A CONSISTENT CROSS-VALIDATORY METHOD FOR DEPENDENT DATA: hv-BLOCK CROSS-VALIDATION

JEFF RACINE

Department of Economics
University of South Florida
4202 East Fowler Avenue
Tampa, Florida USA 33620

813-974-6555
(FAX) 813-974-6510
jracine@bsn01.bsn.usf.edu

Abstract. This paper considers the impact of Shao's (1993) recent results regarding the asymptotic inconsistency of model selection via leave-one-out cross-validation on h-block cross-validation, a cross-validatory method for dependent data proposed by Burman, Chow & Nolan (1994). It is shown that h-block cross-validation is inconsistent in the sense of Shao (1993) and therefore is not asymptotically optimal. A modification of the h-block method, dubbed 'hv-block' cross-validation, is proposed which is asymptotically optimal. The proposed approach is consistent for general stationary observations in the sense that the probability of selecting the model with the best predictive ability converges to 1 as the total number of observations approaches infinity. This extends existing results and yields a new approach which contains leave-one-out cross-validation, leave-n_v-out cross-validation, and h-block cross-validation as special cases. Applications are considered.

1. Introduction

In a recent series of articles (Shao (1993), Shao (1996)), Shao addresses the issue of model selection in a simple linear regression context for independent identically distributed *(iid)* errors, with extensions to nonlinear settings. In these articles the author demonstrates the result that virtually all existing data-driven methods of model selection are "asymptotically inconsistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \to \infty$" (Shao (1993, pg 486)) and "inconsistent in the sense that the probability of selecting the optimal subset of variables does not converge to 1 as $n \to \infty$" (Shao (1996, pg 655)). That is, virtually all data-driven methods of model selection

Date: June 9, 1997.
JEL Classification. C5, C51, C52.
Key words and phrases. Cross-Validation; Dependence; Model-Selection; Prediction.
I would like to thank but not implicate Prabir Burman and Jun Shao for their valuable comments and suggestions. Support from the Division of Sponsored Programs at the University of South Florida is gratefully acknowledged. All errors remain, of course, my own.

will select models which are not optimal in a mean square error sense regardless of the amount of data at hand.

This paper considers the impact of Shao's (1993) results on h-block cross-validation, a cross-validatory method for dependent data recently proposed by Burman et al. (1994) which has wide-spread potential application ranging from selection of the order of an autoregression for linear time series models to selection of the number of terms in nonlinear series approximations for general stationary processes.

Burman et al. (1994, pg 354, Remark 2) state in their paper that "Conditions for asymptotic optimality for this proposal remain open.". In this paper it is shown that h-block cross-validation is inconsistent and not asymptotically optimal in the sense of Shao (1993). A modification of the h-block method, dubbed 'hv-block' cross-validation, is proposed which is asymptotically optimal. This approach provides a general framework for cross-validation which contains leave-one-out cross-validation (Allen (1974), Stone (1974) Stone (1977), Geisser (1975), Whaba & Wold (1975)), leave-n_v-out cross-validation[1] (Breiman, Friedman, Olshen & Stone (1984, Ch. 3,8), Burman (1989), Burman (1990), Burman & Nolan (1992), Geisser (1975), Zhang (1993)), and h-block cross-validation (Burman et al. (1994), Racine (1997)) as special cases. Simulations are conducted to demonstrate the nature of the inconsistency and the value added by the proposed method. A modest application is conducted for determining the order of an autoregression for G7 exchange rate data.

2. Background

Following the notation of Shao (1993), consider a linear model of the form

$$
\begin{aligned}
y_i &= E[y|x_i] + \epsilon_i \\
&= x_i'\beta + \epsilon_i \quad i = 1, \ldots, n,
\end{aligned}
\tag{1}
$$

where y_i is the response, x_i is a $p \times 1$ vector of predictors, β is a $p \times 1$ vector of unknown parameters, ϵ_i is a mean zero disturbance term with constant variance σ^2, and n denotes the sample size.

Let $\alpha \in \mathbb{N}^{p_\alpha}$ denote a subset of $\{1, \ldots, p\}$ of size p_α and let $x_{i\alpha}$ be the subvector of x_i containing all components of x_i indexed by the integers in α. A model corresponding to α shall be called

[1] Similar variants include 'r-fold', 'v-fold', and 'multifold' cross-validation, and the 'repeated learning-testing criterion'.

'model α' for simplicity, and is given by

$$y_i = x'_{i\alpha}\beta_\alpha + \epsilon_i \quad i = 1, \ldots, n. \tag{2}$$

If the true data generating process is linear in that the conditional mean can be expressed as $E[y|x_i] = x_i'\beta$, then the problem of model/variable selection becomes one of finding the 'correct model', that is, the model α for which

$$y_i = E[y|x_i] = x'_{i\alpha}\beta_\alpha + \epsilon_i \quad i = 1, \ldots, n. \tag{3}$$

Starting from the set of predictors $x_i \in \mathbb{R}^p$ assumed to contain those for the correct model, $x_{i\alpha} \in \mathbb{R}^{p_\alpha}$, Shao (1993) considers all $2^p - 1$ possible models of the form (2) corresponding to a subset α and calls the class of such models $\mathcal{M}_\alpha$. Shao (1993) divides all models in this class into two categories:

1. Category I: Those for which a predictor belongs in the set of conditioning predictors for $E[y|x_i]$ but does not appear in $x_{i\alpha}$[2]. This case is sometimes referred to as the 'omission of relevant predictors'.
2. Category II: Those for which the set of conditioning predictors includes all relevant predictors and in addition may include predictors which do not belong in the set of conditioning predictors for $E[y|x_i]$ but appear in $x_{i\alpha}$[3] . This case is sometimes referred to as the 'inclusion of irrelevant predictors'.

Shao (1993, pg 487) defines the 'optimal model' $\mathcal{M}_*$ as the "model in Category II with the smallest dimension", and this optimal model $\mathcal{M}_*$ will possess the smallest expected prediction error of any model α. Cross-validation estimates the expected prediction error for a model α, and cross-validatory model selection proceeds by selecting that model α with smallest estimated expected prediction error. One of the most widely-used variants of cross-validation is leave-one-out cross-validation.

[2] "At least one non-zero component of β is not in β_α" (Shao (1993, pg 487)).
[3] "β_α contains all non-zero components of β" (Shao (1993, pg 487)).

