Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–25

Cross-Validatory Model Selection for Bayesian Autoregressions with Exogenous Regressors

Alex Cooper[*], Dan Simpson[§], Lauren Kennedy[*,†], Catherine Forbes[*], and Aki Vehtari[‡]

Abstract. Bayesian cross-validation (CV) is a popular method for predictive model assessment that is simple to implement and broadly applicable. A wide range of CV schemes is available for time series applications, including generic leave-one-out (LOO) and K-fold methods, as well as specialized approaches intended to deal with serial dependence such as leave-future-out (LFO), h-block, and hv-block.

Existing large-sample results show that both specialized and generic methods are applicable to models of serially-dependent data. However, large sample consistency results overlook the impact of sampling variability on accuracy in finite samples. Moreover, the accuracy of a CV scheme depends on many aspects of the procedure. We show that poor design choices can lead to elevated rates of adverse selection.

In this paper, we consider the problem of identifying the regression component of an important class of models of data with serial dependence, autoregressions of order p with q exogenous regressors ($\mathsf{ARX}(p, q)$), under the logarithmic scoring rule. We show that when serial dependence is present, scores computed using the joint (multivariate) density have lower variance and better model selection accuracy than the popular pointwise estimator. In addition, we present a detailed case study of the special case of ARX models with fixed autoregressive structure and variance. For this class, we derive the finite-sample distribution of the CV estimators and the model selection statistic. We conclude with recommendations for practitioners.

Keywords: model comparison, cross-validation, uncertainty, serial dependence.

1 Overview

Many workflows for constructing predictive Bayesian models require the practitioner to choose the best model among a number of candidates according to their predictive power for the task at hand. Although many predictive model selection methods are available (Vehtari and Ojanen, 2012), among the most popular is cross-validation (CV; Geisser, 1975). CV is flexible and applicable to a wide variety of statistical applications.

arXiv: 2301.08276

[*] Department of Econometrics and Business Statistics, Monash University, Clayton, Australia, alexander.cooper@monash.edu; catherine.forbes@monash.edu

[†] School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, Australia, lauren.a.kennedy@adelaide.edu.au

[‡] Department of Computer Science, Aalto University, Helsinki, Finland, aki.vehtari@aalto.fi

[§] Normal Computing, New York, dan@normalcomputing.ai

https://doi.org/10.1214/23-BA1409

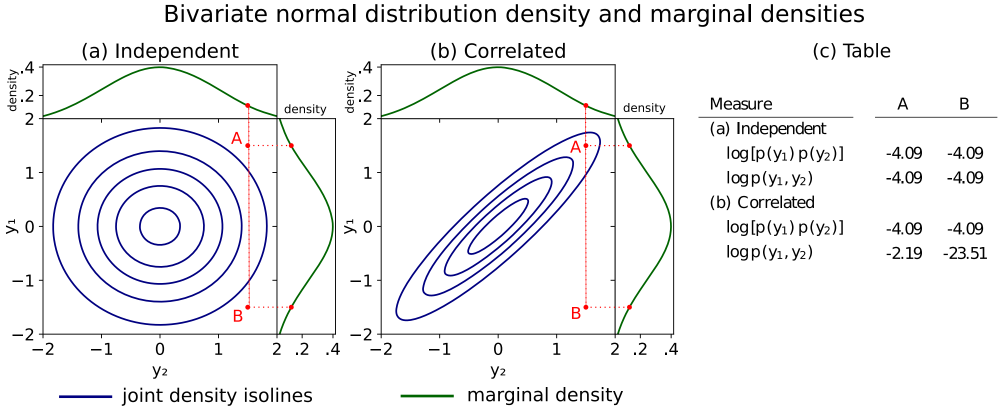
In finite samples, CV-based selection objectives are biased and subject to sampling variation, which leads to uncertainty about model predictive ability and the possibility of adverse model selection (Arlot and Celisse, 2010; Sivula et al., 2020). In the Bayesian context, there is a large literature on the frequency properties (i.e. variability across multiple realizations of a dataset) of model selection rules using information criteria such as the widely-available information criterion (WAIC) and the Bayes factor (e.g., Ward, 2008; Schad et al., 2022). Despite its popularity in Bayesian applications, however, less is known about the frequency properties of dependent CV procedures for Bayesian model selection under log-predictive loss.

Recent work by Sivula et al. (2020) analyzed the frequency properties of leave-one-out CV (LOO-CV) for Bayesian regression models of exchangeable data. The authors identify at least three scenarios that lead to elevated uncertainty in CV model selection, and therefore to an increased probability of adverse model choice. These pathological cases include comparisons between candidate models that produce similar predictions, where models are badly misspecified, and where training data sizes are small.

In this study, we extend the analysis of Sivula et al. (2020) to Bayesian models of serially dependent data. We aim to characterize CV model selection uncertainty for a simple but important class of models: autoregressions of order p with q exogenous regressors, $\mathsf{ARX}(p, q)$. Our goal is to identify the regression component of the model under the logarithmic scoring rule, leaving to one side the related task of identifying the autoregressive component. In this context, a scoring rule is a loss function for assessing the quality of probabilistic predictions (Gneiting and Raftery, 2007). While many scoring rules are available, we focus on the logarithmic scoring rule, for which a measure of predictive performance is the expected log predictive density (elpd) described in Section 2.2.

We address two important aspects of scoring rule design for models of correlated data. First, whether the scoring rule used for model assessment will be univariate or multivariate. Second, for multivariate scoring rules, whether it will be evaluated jointly (as a multivariate predictive density) or pointwise (as univariate marginal densities). We begin with a demonstration of the importance of the latter, showing improved statistical power of model selection with a jointly-evaluated scoring rule. We continue with a detailed case study of model selection under several popular CV schemes. This comparison includes several specific univariate (pointwise) methods, and several joint methods. Throughout, we find that joint methods achieve greater (statistical) efficiency, measured as lower adverse selection rates, and the associated CV estimators tend to have lower variability.

Figure 1 illustrates the importance of joint multivariate assessments when data are correlated. The figure shows two bivariate normal distributions. In one the variates are mutually independent, and in the other they are strongly correlated. The table shows that points A and B have identical pointwise log densities under both distributions, even though point B lies in a region of very low (joint) density in the correlated case. We conclude that in contrast to the joint approach, when correlation is strong the pointwise density fails to detect that point B is in a region of low probability for this model.



