have uncorrelated errors. Underfitting should be tackled beforehand, e.g. by testing residuals for serial correlation. We provide a theoretical proof and additional results of simulation experiments and a real-world example to justify our argument.

2. Cross-validation for the dependent case

Cross-validation for the dependent setting has been studied extensively in the literature, including Györfi et al. (1989), Burman and Nolan (1992) and Burman et al. (1994). Let $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ be a time series. Traditionally, when K-fold CV is performed, K randomly chosen numbers out of the vector $\boldsymbol{y}$ are removed. This removal invalidates the CV in the dependent setting because of the correlation between errors in the training and test sets. Therefore, Burman and Nolan (1992) suggest bias correction, whereas Burman et al. (1994) propose h-block CV whereby the h observations preceding and following the observation are left out in the test set. We call this procedure non-dependent cross-validation, as it leaves out the possibly dependent observations and only considers data points that can be considered to be independent.

However, both bias correction and the h-block CV method have their limitations including inefficient use of the available data.

Let us now consider a purely autoregressive model of order p

$$y_t = g(\boldsymbol{x}_t, \boldsymbol{\theta}) + \varepsilon_t, \tag{1}$$

where ε_t is a shock or disturbance term, $\boldsymbol{\theta}$ is a parameter vector, $\boldsymbol{x}_t \in \mathbb{R}^p$ consists of lagged values of y_t and $g(\boldsymbol{x}_t, \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[y_t|\boldsymbol{x}_t]$. Here $g(\cdot)$ could be a linear or nonlinear function, or even a nonparametric function. Thus, $g(\cdot)$ could be a totally unspecified function of the lagged values of y_t up to pth order.

Here, the lag order of the model is fixed and the time series is *embedded* accordingly, generating a matrix that is then used as the input for a (nonparametric, nonlinear) regression algorithm. The embedded time series with order p and a fixed forecast horizon of $h = 1$ is defined as follows:

$$\begin{bmatrix} y_1 & y_2 & \cdots & y_p & y_{p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{t-p} & y_{t-p+1} & \cdots & y_{t-1} & y_t \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n-p} & y_{n-p+1} & \cdots & y_{n-1} & y_n \end{bmatrix}. \tag{2}$$

Thus each row is of the form $[\boldsymbol{x}_t', y_t]$, and the first p columns of the matrix contain predictors for the last column of the matrix.

Recall the usual K-fold CV method, where the training data is partitioned into K separate sets, say $J = \{J_1, \ldots, J_K\}$. Define $J_{k^-} = \cup_{j \neq k} J_j$, so that for a particular evaluation k in the cross-validation, J_k is used as test set, and the remaining sets combined, J_{k^-}, are used for model fitting. Instead of reducing the training set by removing the h observations preceding and following the observation y_t of the test set, we leave out the entire set of rows corresponding to $t \in J_k$ in matrix (2). Fig. 1 illustrates the procedure.

Provided (1) is true, the rows of the matrix (2) are conditionally uncorrelated because $y_t - \hat{g}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}) = \hat{\varepsilon}_t$ is simply a regression error, and $\{\hat{\varepsilon}_t\}$ are asymptotically independent and identically distributed (i.i.d.) provided g is estimated appropriately. Consequently, omitting rows of the matrix will not affect the bias or consistency of the estimates.

In practice, although we do not know the correct p and other hyper-parameters of the model, if our model is sufficiently large and flexible (and therefore liable to be overfitted), errors will be uncorrelated. In the case of correlated errors, the CV procedure will be biased, but this can be relatively easily tackled by testing the residuals for serial correlation.

It is worth mentioning that this method leaves the entire row related to the chosen test set out instead of test set components only. As a result, we lose much less information embedded in the data in this way than in the h-block CV.

3. The theoretical result

Let $y_1, \ldots, y_n$ be the observations from a stationary process where each y_t has distribution P on $\mathbb{R}$. We consider pure AR(p) models in order to discuss the validity of our method. Without loss of generality and for ease of notation, we focus on the leave-one-out CV because generalization to the K-fold CV is straightforward.