The leave-one-out cross-validation function is given by

$$CV_1 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{i(-i)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}_{(-i)}\right)^2, \tag{4}$$

where $\hat{y}_{i(-i)}$ is the prediction of y_i when the ith observation is deleted from the data set, and where $\hat{\beta}_{(-i)}$ is the least squares estimator obtained by deleting the ith observation from the data set.

Letting $\mathcal{M}_{cv_1}$ denote a model selected via leave-one-out cross-validation and assuming that standard regularity conditions hold, Shao (1993, pg 488) demonstrates that

$$\lim_{n\to\infty} pr(\mathcal{M}_{cv_1} \text{ is in Category I}) = 0, \tag{5}$$

that is, asymptotically the probability of selecting a model which excludes a relevant predictor when the assumed set of predictors nests the relevant set of predictors is zero. Furthermore, he demonstrates that

$$\lim_{n\to\infty} pr(\mathcal{M}_{cv_1} \text{ is in Category II but is not } \mathcal{M}_*) > 0, \tag{6}$$

and therefore that

$$\lim_{n\to\infty} pr(\mathcal{M}_{cv_1} = \mathcal{M}_*) \neq 1, \tag{7}$$

that is, asymptotically the probability of including an irrelevant predictor is not equal to zero unless the dimension of the relevant set of predictors p_α is equal to the dimension of the full set of assumed predictors p. This result has the direct implication that leave-one-out cross-validation will tend to select unnecessarily large models. Finally, it is noted that since numerous criteria such as the widely used Akaike Information Criterion (AIC) of Akaike (1974) are asymptotically equivalent to leave-one-out cross-validation, the same critique also applies.

Result (5) states that the probability of asymptotically selecting a model which will yield biased predictions is zero given the assumptions in Shao (1993). Result (6) states that leave-one-out cross-validation will not be able to discriminate between models which include irrelevant variables but which otherwise are properly specified in that they include all relevant predictors even as $n \to \infty$.

That is, virtually all data-driven methods of model selection will select non mean square error optimal models with positive probability even asymptotically. Note that this holds only in the case where the number of conditioning variables is fixed as $n \to \infty$ (Shao (1993, pg 486, second to last paragraph)) thereby absolving series approximations and some other semi-parametric models whose dimension p increases with n from this critique.

3. Consistent Cross-Validation for Independent Data: Leave-n_v-out Cross-Validation

Shao (1993) considers leave-n_v-out cross-validation in which the model is fit on $n_c = n - n_v$ observations and then the prediction error is determined using the remaining n_v observations not used to fit the model, and this is conducted for all ${}_nC_{n_v}$ possible test sets. Letting $\mathcal{V}$ denote the collection of such subsets, the leave-n_v-out cross-validation function[4] is given by

$$
\begin{aligned}
CV_v &= \frac{1}{{}_nC_{n_v} n_v} \sum_{\text{all } v \in \mathcal{V}} ||Y_v - \hat{Y}_{v(-v)}||^2 \\
&= \frac{1}{{}_nC_{n_v} n_v} \sum_{\text{all } v \in \mathcal{V}} ||Y_v - X_v \hat{\beta}_{(-v)}||^2,
\end{aligned} \tag{8}
$$

where $||a|| = \sqrt{a'a}$, $\hat{\beta}_{(-v)}$ is the least squares estimator obtained by deleting v observations from the data set, and (Y_v, X_v) is the deleted sample. As a practical matter this approach becomes intractable as $n \to \infty$ since $\lim_{n\to\infty} {}_nC_{n_v} = \infty$ for finite n_v, and numerous simplifications have been proposed to circumvent this problem such as Shao's (1993, pg 488) balanced-incomplete cross-validation, Zhang's (1993, pg 307) r-fold cross-validation, and the repeated learning-testing method (Burman (1989), Burman (1990), Zhang (1993)).

The conditions required for the consistency of CV_v are given by Shao (1993, pp 488-489) and are

$$
\begin{aligned}
&\lim_{n\to\infty} n_c = \infty \\
&\lim_{n\to\infty} \frac{n_c}{n} = 0 \\
&\lim_{n\to\infty} \frac{n_v}{n} = 1.
\end{aligned} \tag{9}
$$

[4] Zhang (1993, pg 300) refers to this as 'deleting-d multifold cross-validation'.

One candidate for n_c is to let n_c be the integer part of n^δ. Letting the resulting degrees of freedom be strictly positive it is possible to obtain the following bounds on δ,

$$\frac{\log(p)}{\log(n)} < \delta < 1, \tag{10}$$

which follows directly from the condition $n_c - p > 0$ where n_c is the integer part of n^δ. Clearly δ satisfying these restrictions satisfies the conditions given in Equation (9).

It is important to note that the size of the training set n_c must not grow too fast, and the size of the test set n_v must increase as the sample size increases and cannot be a singleton. We now consider the impact of these results on a cross-validatory method for dependent data proposed by Burman et al. (1994) known as h-block cross-validation.

4. Cross-Validation for Dependent Data: h-block Cross-Validation

Let the matrix of n observations on the response and the p predictors be given by $Z = (Y, X)$. We assume that Z denotes a set of jointly dependent stationary observations. Therefore, the covariance between z_i and z_{i+j} depends only on j and approaches 0 as $|j - i| \to \infty$. Removing h observations either side of z_i will yield a new series which will be nearly independent of z_i as h increases. We remove the ith (vector-valued) observation and h observations on either 'side' of the ith thereby removing $2h+1$ observations from the sample. When these $2h+1$ observations are removed from a data set, the resulting data matrix will be denoted by $Z_{(-i:h)} = (Y_{(-i:h)}, X_{(-i:h)})$, while the matrix of removed observations will be denoted by $Z_{(i:h)} = (Y_{(i:h)}, X_{(i:h)})$. If $h = 0$ this will have the effect of removing only the ith observation from the sample.

The h-block cross-validation is given by

$$\begin{aligned} CV_h &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{i(-i:h)}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}_{(-i:h)}\right)^2. \end{aligned} \tag{11}$$

where $\hat{\beta}_{(-i:h)} = (X'_{(-i:h)}X_{(-i:h)})^{-1}X'_{(-i:h)}Y_{(-i:h)}$ is the least squares estimator obtained by removing the ith observation and the h observations either side of the ith. Note that when $h = 0$ this simplifies to leave-one-out cross-validation.

Burman et al. (1994) consider both this h-block cross-validation function and a modified h-block cross-validation function. Their modification was motivated by cases in which the ratio of the number of parameters to the sample size (p/n) is not negligible and corrects for underuse of the sample in small-sample settings. Since this paper is concerned with asymptotic behavior the ratio p/n is negligible and their finite-sample modification is not addressed in this setting.