Measure	A	B
(a) Independent		
$\log[p(y_1)\,p(y_2)]$	-4.09	-4.09
$\log p(y_1, y_2)$	-4.09	-4.09
(b) Correlated		
$\log[p(y_1)\,p(y_2)]$	-4.09	-4.09
$\log p(y_1, y_2)$	-2.19	-23.51

Figure 1: Illustration of the distinction between joint and pointwise log density measures in correlated models. Both plots show bivariate normal densities centered at the origin with unit marginal variance and correlation coefficients of 0 (Panel (a)) and 0.9 (Panel (b)). Panel (c) tabulates joint and pointwise log densities evaluated at the marked points A and B, indicated in red on the plots. In the correlated case, only the joint log density, $\log p(y_1, y_2)$, identifies point B as having low log density. In contrast, the pointwise density $\log p(y_1, y_2)$ is the same for both.

To further motivate our approach, Figure 2 previews the results of a model selection experiment we will describe in more detail in Section 3. The figure compares the distribution of CV model selection statistics for selecting between two candidate ARX models. The panels show increasing degrees of dependence from left to right. As dependence increases, the pointwise model selection statistic shows an increasing rate of adverse model selection (red bars in the vertical margin). In contrast, there is little change for the jointly-evaluated case (red bars in the horizontal margin). For a full description of this experiment, please refer to Section 3.

Several CV strategies are available for models of serially dependent data, and there seems to be little agreement in the literature about which one practitioners should adopt, especially in the Bayesian modeling literature. Furthermore, much of the existing literature addresses different blocking strategies, but does not make a distinction between joint and pointwise evaluation of the scoring rule. Moreover, we speculate that different CV schemes will be useful when assessing different aspects of such a model. Even when the analytical focus is not the autoregressive component, joint predictive measures appear to be useful for identifying the regression components.

In contrast to much of the existing literature on CV for autoregressive models, including many large-sample consistency results (e.g., Bergmeir et al., 2018; Racine, 2000), our emphasis is on the frequency properties of the CV estimator in finite samples. Further to the three problematic scenarios identified by Sivula et al. (2020), our results suggest that strong serial dependence and a cross-validatory objective function that does not capture model dynamics (e.g. pointwise objectives) can pose difficulties for

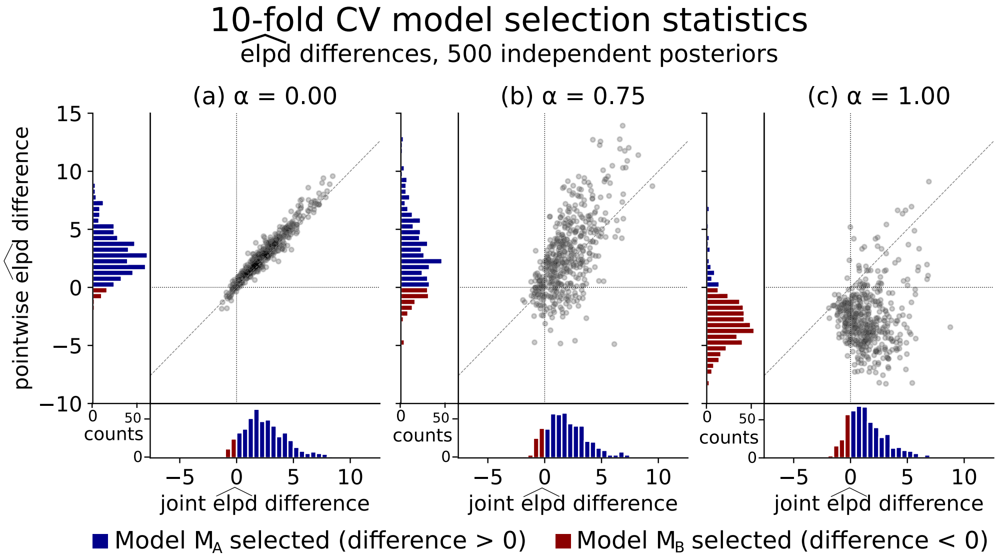


Figure 2: Joint log-score differences versus pointwise log score differences, computed using 10-fold-CV in a model selection statistics for 500 independent posteriors of the 'hard' case (see Section 3). The DGP is a stationary $\mathsf{ARX}(2,3)$ and candidate models are $M_A : \mathsf{ARX}(1,2)$ and $M_B : \mathsf{ARX}(1,1)$. Model selection statistics are expected log pointwise predictive density (elpd) differences. The DGP has autoregressive parameter $\phi_* = \alpha(0.75, 0.2)$ so that α selects increasing serial dependence from left to right. Adverse selection increases for the pointwise method (vertical axis) as dependence increases. See Section 3 for a full description.

CV-based model selection, even when the goal is not limited to identification of the autoregressive component of the model. Our results stand as a counterpoint to large-sample consistency results and suggest that mere consistency of the estimator is not enough. That is, under certain choices of the CV scheme the variance of the model selection statistic can be very high, leading to elevated rates of adverse selection.

1.1 Contributions

We present novel results for procedures that use CV methods under the logarithmic scoring function when serial dependence is present. Working with the logarithmic scoring function and focusing on identifying the regression component of the model, we demonstrate that:

- Under serial dependence, CV schemes should be designed to account for the presumed dependence structure of the data in order to achieve good model selection performance;

- When serial dependence is strong, performance measures evaluated jointly are much more (statistically) efficient than pointwise counterparts;
- When the sample size is finite, there is a U-shaped relationship between certain CV scheme hyperparameters and the adverse selection rate;
- We present novel results on the variability of Bayesian CV procedures for $\mathsf{ARX}(p, q)$ models. To our knowledge, these are among the first results describing the finite-sample uncertainty of CV methods under serial dependence, particularly in a Bayesian setting.

We offer the following advice for practitioners working with models of serially-dependent data. Broadly speaking, since CV methods based on joint scoring rules are usually more complex to implement, their improved efficiency should be traded off against implementation burden. Following model criticism of each candidate model ahead of model selection, we recommend:

- Where measured serial dependence in the data is not very strong, simpler pointwise CV methods (like LOO-CV and LFO-CV) can be used as a first-pass, and relied upon where the results are clear (see Sivula et al., 2020, for criteria);
- Otherwise, if serial dependence is strong or results are unclear then joint CV model selection methods should be implemented instead.
- Even when the actual predictive task requires a univariate prediction (like a one-step-ahead prediction), for model selection it may be better to use a CV scheme that leaves out multiple observations, combined with a multivariate scoring rule.

The remainder of the paper proceeds as follows. In Section 2 we describe the model class and summarize CV-based model selection and some relevant literature, highlighting some key challenges associated with CV for dependent data. Section 3 presents a short simulation experiment, and Section 4 presents a detailed case study of CV model selection in a simplified form of ARX model, focusing on the properties of CV under dependence and demonstrating where challenges can arise. Finally, Section 5 discusses the results and concludes. See `https://github.com/kuperov/arx` for code and experiments.

2 Background

In this section, we briefly review CV model selection and review some relevant literature. We will suppose we have observed a data vector $y = (y_1, \dots, y_T)$, presumed to be drawn from a joint distribution p_{true}, the (typically unknown) data-generating process (DGP). Our goal is to construct predictions by first selecting the best available model M^* from some set $\mathcal{M}$ of candidates (or candidate model families identified up to a parameter). This selection is made according to candidate models' ability to predict as-yet unseen realizations of the process, that is, by their out-of-sample predictive power.

To simplify our analysis, we will consider only pairwise comparisons (i.e. $|\mathcal{M}| = 2$), but we do explicitly allow that $M_{\text{true}} \notin \mathcal{M}$, i.e. the model associated with p_{true} is not in $\mathcal{M}$.

2.1 Autoregressions with exogenous regressors

We will write $\mathsf{ARX}(p, q)$ for an autoregression with p lags of the dependent variable and q exogenous regressors. Throughout we assume $p \ll T$ and $q \ll T$, and that T is 'small', corresponding to common applied settings where data are limited.

The $\mathsf{ARX}(p, q)$ class is a key building block for time series models in a wide range of scientific, policy, and business applications. For instance, autoregressions underpin the popular vector autoregression (VAR) models used by macroeconomic policymakers (e.g., Sims, 1980) and spatial epidemiological studies (Lee, 2011).

The $\mathsf{ARX}(p, q)$ model is conditionally normal,

$$p(y_t|\phi, \beta, y_{t-1}, \ldots, y_{t-p}) = \mathcal{N}\big(y_t|\phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + z_t^\top \beta, \sigma^2\big), \tag{2.1}$$

for $t = 1, \ldots, T$, where the first element of the $q \times 1$ vector of exogenous variables z_t is 1. For simplicity, we will initialize the sequence from zero, so that $y_{1-p} = \cdots = y_0 = 0$.

In comparison to the linear regression (LR) model studied by Sivula et al. (2020), the dependence structure of the ARX class substantially complicates the analysis. Naturally, one could view the lags of y_t as explanatory variables which would mean the $\mathsf{ARX}(p, q)$ model is identical to that of the Gaussian linear regression (LR) analyzed by Sivula et al. (2020). However, to restrict our attention to models in the stationary regime, we must either impose informative priors on the autoregressive parameters (as we do in Section 3) or fix them (as we do in Section 4). In both Sections 3 and 4 we have allowed analytical convenience to guide the choice of prior.

It is worth emphasizing that all the results we present in this paper depend on Z, the $T \times q$ matrix of exogenous covariates. Our results do not need to make any assumptions about the distribution of Z, since they are assumed known and fixed. In our experiments, we construct Z by drawing a matrix of independent standard normal variates, a matrix we keep fixed across all replicates of each experiment.

2.2 Predictive model selection

When the goal of a modeling exercise is prediction, it is natural to use predictive performance as a measure of model goodness or 'utility'. Predictive performance can be assessed using a scoring rule, a function that produces a numerical assessment of a probabilistic prediction against actual observations (Gneiting and Raftery, 2007). Since the choice of scoring rule governs the selection of M^*, an ideal choice for a scoring rule would be tailored to the modeling task at hand. However, in the absence of a specific application, general-purpose scoring rules are available.

We focus on the popular logarithmic scoring rule, which enjoys the mathematical properties of being local and strictly proper and is closely related to the KL divergence

(Gneiting and Raftery, 2007; Vehtari and Ojanen, 2012; Dawid, 1984). Under the logarithmic score, we call the expected score for some model M_ℓ the *expected log joint predictive density* (eljpd),

$$\begin{aligned} \mathrm{eljpd}(M_\ell|y) &= \mathbb{E}_{\tilde{y}\sim p_{\mathrm{true}}}\left[\log \int p(\tilde{y}|\theta, M_\ell)p(\theta \mid y, M_\ell)\,\mathrm{d}\theta\right] && (2.2)\\ &= \mathbb{E}_{\tilde{y}\sim p_{\mathrm{true}}}\left[\log p(\tilde{y}|y, M_\ell)\right] && (2.3)\end{aligned}$$

where 'joint' refers to the fact that the multivariate predictive $p(\tilde{y}|y, M_\ell)$ is a joint density. Here, the $T \times 1$ random variable $\tilde{y}$ is independent of the data y.

If p_{true} were known or unlimited independent replicates $\tilde{y} \sim p_{\mathrm{true}}$ were available so that (2.2) could be evaluated, the utility-maximizing model M^* could be selected by 'external validation' (Gelman et al., 2014) of the model M_ℓ joint predictive $p(\cdot|y, M_\ell)$ with respect to p_{true},

$$M^* := \arg\max_{M_\ell\in\mathcal{M}} \mathrm{eljpd}(M_\ell|y) = \arg\max_{M_\ell\in\mathcal{M}} \mathbb{E}_{\tilde{y}\sim p_{\mathrm{true}}}\left[\log p(\tilde{y}|y, M_\ell)\right]. \tag{2.4}$$

In many cases it is computationally convenient to compute (2.3) in a pointwise fashion, which yields the expected log *pointwise* predictive density (elppd),

$$\begin{aligned} \mathrm{elppd}(M_\ell|y) &= \mathbb{E}_{\tilde{y}\sim p_{\mathrm{true}}}\left[\log \prod_{t=1}^{T} \int p(\tilde{y}_t|\theta)p(\theta \mid y, M_\ell)\,\mathrm{d}\theta\right] && (2.5)\\ &= \mathbb{E}_{\tilde{y}\sim p_{\mathrm{true}}}\left[\sum_{t=1}^{T} \log p(\tilde{y}_t|y, M_\ell)\right], && (2.6)\end{aligned}$$