4.1. **Inconsistency of h-Block Cross-Validation.** For h-block cross-validation, the size of the training set is $n_c = n - 2h - 1$. Györfi, Härdle, Sarda & Vieu (1989) require that $h/n \to 0$ as $n \to \infty$, which therefore implies that

$$\lim_{n\to\infty} n_c = \infty, \tag{12}$$

and that

$$\lim_{n\to\infty} \frac{n_c}{n} = 1, \tag{13}$$

while Burman et al. (1994) recommend taking h as a fixed fraction of n, that is, $h/n = \gamma$ for some $0 < \gamma < 1/2$, which again implies that

$$\lim_{n\to\infty} \frac{n_c}{n} = 1 - 2\gamma > 0 \text{ since } 0 < \gamma < \frac{1}{2}. \tag{14}$$

Note that h-block cross-validation retains a leave-one-out aspect in that the test set is a singleton. If we denote the test set again by n_v, then $n_v = 1$ for h-block cross-validation. Therefore, it is observed that for h-block cross-validation,

$$\lim_{n\to\infty} \frac{n_v}{n} = 0. \tag{15}$$

Given that $\lim_{n\to\infty} n_c/n > 0$ and $\lim_{n\to\infty} n_v/n = 0$ the results of Shao (1993) imply that h-block cross-validation is inconsistent. Also, since the h-block algorithm reduces to leave-one-out cross-validation for *iid* observations when $h = 0$, then in this case $\lim_{n\to\infty} n_c/n = 1$ and $\lim_{n\to\infty} n_v/n = 0$ again implying inconsistency. The results of Shao (1993) therefore imply that h-block cross-validation is inconsistent and will tend to select unnecessarily large models. This result is borne out by simulations summarized in Section 6. In addition, leave-n_v-out cross-validation is inappropriate for dependent data since the training and test sets will be highly dependent since they are contiguous, and randomization is clearly inappropriate since the temporal ordering of the data contains the stochastic structure of the model.

In order for h-block cross-validatory model-selection to be consistent it must be modified in such a way as to satisfy the conditions in Equation (9) while maintaining near-independence of the training and test data.

5. Consistent Cross-Validation for Dependent Data: hv-Block Cross-Validation

Fortunately there is a straightforward modification of the h-block approach which will be consistent. Instead of the test set being a singleton we will train on a set of observations of size n_c and test on a set of size n_v while maintaining near-independence of the training and test data via h-blocking. By placing restrictions on the relationship between the training set, test set, the size of the h-block, and the sample size, we can thereby obtain a consistent cross-validation procedure for general stationary processes.

We first remove v observations either side of z_i yielding a test set of size $2v+1$, and then we remove another h observations on either side of this test set with the remaining $n-2v-2h-1$ observations forming the training set. The value of v controls the size of the test set with $n_v = 2v+1$, and the value of h controls the dependence of the training set of size $n_c = n-2h-n_v$ and the test set of size n_v. We thereby remove the ith (vector-valued) observation and $v+h$ observations on either 'side' of the ith thereby removing $2h+2v+1$ observations from the sample. When these $2h+2v+1$ observations are removed from a data set, the resulting data matrix will be denoted by $Z_{(-i:h,v)} = (Y_{(-i:h,v)}, X_{(-i:h,v)})$, while the matrix of removed observations will be denoted by $Z_{(i:h,v)} = (Y_{(i:h,v)}, X_{(i:h,v)})$ and the matrix of test observations will be denoted by $Z_{(i:v)} = (Y_{(i:v)}, X_{(i:v)})$.

The resultant cross-validation function will then be given by

$$
\begin{aligned}
CV_{hv} &= \frac{1}{(n-2v)n_v}\sum_{i=v}^{n-v} ||Y_{(i:v)} - \hat{Y}_{(i:v)(-i:h,v)}||^2 \\
&= \frac{1}{(n-2v)n_v}\sum_{i=v}^{n-v} ||Y_{(i:v)} - X_{(i:v)}\hat{\beta}_{(-i:h,v)}||^2,
\end{aligned}
\tag{16}
$$

where $||a|| = \sqrt{a'a}$ and where $\hat{\beta}_{(-i:h,v)} = (X'_{(-i:h,v)}X_{(-i:h,v)})^{-1}X'_{(-i:h,v)}Y_{(-i:h,v)}$.

The parameter h controls the dependence of the test and training sets and is set to insure near independence of these sets. These sets need not be completely independent for cross-validation to work, however (Burman et al. (1994, pg 354)). The parameter v controls the relationship between the training set, test set, and the sample size.

Consider again setting n_c to be the integer part of n^δ. For positive degrees of freedom ($n_c - p > 0$) we obtain a lower bound on δ, $\log(p)/\log(n) < \delta$. If $\delta < 1$ then $n_c/n = n^{\delta-1}$ approaches 0 as n increases. As well, $n_v/n = 1 - n^\delta/n - 2h/n$ approaches 1 if $\delta < 1$ for fixed h and for h less than $o(n)$. Thus,

$$\lim_{n\to\infty} n_c = \infty, \tag{17}$$

and

$$\lim_{n\to\infty} \frac{n_c}{n} = 0, \tag{18}$$

and

$$\lim_{n\to\infty} \frac{n_v}{n} = 1. \tag{19}$$

Therefore, hv-block cross-validation is consistent if n_c is the integer part of n^δ and where $\log(p)/\log(n) < \delta < 1$, $v = (n - n^\delta - 2h - 1)/2$, and the conditions of Burman et al. (1994) or Györfi et al. (1989) are met.

An attractive feature of the proposed hv-block approach is that it contains many existing forms of cross-validation as special cases. For instance, if $h = v = 0$ this will have the effect of removing only the ith observation from the sample and this simplifies to leave-one-out cross-validation. If $h = 0$ and $v > 0$, this simplifies to leave-n_v-out cross validation in which the number of test samples is $n - n_v$ rather than the standard ${}_nC_{2v+1}$ possible samples. This is attractive since it results in feasible leave-n_v-out cross-validation for large samples, but note that the data must be pointwise randomized prior to applying this approach for *iid* data[5]. If $h > 0$ and $v = 0$ this simplifies to

[5]Shao pointed out that this is very similar to his proposed 'Monte Carlo CV' method in Shao (1993).

h-block cross-validation, while if $h > 0$ and $v > 0$ this new approach will be dubbed 'hv-block' cross-validation.

6. Simulations

Shao (1993) considered a limited simulation experiment with the sample size held fixed at $n = 40$. The purpose of those experiments was to investigate the relative performance of leave-n_v-out cross-validation and leave-one-out cross-validation for *iid* data. He considered two variants of leave-n_v-out cross-validation, Monte Carlo cross-validation (bootstrap resampling and averaging the squared predicting errors), and analytic approximate cross-validation (See Shao (1993, pg 490) for details). He found a negligible difference between the performance of leave-one-out cross-validation and analytic approximate cross-validation, while Monte Carlo cross-validation performed much better in some situations but worse when the optimal model was the largest model considered. Similarly, Shao (1996) noted the same inconsistency in the context of bootstrap model selection again for $n = 40$, and again his modified bootstrap model selection procedure performed much better than the unmodified one in some situations but worse when the optimal model is the largest model considered.

In the following sections we address the issues raised by Shao (1993) from an empirical perspective. We first consider the finite-sample relevance of his findings as $n \to \infty$ for cross-validatory model selection for *iid* data, and then consider their impact on cross-validatory model selection for dependent data.

6.1. Inconsistency of CV_1 for *iid* Data.

In order to validate Shao's (1993) results in an *iid* setting it is clearly necessary to examine the performance of leave-one-out cross-validation as $n \to \infty$. Following Shao (1993), linear models were considered with $p = 5$. Five models were compared, two from Category I, two from Category II, and one which was optimal. The data was simulated from

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \tag{20}$$

where the $x_{ij} \sim U[0,1]$ for $j = 1, \ldots, 3$ and $i = 1, \ldots, n$, $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.5$, and $(\beta_1, \beta_2, \beta_3) = (1, 1, 1)$. The sample size was increased to determine whether leave-one-out cross-validation displays any finite-sample symptoms of the asymptotic inconsistency which would require Shao's (1993) modification involving leave-n_v-out cross-validation whereby $n_c/n \to 0$ and $n_v/n \to 1$ as $n \to \infty$.

The five models were

$$\begin{aligned}
&\text{Model 1:} \quad y_i = \beta_1 x_{i1} + \epsilon_i \\
&\text{Model 2:} \quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \\
&\text{Model 3:} \quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \\
&\text{Model 4:} \quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \\
&\text{Model 5:} \quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i,
\end{aligned} \tag{21}$$

and clearly models 1 and 2 are from Category I, 4 and 5 from Category II, and 3 is optimal.

The simulations were run using the feasible algorithm of Racine (1997), and results from 1,000 simulations are given in Table 1 below.

	Category I		$\mathcal{M}_*$	Category II		$pr(\mathcal{M}_{cv} \neq \mathcal{M}^*)$	
n	Model 1	Model 2	Model 3	Model 4	Model 5	pr(I)	pr(II)
50	0.000	0.000	0.774	0.138	0.088	0.000	0.226
100	0.000	0.000	0.799	0.116	0.085	0.000	0.201
250	0.000	0.000	0.781	0.134	0.085	0.000	0.219
500	0.000	0.000	0.783	0.135	0.082	0.000	0.217
1000	0.000	0.000	0.788	0.139	0.073	0.000	0.212
2500	0.000	0.000	0.798	0.124	0.078	0.000	0.202
5000	0.000	0.000	0.791	0.146	0.063	0.000	0.209

TABLE 1. Inconsistency of CV_1 for *iid* Data. Each row represents the relative frequency of model selection via leave-one-out cross-validation for a given sample size, and 1,000 simulations were considered for each sample size.

For a given sample size, the empirical probability of choosing a model from Category I is the sum of the entries from column 2 and 3, the probability of choosing a model from Category II is the sum of columns 5 and 6, while the probability of choosing the optimal model is the entry in column 4.

The rightmost column in Table 1 contains $pr(\mathcal{M}_{cv}$ is in Category II but is not $\mathcal{M}_*)$ denoted by $pr(\text{II})$ in the table. For this example this probability remains constant at around 1/5 as $n \to \infty$. It is worth noting that results were generated for $n = 100,000$ and this probability did fall below 1/5. Therefore, $pr(\mathcal{M}_{cv_1} = \mathcal{M}_*) \neq 1$ as $n \to \infty$, hence the inconsistency result of Shao (1993) for leave-one-out cross-validation for *iid* data appears to hold up under simulation.

6.2. **Consistency of CV_v for *iid* Data.** We now consider the performance of leave-n_v-out cross-validation using the proposed algorithm with $h = 0$ in the same setting as above and arbitrarily set $\delta = 0.50$ so that conditions for consistency are met. The simulations were run using a modified version of the approach of Racine (1997), and results from 1,000 simulations are given in Table 2 below.

	Category I		$\mathcal{M}_*$	Category II		$pr(\mathcal{M}_{cv} \neq \mathcal{M}^*)$	
n	Model 1	Model 2	Model 3	Model 4	Model 5	$pr(\text{I})$	$pr(\text{II})$
50	0.029	0.241	0.683	0.041	0.006	0.270	0.047
100	0.000	0.065	0.871	0.056	0.008	0.065	0.064
250	0.000	0.016	0.945	0.038	0.001	0.016	0.039
500	0.000	0.002	0.958	0.037	0.003	0.002	0.040
1000	0.000	0.000	0.971	0.029	0.000	0.000	0.029
2500	0.000	0.000	0.986	0.012	0.002	0.000	0.014
5000	0.000	0.000	0.992	0.008	0.000	0.000	0.008

TABLE 2. Consistency of CV_v for *iid* Data, $\delta = 0.50$. Each row represents the relative frequency of model selection via leave-n_v-out cross-validation for a given sample size, and 1,000 simulations were considered for each sample size.

The expectation is for $pr(\mathcal{M}_{cv}$ is in Category II but is not $\mathcal{M}_*)$ in Table 2 to approach 0 as $n \to \infty$, and for this example this occurs. As well, $pr(\mathcal{M}_{cv_1} = \mathcal{M}_*) \to 1$ as $n \to \infty$ for leave-n_v-out cross-validation, hence the consistency result of Shao (1993) for *iid* data appears to hold up under simulation provided that n_c and n_v satisfy the conditions from Equation (9).