The pointwise predictive $p(\tilde{y}_t \mid \theta, M_\ell)$ that appears in (2.6) is simply the multivariate predictive with all but one $\tilde{y}$ element marginalized out,

$$p(\tilde{y}_t|y, M_\ell) = \int \cdots \int p(\tilde{y}|y, M_\ell) \prod_{s\neq t} \mathrm{d}\tilde{y}_s. \tag{2.7}$$

The resulting utility measure for model $M_\ell \in \mathcal{M}$ given observed data y can be computed using the model joint predictive density. We will often want to discuss both classes of expected predictive densities in a generic sense, in which case we will use the umbrella term *expected log predictive density* (elpd).

We adopt (2.4) as our benchmark for the preferred model. From this perspective, the pointwise density (elppd) is useful to the extent that it is a computationally convenient approximation of the joint density (eljpd). It is important to note that while elppd and eljpd are both useful for making comparisons against similarly-constructed measures, they are fundamentally different quantities. See, for instance, Madiman and Tetali (2010) for inequalities between joint and pointwise densities.

When observations are conditionally independent given global model parameters, it is often the case that the elppd and eljpd are close or even identical. However, under serial and other forms of dependence, this is rarely the case because the eljpd captures additional information about serial dependence of the observations not reflected by the pointwise measure.

Unfortunately, the expected utility maximization framework described above suffers a crucial drawback: p_{true} is rarely ever known in practice, and thus the elpd must be estimated purely from observed data. While one might be tempted to simply substitute $p(y \mid y, M_\ell)$ into (2.3) or (2.6), this will lead to a positively biased (over-optimistic) estimate due to model overfit (Vehtari and Ojanen, 2012; Gelman et al., 2014). Instead, we need a method for estimating elppd and eljpd using only the available data.

2.3 Cross-validation

CV is a method for estimating the elpd purely from observed data by data splitting and repeated re-fits of the model. Suppose for a moment that independent replicates of the data $\tilde{y}^{(s)} \sim p_{\text{true}}$, $s = 1, \ldots, S$, were available, and the predictive were able to be evaluated pointwise. Then the utility (2.6) under model M_ℓ could be targeted by the following Monte Carlo estimator,

$$\widehat{\text{elppd}}(M_\ell \mid y) = \frac{1}{S} \sum_{s=1}^{S} \sum_{t=1}^{T} \left[\log p\left(\tilde{y}_t^{(s)} \mid y, M_\ell\right) \right]. \tag{2.8}$$

In applications where such replicates are unavailable, CV estimators exploit the fact that the data y are distributed according to p_{true}, even if p_{true} is itself unknown. CV proceeds by repeatedly splitting the data into disjoint testing and training data subsets, estimating the model on the training set, then constructing an estimator using pointwise predictions for the testing set. The CV estimator for elppd$(M_\ell | y)$, which divides y into K test sets, can be defined as

$$\widehat{\text{elppd}}_{CV}(M_\ell \mid y) = \frac{T}{K} \sum_{k=1}^{K} \frac{1}{|\mathsf{test}_k|} \sum_{t \in \mathsf{test}_k} \log p(y_t \mid y_{\mathsf{train}_k}, M_\ell), \tag{2.9}$$

where test_k denotes the subset of y to be evaluated under the predictive, train_k denotes the subset of y to be used to train the data. The scaling factors normalize the measure to 'sum scale'. The corresponding joint measure is given by

$$\widehat{\text{eljpd}}_{CV}(M_\ell \mid y) = \frac{T}{K} \sum_{k=1}^{K} \frac{1}{|\mathsf{test}_k|} \log p(y_{\mathsf{test}_k} \mid y_{\mathsf{train}_k}, M_\ell). \tag{2.10}$$

We stress that CV schemes with multivariate test sets, like *hv*-block and K-block, can be evaluated in either a joint or pointwise fashion. In comparison, univariate schemes like LOO can only be evaluated pointwise.

The CV scheme blocking design is fully described by the triple $(K, \{\mathsf{test}_k\}_{k=1}^{K}, \{\mathsf{train}_k\}_{k=1}^{K})$. Classic LOO, for instance, has $K = T$, $\mathsf{test}_k = \{k\}$ and train_k includes all but the kth element.

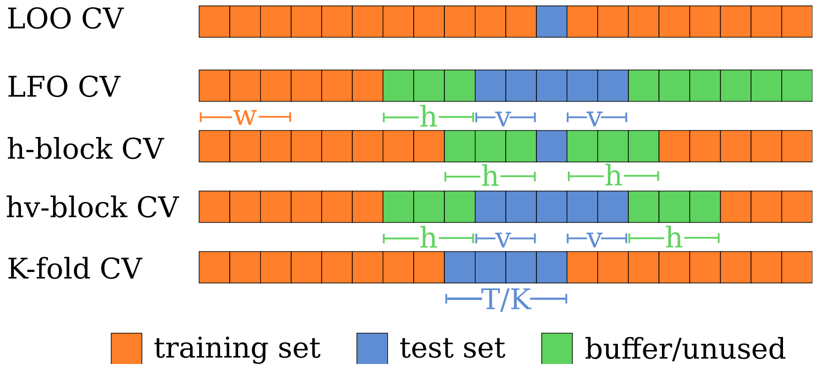


Figure 3: Cross-validation blocking schemes described in Section (2.4) for a sequence of length $T = 20$. The various schemes have hyperparameters $h = 3$ (h- and hv-block CV, LFO), $v = 2$ (hv-block CV and LFO), $K = 5$ (K-fold CV), and $w = 3$ (LFO).

Model selection using cross-validation selects the model with the greatest estimated utility—or at least the simplest model similar to the best model. For a pairwise comparison between $\mathcal{M} = \{M_A, M_B\}$, the CV estimate of the utility-maximizing objective is the sign of the difference

$$\widehat{\text{elppd}}_{CV}(M_A, M_B \mid y) = \widehat{\text{elppd}}_{CV}(M_A \mid y) - \widehat{\text{elppd}}_{CV}(M_B \mid y). \tag{2.11}$$

We have omitted from this formulation of the model selection objective the bias correction term that is sometimes included to account for the fact that there are fewer elements in the training set for each CV fold than in the full-data posterior. Typically a first-order correction is used, and it is usually very small (see Gelman et al., 2014).