6.3. **Inconsistency of CV_h for Dependent Data.** One of the goals of this paper is to determine whether the inconsistency result of Shao (1993) also applies to h-block cross-validation, a cross-validatory method for dependent data. The following example considers the performance of h-block cross-validation as $n \to \infty$, and interest lies in $pr(\mathcal{M}_{cv}$ is in Category II) as $n \to \infty$. If this inconsistency result holds, we would expect that $pr(\mathcal{M}_{cv}$ is in Category II but is not $\mathcal{M}_*) > 0$ as $n \to \infty$.

Linear time-series models were considered with $p = 5$ where again five models were compared, two from Category I, two from Category II, and one which was optimal. The data was simulated from an AR(3) process given by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \epsilon_t, \tag{22}$$

where $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.15$, and $(\phi_1, \phi_2, \phi_3) = (0.9, -0.8, 0.7)$[6]. The sample size was increased to determine whether h-block cross-validation displays any finite-sample symptoms of the asymptotic inconsistency.

The five models were

$$\begin{aligned}
&\text{Model 1:} \quad y_i = \phi_1 y_{i-1} + \epsilon_i \\
&\text{Model 2:} \quad y_i = \phi_1 y_{i-1} + \phi_2 y_{i-2} + \epsilon_i \\
&\text{Model 3:} \quad y_i = \phi_1 y_{i-1} + \phi_2 y_{i-2} + \phi_3 y_{i-3} + \epsilon_i \\
&\text{Model 4:} \quad y_i = \phi_1 y_{i-1} + \phi_2 y_{i-2} + \phi_3 y_{i-3} + \phi_4 y_{i-4} + \epsilon_i \\
&\text{Model 5:} \quad y_i = \phi_1 y_{i-1} + \phi_2 y_{i-2} + \phi_3 y_{i-3} + \phi_4 y_{i-4} + \phi_5 y_{i-5} + \epsilon_i,
\end{aligned} \tag{23}$$

and again models 1 and 2 are from Category I, 4 and 5 from Category II, and 3 is optimal.

The size of the h-block was set using $h = np$ with $\gamma = 0.25$ in accordance with the suggestion of Burman et al. (1994, pg 356).

The rightmost column in Table 3 contains $pr(\mathcal{M}_{cv}$ is in Category II but is not $\mathcal{M}_*)$ denoted by $pr(\text{II})$ in the table. Note that $pr(\mathcal{M}_{cv}$ is in Category II but is not $\mathcal{M}_*)$ does not go below roughly 1/3 as $n \to \infty$. As well, $pr(\mathcal{M}_{cv_1} = \mathcal{M}_*) \neq 1$ as $n \to \infty$, hence the inconsistency result of Shao (1993) appears to also hold for h-block cross-validation with dependent data, as outlined in Section 4.1.

[6] Shao has commented that the ratio of ϕ/σ (the average ϕ) controls the convergence speed of the cross-validation algorithm. The DGP for the *iid* case had $\beta/\sigma = 2$ for the average value of β. Therefore, for comparison of simulation results I choose $\sigma = 0.15$ so that the average $\phi/\sigma \approx 2$ and the results should be comparable in both cases.

	Category I		$\mathcal{M}_*$	Category II		$pr(\mathcal{M}_{cv} \neq \mathcal{M}^*)$	
n	Model 1	Model 2	Model 3	Model 4	Model 5	pr(I)	pr(II)
50	0.005	0.008	0.669	0.178	0.140	0.013	0.318
100	0.000	0.000	0.637	0.217	0.146	0.000	0.363
250	0.000	0.000	0.654	0.200	0.146	0.000	0.346
500	0.000	0.000	0.668	0.177	0.155	0.000	0.332
1000	0.000	0.000	0.656	0.190	0.154	0.000	0.344
2500	0.000	0.000	0.654	0.200	0.146	0.000	0.346
5000	0.000	0.000	0.657	0.211	0.132	0.000	0.343

TABLE 3. Inconsistency of CV_h for Dependent Data, $\gamma = 0.25$. Each row represents the relative frequency of model selection via h-block cross-validation for a given sample size, and 1,000 simulations were considered for each sample size.

6.4. **Consistency of CV_{hv} for Dependent Data.** We now consider the performance of the proposed hv-block cross-validation in the same setting as above and arbitrarily set $\gamma = 0.25$ and $\delta = 0.50$ so that conditions for consistency are met. The simulations were run using a modified version of the algorithm of Racine (1997), and results from 1,000 simulations are given in Table 4 below.

	Category I		$\mathcal{M}_*$	Category II		$pr(\mathcal{M}_{cv} \neq \mathcal{M}^*)$	
n	Model 1	Model 2	Model 3	Model 4	Model 5	pr(I)	pr(II)
50	0.076	0.033	0.809	0.068	0.014	0.109	0.082
100	0.013	0.000	0.878	0.091	0.018	0.013	0.109
250	0.000	0.000	0.920	0.064	0.016	0.000	0.080
500	0.000	0.000	0.931	0.056	0.013	0.000	0.069
1000	0.000	0.000	0.955	0.041	0.004	0.000	0.045
2500	0.000	0.000	0.964	0.035	0.001	0.000	0.036
5000	0.000	0.000	0.987	0.012	0.001	0.000	0.013

TABLE 4. Consistency of CV_{hv} for Dependent Data, $\gamma = 0.25$, $\delta = 0.50$. Each row represents the relative frequency of model selection via hv-block cross-validation for a given sample size, and 1,000 simulations were considered for each sample size.

If the proposed hv-block method is consistent in the sense of Shao (1993), the expectation would be for $pr(\mathcal{M}_{cv}$ is in Category II but is not $\mathcal{M}_*)$ in Table 4 to approach 0 as $n \to \infty$, and for this example this occurs. As well, $pr(\mathcal{M}_{cv_1} = \mathcal{M}_*) \to 1$ as $n \to \infty$, hence the consistency result of Shao (1993) for hv-block cross-validation for dependent data appears to hold up under simulation when γ and δ satisfy the conditions required for consistency as outlined in Section 5.

7. Discussion of Simulation Results

The above simulations validate the findings of inconsistency of Shao (1993) for leave-one-out cross-validation for *iid* data, and confirm that they are also relevant for h-block cross-validation for dependent data. In addition, the simulations validate the consistency of leave-n_v-out cross-validation for *iid* data and of the proposed hv-block cross-validation for dependent data.

Some regularities appear in the simulations which are noteworthy. First, for small sample sizes such as $n = 50$, the consistent approaches CV_v and CV_{hv} have a larger probability of the selected model lying in Category I than the inconsistent approaches CV_1 and CV_h for the examples considered above. That is, in small samples the consistent approaches tend to choose underparameterized models while the inconsistent approaches tend to choose overparameterized models. As well, it appears that this probability increases with δ and is therefore inversely related to the training data size n_v. However, it is likely that this is simply a reflection of not using the finite-sample corrections for leave-n_v-out cross-validation of Burman (1989) and that for h-block cross-validation found in Burman et al. (1994). As the focus of this paper lies with the asymptotic inconsistency of h-block cross-validation, this correction has not been incorporated here, and it remains an open question whether the corrections will improve the performance of the consistent approaches CV_v and CV_{hv} in very small samples, though preliminary investigation suggests that this will indeed be the case. Second, the inconsistency result for h-block cross-validation manifests itself in moderate sample-size settings and is clearly more than an asymptotic curiosity. Third, the proposed consistent hv-block approach performs much better than it's inconsistent counterpart for all sample sizes.

What guidelines can be offered for optimal settings of γ and δ for the proposed approach? For the examples given, $\gamma = 0.25$ and $\delta = 0.50$ appear to be reasonable but clearly will be application-specific. Though no general solution to this problem is offered at this time, Appendix B examines the effects of various settings for γ and δ on the algorithm's performance. A general solution to this problem remains the subject of future work in the area.

8. Application: Autoregression Order for G7 Exchange Rate Series

The modest aim of this application is simply to gauge whether the proposed algorithm makes a difference, that is, will there be dramatic differences in the models selected in applied settings or will all approaches tend to select the same model.

It is of interest to compare the performance of the proposed hv-block approach with a traditional model selection criterion and with variants such as leave-one-out, v-block, and h-block cross-validation. The traditional criterion used here for comparison purposes is the well-known Akaike Information Criterion (AIC) (Akaike (1974)). The leave-one-out and v-block variants are clearly inappropriate for dependent data, while the h-block approach has been shown to be inconsistent in Section 6.3. As well, it would be interesting to see how the leave-one-out, v-block, and AIC methods would perform when in fact the data form a dependent sequence, and Appendix A considers their relative performance when the data form a simulated dependent sequence. As is seen in Appendix A, these methods do not perform nearly as well as the proposed hv-block approach, tending to select overly parameterized models as would be expected based on the results of Shao (1993).

We now consider the problem of selecting the order of an autoregression for exchange rate data using cross-validatory methods. Exchange rate modeling has a long history going back decades to early work on efficient markets by Fama (1965) and Cootner (1964), while Meese & Rogoff (1983) remains the seminal reference for out-of-sample point prediction of nominal exchange rates. Monthly nominal exchange rate data for the G7 countries was obtained from Citibase for time periods 1972:2 through 1994:10 (all rates were quoted relative to the US\$). The range of models from which a model was selected went from an AR(0) (regression on a constant) through an AR(6) model (that is, $y_t = \phi_0 + \sum_{j=1}^{k} \phi_j y_{t-j} + \epsilon_t$ for $k = 0, \ldots 6$). The model selection results for the various criteria are summarized in Table 5 below, while Appendix C presents the partial autocorrelation (PAC) functions for each series for comparison purposes.

The results contained in Appendix A and Section 6.4 suggest that the proposed hv-block approach performs best for dependent data followed by the v-block, AIC, and then the h-block approach. The tendency for the non hv-block approaches is to select over-parameterized exchange rate models as would be expected based on the results of Shao (1993) and which can be seen by an examination

Country	AIC	Leave-one-out ($\gamma = 0, \delta = 1$)	v-Block ($\gamma = 0, \delta = 0.5$)	h-block ($\gamma = 0.25, \delta = 1$)	$hv - block$ ($\gamma = 0.25, \delta = 0.5$)
Canada	AR(2)	AR(2)	AR(4)	AR(2)	AR(1)
England	AR(4)	AR(4)	AR(3)	AR(3)	AR(1)
France	AR(4)	AR(4)	AR(1)	AR(2)	AR(1)
Germany	AR(2)	AR(2)	AR(1)	AR(2)	AR(1)
Italy	AR(2)	AR(2)	AR(1)	AR(2)	AR(1)
Japan	AR(4)	AR(4)	AR(1)	AR(3)	AR(1)

TABLE 5. Autoregressive model order selected by various criteria.

of Table 5. For example, when modeling the England/US series, the AIC and leave-one-out criteria select AR(4) models, the v-block and h-block criteria select AR(3) models, while the proposed hv-block approach selects an AR(1) model.

The models selected via hv-block cross-validation given in Table 5 are consistent with model selection based on the PAC function, as can be seen from the PAC functions presented in Appendix C. Based solely on the PAC function, one would conclude that an AR(1) specification is appropriate for Canada, Germany, Italy, and Japan, while at most an AR(2) specification would be appropriate for England and France. However, cross-validation goes farther than a simple examination of the PAC function by in effect comparing out-of-sample prediction errors across AR(P) models and would therefore be expected to be a preferable model selection criterion for prediction purposes.

This modest application suggests that the use of the proposed hv-block method for model selection can make a noticeable difference in applied settings and thus can constitute a valuable addition to the tools employed by applied researchers in time-series settings.

9. CONCLUSION

Cross-validatory model selection for dependent processes has widespread potential application ranging from selection of the order of an autoregression for linear time series models to selection of the number of terms in nonlinear series approximations for general stationary processes. In this paper we consider the impact of Shao's (1993) result of the inconsistency of linear model selection via cross-validation on h-block cross-validation, a cross-validatory method for dependent data. It is shown that model selection via h-block cross-validation is inconsistent and therefore not asymptotically optimal. A modification of the h-block method, dubbed 'hv-block' cross-validation, is proposed which is asymptotically optimal. The proposed approach is consistent in the sense that the probability of selecting the model with the best predictive ability converges to 1 as the total number of observations approaches infinity.

The proposed hv-block cross-validatory approach to model selection extends existing results and contains numerous variants of cross-validation including leave-one-out cross-validation, leave-n_v-out cross-validation, and h-block cross-validation as special cases.

A number of simulations and applications are conducted to examine the inconsistency of leave-one-out cross-validation and h-block cross-validation and to validate consistency of leave-n_v-out cross-validation for *iid* data and the proposed hv-block cross-validation for general stationary data. The simulations and applications provide insight into the nature of the various methods and suggest that the proposed method outperforms existing variants of cross-validation in time-series contexts and can be of value in applied settings.