Under correct model specification, the summands in the CV estimator (2.9) will usually be weakly correlated. Under relatively mild regularity conditions $\widehat{\text{elppd}}_{CV}$ should converge to the expected utility elppd as T grows large (see Bergmeir et al., 2018, for an analysis of LOO-CV, for instance).

2.4 CV schemes for serial dependence

In models of cross-sectional data where all observations can be assumed conditionally independent, the data structure imposes relatively few constraints on the sequence of training and test sets used for CV.

Under serial dependence, care is needed to ensure the contributions to (2.6) are mutually independent, or at least as independent as we can make them. To this end, a number of CV schemes have been developed specifically for models of serially dependent data.

A key consideration when selecting a CV scheme is the nature of the intended prediction task, for instance, whether the model will be used for one-step-ahead or M-step-ahead predictions.

Existing analyses of these schemes refer to specific contexts that do not necessarily align with our Bayesian framework. Most use different scoring functions and all but a few are analyzed with reference to classical models that yield point predictions. For our purpose, what we take from this earlier work is the design of the blocking scheme, i.e. the choices of $(K, \{\mathsf{test}_k\}_{k=1}^K, \{\mathsf{train}_k\}_{k=1}^K)$. These are summarized below and illustrated in Figure 3.

Burman et al. (1994) present h-block CV, an adaptation of CV for stationary dependent sequences. In order to nearly eliminate the bias arising from dependence between train and test sets in LOO-CV, their procedure deletes a buffer of size h around the training set, while retaining just a single test observation (see Figure 3). This reduces the size of the training set by $2h$ observations, but still leaves a total of n test sets. They propose h be a fixed proportion of the data length. Although this is 'conservative' (the authors refer to Györfi et al. (1989), whose results allow consistency if $h/T \longrightarrow 0$ so long as certain conditions on the underlying dependence structure are met), they argue this is appropriate because in practice the dependence structure is typically unknown. Since each h-block fold has a single test element, it is by definition a pointwise CV framework.

Racine (2000) proposes hv-block CV as an extension of h-block CV, which increases the test set dimension from 1 to $2v + 1$. The author claims this provides selection consistency in a wider range of circumstances, including nested models, which may be of interest in the case where model identification is the goal. Since hv-block has a multivariate test set, it can be evaluated jointly.

Another blocking scheme specific to serially dependent data is leave-future-out (LFO-CV; Bürkner et al., 2020). LFO-CV trains the model only on past observations, starting with a warmup period $0 < w \ll T$, and leaves future observations unused. To be comparable to the other methods, our implementation of LFO mirrors the structure of hv-block CV (Figure 3). That is, we write $\mathrm{LFO}(h, v, w)$ for a LFO scheme that includes a halo $h \geq 0$, size parameter v, and initial buffer w. This generalized form of LFO contains the usual formulation, which is $\mathrm{LFO}(0, 0, w)$.

We have also included K-fold CV. This method is not specific to serial dependence, although it is commonly applied under serial dependence in the literature (Cerqueira et al., 2020; Bergmeir and Benítez, 2012; Bergmeir et al., 2018). We use a variant of K-fold CV that partitions the sample into K contiguous sub-blocks each roughly of size T/K. Typical values for K are 5 or 10. The predictive may be evaluated jointly or pointwise, depending on the context.

3 A model selection experiment

In this section, we illustrate the behavior of CV model selection under serial dependence by repeatedly performing a model selection experiment on simulated data. We have chosen this experiment because we believe it is illustrative of general behaviors of CV for autoregressive models. We use a sequence of experiments, where we control the degree of serial dependence.

Consider the following model selection problem. Let the vector $y = y_1, \dots, y_T$ be distributed according to an $\mathsf{ARX}(2,3)$ of the following form,

$$\text{DGP}: \; y_t = \alpha \begin{pmatrix} 0.75 \\ 0.2 \end{pmatrix}^\top \begin{pmatrix} y_{t-1} \\ y_{t-2} \end{pmatrix} + \beta_{*1} + \beta_{*2} z_{2t} + \beta_{*3} z_{3t} + \sigma_* \varepsilon_t, \tag{3.1}$$

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The fixed parameter $\alpha \in [0,1]$ allows us to select the degree of serial dependence. When $\alpha = 0$ the observations are mutually independent, and $\alpha = 1$ generates a highly persistent series for the model in (3.1). All $\alpha \in [0,1]$ generate stationary series. While the upper bound $\alpha = 1$ is arbitrary, corresponding to $\phi_* = (0.75, 0.2)$, it is a useful upper limit for our experiments that generates a persistent but nonetheless stationary series. We observe similar results for other choices of the upper bound and ratio between elements of ϕ. For simplicity, we fix initial conditions $y_t = 0$ for $t \leq 0$.

The experiment selects between two candidate models:

$$M_A: \; y_t = \phi_1 y_{t-1} + \beta_1 + \beta_2 z_{2t} + \varepsilon_t \qquad \mathsf{ARX}(1,2) \tag{3.2}$$

$$M_B: \; y_t = \phi_1 y_{t-1} + \beta_1 + \varepsilon_t \qquad \mathsf{ARX}(1,1). \tag{3.3}$$

We choose the analytically-convenient (although non-conjugate) prior,

$$\beta | \sigma^2 \sim \mathcal{N}\big(\beta_*, \sigma^2 I_p\big), \quad \sigma^2 \sim \mathcal{IG}(a_0, b_0), \quad \phi \sim \mathcal{BE}_{(-1,1)}(c_0, d_0), \tag{3.4}$$

where $a_0 = b_0 = c_0 = d_0 = 1$. $\mathcal{IG}$ denotes the inverse-gamma density and $\mathcal{BE}_{(-1,1)}$ the beta distribution scaled to have support on $(-1,1)$.

This model is 'fully Bayesian' is the sense that we regard all three parameters (β, σ^2, and ϕ) as unknown, and we allow them all to be estimated. For computational tractability, we have chosen conjugate priors for β and σ^2, and we will conduct inference only on stationary $\mathsf{ARX}(1,q)$ models. In this special case, ϕ is univariate with support on the interval $(-1,1)$. We center the β prior on the truth β_* to avoid distortions as the effective sample size changes with α.

Both candidate models are 'misspecified' in the sense that neither has the same functional form as (3.1). However, while both models omit the second y_t lag and effect β_3, the candidate M_A is nonetheless the better model in the sense that it is closer in KL divergence to the DGP. This is because it includes β_2, which is also omitted by M_B.

We will work with two vectors of true DGP parameters β_*, distinguished by the relative ease with which CV is able to select the better model in our experiments:

$$\text{'easy' case: } \beta_*^{\text{easy}} = (1,2,1), \qquad \text{'hard' case: } \beta_*^{\text{hard}} = \left(1, \frac{1}{2}, 1\right).$$

These arbitrary parameter values were chosen for convenience in the context of our experiments and simulated covariates. β_*^{easy} is an example of the case where CV has little difficulty separating M_A and M_B under logarithmic loss, and under β_*^{hard} model identification by CV is much more challenging. Naturally, the relative difficulty of β_*^{easy} and β_*^{hard} depends on a range of factors, including the covariates z_{it}, noise variance, and data length. This setup does not make any assumptions about the distribution of the matrix Z with elements z_{it}, other than that it is known. In our simulations, we have drawn $z_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, which remain fixed throughout.

Although the posteriors associated with the above candidate models have no closed form, we are able to estimate them relatively computationally cheaply using one-dimensional quadrature or MCMC; see Appendix E for the complete computational details (Cooper et al., 2024). The availability of efficient estimation procedures is important for our experiments because we perform CV by brute force for each simulation draw. That is, we avoid computational shortcuts like importance sampling to ensure our results are not being driven by approximation error.

Figure 2 summarizes the results of this model selection experiment. It plots model selection statistic for 500 independent simulated data sets for the 'hard' model variant, comparing model selection statistics from two variants of 10-fold CV. The vertical axis shows the CV objective evaluated pointwise, that is $\widehat{\text{elppd}}_{\text{CV}}(M_A, M_B|y)$, and the horizontal axis shows the objective evaluated jointly, that is $\widehat{\text{eljpd}}_{\text{CV}}(M_A, M_B|y)$. Adverse selection for both measures is indicated by negative selection statistics, that is those that would selection M_B. Points lying in the first quadrant, for instance, represent correct selection by both joint and pointwise methods.

Some stylized facts are evident in the Figure 2. First, when data are mutually independent (the dependence parameter $\alpha = 0$, left panel) the marginal distribution and relative performance of both methods is approximately equal. However, as $\alpha \to 1$ we see that the variance of the pointwise estimates grows sharply and the location of the distribution shifts in a negative direction, indicating a sharp increase in the adverse selection rate. In contrast, the joint estimates are little changed as α varies. Further note that many LOO estimates fall in the fourth quadrant, indicating that the wrong model would be selected in those cases.

An equivalent experiment with the 'easy' parameter variant (not shown) displays a similar increase in the variability of the pointwise model selection statistic, but because the entire distribution lies far enough from the x-axis there is no appreciable increase in the adverse selection rate.

Figure 4 offers one explanation for the rising variance of the pointwise CV statistic. It plots the true elppd and eljpd for the 500 posteriors in Figure 2, i.e. the underlying quantities that the joint and pointwise CV estimators are targeting. The change in distributions evident in both figures is quite similar, suggesting that it is a difference between the underlying quantities, rather than some problem with the CV estimators under serial dependence, that is driving the differences between joint and pointwise 10-fold CV.

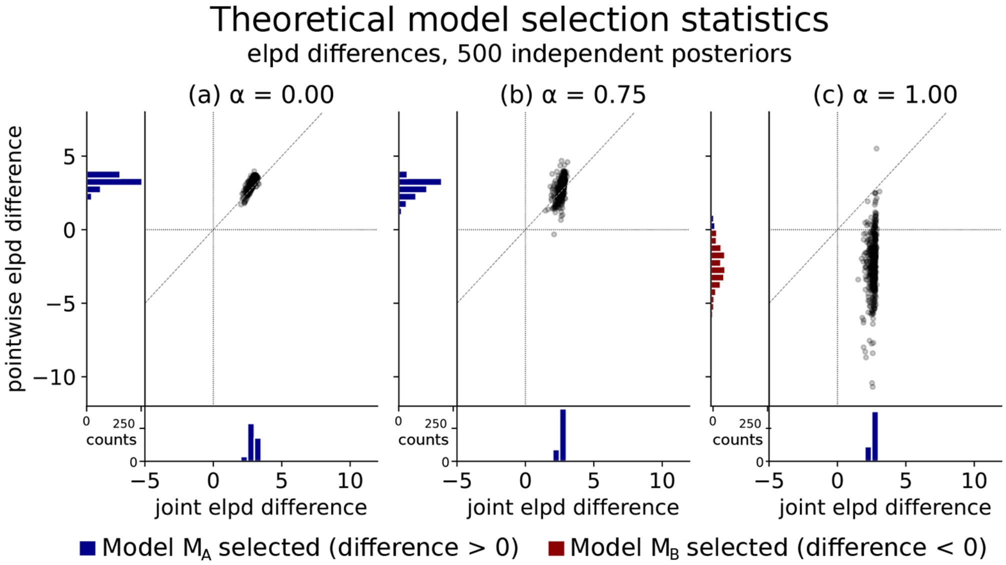


Figure 4: Joint and pointwise theoretical model selection statistics under log score for 500 independent posteriors, as dependence increases from left to right. This is the 'hard' example described in Section 4: the DGP is an $\mathsf{ARX}(2,3)$ and the candidates are M_A : $\mathsf{ARX}(1,2)$ and M_B : $\mathsf{ARX}(1,1)$. Positive values select the M_A, the better model. The gray line is the 45 degree line where both estimators are equal. Marginal histograms are also shown, with red bars indicating adverse selection.