Much work remains to be done before a seamless approach towards model selection can exist, however. Optimal choice of δ and γ for hv-block cross-validation is clearly an important issue, and a sound framework regarding optimal settings for δ and γ remains the subject of future work in this area.

Appendix A. Relative Performance of Algorithms for Dependent Data

The following chart presents simulation results for leave-one-out, h-block, and v-block variants for the DGP given in Section 6.4. For the purpose of comparison with a traditional model-selection criteria the well-known Akaike Information Criterion (AIC) of Akaike (1974) is also computed.

	Category I		$\mathcal{M}_*$	Category II		$pr(\mathcal{M}_{cv} \neq \mathcal{M}^*)$	
n	Model 1	Model 2	Model 3	Model 4	Model 5	pr(I)	pr(II)
			AIC				
50	0.000	0.000	0.783	0.137	0.080	0.000	0.217
100	0.000	0.000	0.774	0.145	0.081	0.000	0.226
250	0.000	0.000	0.766	0.154	0.080	0.000	0.234
500	0.000	0.000	0.811	0.116	0.073	0.000	0.189
1000	0.000	0.000	0.778	0.132	0.090	0.000	0.222
2500	0.000	0.000	0.774	0.140	0.086	0.000	0.226
5000	0.000	0.000	0.797	0.126	0.077	0.000	0.203
			Leave-one-out ($\gamma = 0.0, \delta = 1.0$)				
50	0.000	0.000	0.794	0.122	0.084	0.000	0.206
100	0.000	0.000	0.767	0.148	0.085	0.000	0.233
250	0.000	0.000	0.769	0.147	0.084	0.000	0.231
500	0.000	0.000	0.807	0.115	0.078	0.000	0.193
1000	0.000	0.000	0.771	0.136	0.093	0.000	0.229
2500	0.000	0.000	0.779	0.136	0.085	0.000	0.221
5000	0.000	0.000	0.797	0.127	0.076	0.000	0.203
			h-block ($\gamma = 0.25, \delta = 1.0$)				
50	0.005	0.008	0.669	0.178	0.140	0.013	0.318
100	0.000	0.000	0.637	0.217	0.146	0.000	0.363
250	0.000	0.000	0.654	0.200	0.146	0.000	0.346
500	0.000	0.000	0.668	0.177	0.155	0.000	0.332
1000	0.000	0.000	0.656	0.190	0.154	0.000	0.344
2500	0.000	0.000	0.654	0.200	0.146	0.000	0.346
5000	0.000	0.000	0.657	0.211	0.132	0.000	0.343
			v-block ($\gamma = 0.0, \delta = 0.5$)				
50	0.178	0.038	0.737	0.044	0.003	0.216	0.047
100	0.074	0.015	0.833	0.065	0.013	0.089	0.078
250	0.001	0.001	0.935	0.045	0.018	0.002	0.063
500	0.000	0.000	0.924	0.061	0.015	0.000	0.076
1000	0.000	0.000	0.929	0.055	0.016	0.000	0.071
2500	0.000	0.000	0.933	0.053	0.014	0.000	0.067
5000	0.000	0.000	0.949	0.045	0.006	0.000	0.051

It can be seen by a comparison with Table 4 that the proposed hv-block method appears to outperform all variants of cross-validation when the data form a dependent sequence.

Appendix B. Performance of hv-Block Cross-Validation: Selection of γ and δ

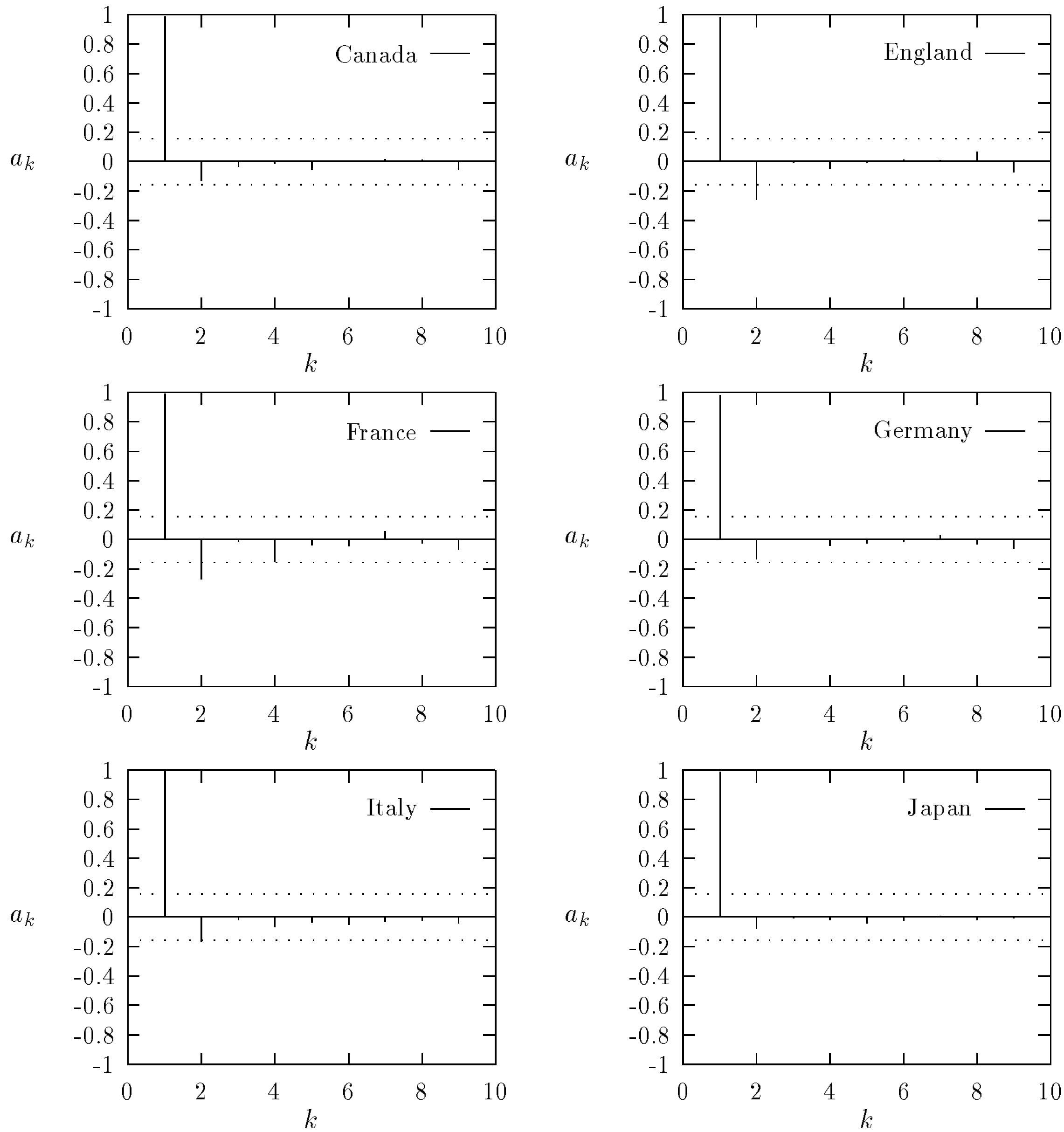
It is of interest to consider the effect of changing the values of γ and δ upon the performance of the hv-block algorithm. The following chart presents simulation results for the DGP given in Section 6.4.