4 Detailed case study

In this section, we focus on the problem of identifying the regression component within the ARX class. We will work with a simplified version of the $\mathsf{ARX}(p,q)$, where we regard only the regression parameter β as random, with a Gaussian prior centered on the truth, $\beta \sim \mathcal{N}(\beta_*, \Sigma_0)$. This simplification allows us to focus our attention narrowly on the task of identifying β. It also makes available analytical expressions for the posterior distribution, as well as the distribution of the theoretical elppd and eljpd, associated CV estimators, and model selection statistics.

This approach circumvents a key challenge associated with analyzing the variability of CV procedures: in general there are no closed-form expressions for the variance of elpd measures (Bengio and Grandvalet, 2004). However, when ϕ and σ^2 are fixed, a closed form does exist. The existence of closed form expressions allows us to derive the finite-sample distribution of the elpd for all of the CV procedures we consider.

4.1 Utility measures and model selection statistic

In this simple case, we can obtain exact distributions for the joint eljpd and pointwise elppd for the ARX class with fixed ϕ_* and σ^2_*, as well as their associated CV estimators.

The results in this section extend earlier results for the elppd and LOO-CV derived by Sivula et al. (2020), in the case of i.i.d. Gaussian linear regressions with fixed variance.

For arbitrary ARX models M_A and M_B, we show that the stochastic variables $\mathrm{eljpd}(M_A \mid y)$ and $\widehat{\mathrm{eljpd}}_{\mathrm{CV}}(M_A \mid y)$ (for all of the CV schemes listed in Subsection 2.4), as well as the model selection criteria $\mathrm{eljpd}(M_A, M_B \mid y)$ and $\widehat{\mathrm{eljpd}}_{\mathrm{CV}}(M_A, M_B \mid y)$, all have generalized χ^2 distributions with parameters that depend on parameters of the DGP, the posited model, and the exogenous covariates.

Following the setup in Section 3, suppose that y is distributed as $\mathsf{ARX}(p_*, q_*)$ as described above, and suppose an experimenter posits a candidate $\mathsf{ARX}(p_\ell, q_\ell)$ model M_ℓ for y. This and other results are proven in Appendix F.

Proposition 1 (Quadratic polynomial form of utility measures)**.** *Let y be distributed according to an $\mathsf{ARX}(p_*, q_*)$ process, and let M_ℓ be the simplified $\mathsf{ARX}(p_\ell, q_\ell)$ model described in Section 2.1. Then the theoretical pointwise and joint measures* $\mathrm{eljpd}(M_\ell|y)$ *and* $\mathrm{elppd}(M_\ell|y)$ *respectively defined in* (2.3) *and* (2.6)*, as well as the corresponding CV estimates* $\widehat{\mathrm{eljpd}}_{\mathrm{CV}}(M_\ell|y)$ *and* $\widehat{\mathrm{elppd}}_{\mathrm{CV}}(M_\ell|y)$*, can be expressed as second-degree vector polynomials in y,*

$$\omega_\ell(y) = y^\top A_\ell y + y^\top b_\ell + c_\ell, \tag{4.1}$$

for nonrandom coefficients A_ℓ (a $T \times T$ matrix), b_ℓ (a T-vector), and scalar c_ℓ. The coefficients are functions of $\phi^{(\ell)}, \sigma_\ell^2, Z_\ell$, and the CV blocking scheme parameters. All are defined in Appendix D.

A reviewer pointed out the similarity between the quadratic polynomial form of (4.1) and that of the likelihood ratio test for non-nested models by Vuong (1989). Although the latter's results are frequentist and asymptotic in character, the similarity is nonetheless interesting considering that the intent of these test statistics are so similar.

4.2 'Oracle' plug-in values for fixed parameters

Our candidate models require appropriate choices for the noise variance parameter σ^2 and autoregressive parameter ϕ. Especially when the model is misspecified, it would not necessarily be optimal to use the true DGP value $\sigma_\ell^2 = \sigma_*^2$, in the sense that this choice would not produce the best possible predictions with respect to log score for the chosen model class.

Inference requires suitable choices for σ_ℓ^2 and $\phi^{(\ell)}$ that would correspond as closely as possible to the behavior of an inference procedure where ϕ and σ^2 are unknown.

Suppose some hypothetical Oracle happens to know the true DGP, and offers to select the best-performing autoregressive and variance parameters ϕ_ℓ and σ_ℓ^2 for our particular model and covariates. Naturally, this choice will be independent of any specific realization of y since we are interested in utility distributions across all potential values of y. Consider two approaches our Oracle might use for selecting these parameters. She might choose to minimize the distance (in KL divergence) between the DGP and the model K predictive. Alternatively, she could directly target the objective function by maximizing the achievable $\mathbb{E}[\mathrm{elpd}(M_\ell \mid y)]$.

Under the logarithmic loss function, it follows from Dawid (1984) that these two options represent equivalent calculations. That is, maximizing the loss function,

$$\big(\widehat{\phi_\ell}, \widehat{\sigma_\ell^2}\big) := \arg \max_{\substack{\sigma_\ell^2 \in \mathbb{R}_+ \\ \phi \in \Phi_\ell}} \mathbb{E}\big[\mathrm{elpd}(M_\ell \mid y)\big]. \tag{4.2}$$

and minimizing the expected KL divergence between the DGP and model predictive,

$$\big(\widehat{\phi_\ell}, \widehat{\sigma_\ell^2}\big) := \arg \min_{\substack{\sigma_\ell^2 \in \mathbb{R}_+ \\ \phi_\ell \in \Phi_\ell}} \mathbb{E}\Big[\mathbb{D}\big(p_{\mathrm{true}}(\tilde{y}) \,\|\, p\big(\tilde{y} \mid y, \sigma_\ell^2, \phi^{(\ell)}, M_\ell\big)\big)\Big], \tag{4.3}$$

yield the same answer. In the above $\Phi_\ell \subset \mathbb{R}^{p_\ell}$ denotes the parameter space for ϕ associated with stationary $\mathsf{ARX}(p_\ell, \cdot)$ models. In our experiments, we solve the optimization (4.3) using the Nelder-Mead algorithm.

4.3 Distribution of the model selection statistic

The purpose of conducting CV on the candidate models is to determine which candidate has the greatest predictive performance, in our case under log score. The CV model selection statistic (2.11) is the difference between the estimated scores of the two models.

The form of the model selection statistic used in this experiment is characterized by the following corollary, implied by Proposition 1.