	Category I		$\mathcal{M}_*$	Category II		$pr(\mathcal{M}_{cv} \neq \mathcal{M}^*)$	
n	Model 1	Model 2	Model 3	Model 4	Model 5	pr(I)	pr(II)
hv-block ($\gamma = 0.1, \delta = 0.75$)							
50	0.013	0.007	0.753	0.138	0.089	0.020	0.227
100	0.001	0.000	0.763	0.153	0.083	0.001	0.236
250	0.000	0.000	0.814	0.129	0.057	0.000	0.186
500	0.000	0.000	0.852	0.108	0.040	0.000	0.148
1000	0.000	0.000	0.876	0.094	0.030	0.000	0.124
2500	0.000	0.000	0.898	0.087	0.015	0.000	0.102
5000	0.000	0.000	0.944	0.048	0.008	0.000	0.056
hv-block ($\gamma = 0.1, \delta = 0.5$)							
50	0.131	0.034	0.783	0.049	0.003	0.165	0.052
100	0.036	0.002	0.870	0.077	0.015	0.038	0.092
250	0.000	0.000	0.921	0.067	0.012	0.000	0.079
500	0.000	0.000	0.943	0.047	0.010	0.000	0.057
1000	0.000	0.000	0.957	0.039	0.004	0.000	0.043
2500	0.000	0.000	0.973	0.026	0.001	0.000	0.027
5000	0.000	0.000	0.992	0.008	0.000	0.000	0.008
hv-block ($\gamma = 0.25, \delta = 0.75$)							
50	0.014	0.012	0.686	0.171	0.117	0.026	0.288
100	0.001	0.000	0.678	0.191	0.130	0.001	0.321
250	0.000	0.000	0.721	0.170	0.109	0.000	0.279
500	0.000	0.000	0.779	0.146	0.075	0.000	0.221
1000	0.000	0.000	0.805	0.137	0.058	0.000	0.195
2500	0.000	0.000	0.850	0.115	0.035	0.000	0.150
5000	0.000	0.000	0.892	0.085	0.023	0.000	0.108
hv-block ($\gamma = 0.25, \delta = 0.5$)							
50	0.076	0.033	0.809	0.068	0.014	0.109	0.082
100	0.013	0.000	0.878	0.091	0.018	0.013	0.109
250	0.000	0.000	0.920	0.064	0.016	0.000	0.080
500	0.000	0.000	0.931	0.056	0.013	0.000	0.069
1000	0.000	0.000	0.955	0.041	0.004	0.000	0.045
2500	0.000	0.000	0.964	0.035	0.001	0.000	0.036
5000	0.000	0.000	0.987	0.012	0.001	0.000	0.013

It can be seen that the default values of $\gamma = 0.25$ and $\delta = 0.50$ suggested by existing literature on h-block and v-block cross-validation appear to be reasonable for the DGP considered.

Appendix C. Partial Autocorrelation Functions for G7 Exchange Rates

The partial autocorrelation (PAC) functions for the nominal exchange rate series modeled in Section 8 are graphed below along with 99% confidence intervals around $a_k = 0$. Based on the PAC function, one would conclude that an AR(1) specification is appropriate for Canada, Germany, Italy, and Japan, while *at most* an AR(2) specification would be appropriate for England and France.



References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions in Automaic Control* **19**(6), 716–723.

Allen, D. M. (1974), 'The relationship between variable selection and data augmentation and a method for prediction', *Technometrics* **16**, 125–127.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Burman, P. (1989), 'A comparative study of ordinary cross-validation, v-fold cross-validation, and the repeated learning-testing methods', *Biometrika* **76**, 503–514.

Burman, P. (1990), 'Estimation of optimal transformations using v-fold cross-validation and repeated learning-testing methods', *Sankhyā* **52**, 314–345.

Burman, P., Chow, E. & Nolan, D. (1994), 'A cross-validatory method for dependent data', *Biometrika* **84**, 351–358.

Burman, P. & Nolan, D. (1992), 'Data dependent estimation of prediction functions', *Journal of Time Series Analysis* **13**, 189–207.

Cootner, P. (1964), *The Random Character of Stock Market Prices*, MIT Press, Cambridge, MA.

Fama, E. F. (1965), 'The behavior of stock market prices', *Journal of Business* **38**, 34–105.

Geisser, S. (1975), 'A predictive sample reuse method with application', *Journal of the American Statistical Association* **70**, 320–328.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989), *Nonparametric Curve Estimation from Time Series*, Springer-Verlag.

Meese, R. & Rogoff, K. D. (1983), 'Empirical exchange rate models of the seventies: Do they fit out of sample?', *Journal of International Economics* **14**, 3–24.

Racine, J. (1997), 'Feasible cross-validatory model selection for general stationary processes', *Journal of Applied Econometrics* **12**, 169–179.

Shao, J. (1993), 'Linear model selection by cross-validation', *Journal of the American Statistical Association* **88**(422), 486–495.

Shao, J. (1996), 'Bootstrap model selection', *Journal of the American Statistical Association* **91**(434), 655–665.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions (with discussion)', *Journal of the Royal Statistical Society* **36**, 111–147.

Stone, M. (1977), 'Cross-validation: a review', *Math. Oper. Statist.* **9**, 127–139.

Whaba, G. & Wold, S. (1975), 'A completely automatic french curve: Fitting spline functions by cross-validation', *Communications in Statistics* **4**, 1–17.

Zhang, P. (1993), 'Model selection via multifold cross-validation', *Annals of Statistics* **21**, 299–313.