Corollary 2 (Form of model selection objectives and CV estimates)**.** *Let y be distributed according to an $\mathsf{ARX}(p_*, q_*)$ process, and let both M_A and M_B be simplified $\mathsf{ARX}(p_A, q_A)$ and simplified $\mathsf{ARX}(p_B, q_B)$, respectively. Then the theoretical model selection statistics $\mathrm{eljpd}(M_A, M_B \mid y)$ and $\mathrm{elppd}(M_A, M_B \mid y)$, and their corresponding CV estimators $\widehat{\mathrm{eljpd}}_{\mathrm{CV}}(M_A, M_B \mid y)$ and $\widehat{\mathrm{elppd}}_{\mathrm{CV}}(M_A, M_B \mid y)$, can be expressed as second-degree polynomials in y.*

Proposition 1 and Corollary 2 imply that all of the quantities of interest follow generalized χ^2 distributions (see Definition 6 in Appendix D.4). Proposition 3 describes the mean and variance, and further states the parameters of this distribution. The associated distribution function $F_{\omega(y)}(w) = \Pr(\omega(y) < w)$ must be approximated numerically. This can be done by simulation or the method of Davies (1973).

Proposition 3 (Distribution of $\omega(y)$)**.** *Let $\omega(y)$ be a quadratic polynomial in y with coefficients A, b, and c as described in Proposition 1 and Corollary 2, where $A = A^\top$. Then $\omega(y)$ has mean and covariance:*

$$\mathbb{E}\big[\omega(y)\big] = \sigma_*^2 \operatorname{tr}(AV_*) + m_*^\top A m_* + b^\top m_* + c, \tag{4.4}$$

$$\operatorname{var}\big(\omega(y)\big) = 2\sigma_*^4 \operatorname{tr}\big(A^2 V_*^2\big) + \sigma_*^2 b^\top V_* b + \sigma_*^2 4 b^\top V_* A m_* + 4\sigma_*^2 m_*^\top A V_* A m_*, \tag{4.5}$$

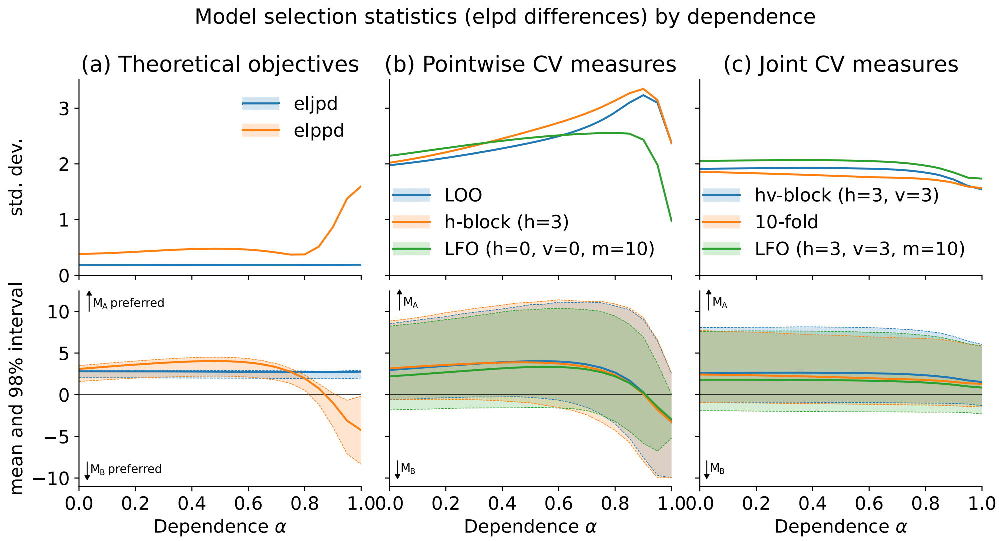


Figure 5: Model selection objectives for the 'hard' case ($\beta_* = \beta^{\text{hard}}$). Column (a) shows the theoretical model selection objectives, and columns (b) and (c) the associated CV estimates. The top row plots the standard deviation of the relevant $\omega(y)$ for the corresponding column. The bottom row shows its mean and 98% interval. The model parameter α governs the degree of serial dependence. Notice that the adverse selection rate for both joint and pointwise methods is close to zero for all but the strongest dependence. This model selection experiment compares $M_A : \mathsf{ARX}(1,2)$ vs $M_B : \mathsf{ARX}(1,1)$ under an $\mathsf{ARX}(2,3)$ DGP, as described in Section 4. The autoregressive parameter is $\phi_* = \alpha(0.75, 0.2)$, and data length $T = 100$. See also Figure 8 for the 'easy' case.

for a fixed T-vector m_* and fixed $T \times T$ matrix V_*. Moreover, it has a generalized χ^2 distribution,

$$\omega(y) \sim \tilde{\chi}^2(\boldsymbol{\lambda}, \mathbf{1}, \boldsymbol{\delta}, \mu, \sigma),$$

where $\boldsymbol{\lambda}$ is the vector of $k \leq T$ nonzero eigenvalues in the eigendecomposition $Q\Lambda Q^{-1}$ of $\sigma_*^2 L_{\phi_*}^{-1} A L_{\phi_*}^{-\top}$, $\mathbf{1}$ is a k-vector of ones, $\boldsymbol{\delta}$ is the k-vector with elements $\delta_j = \tilde{b}_j/(2\lambda_j)$, where $\tilde{b} = Q L_{\phi_*}^{-\top}(2\sigma_*^2 A m_* + \sigma_* b)$, $\mu = m_*^\top A m_* + \sigma_* m_*^\top b + c - \frac{1}{4}\sum_{j=1}^{k} \tilde{b}_j^2 \lambda_j^{-2}$, and $\sigma^2 = \sum_{j=k+1}^{T} \tilde{b}_j^2$ if $T > k$, or 0 otherwise.

Several interesting features of the distribution of $\omega(y)$ are evident in Figure 5, which summarizes the results for the 'hard' ($\beta_* = \beta_*^{\text{hard}}$) case under increasing dependence α. These features are consistent with the findings of Section 3. First, notice the striking difference between the behavior of the pointwise and joint theoretical elpds as α increases (column (a)). When data are mutually independent ($\alpha = 0$), both distributions are basically equal. For the joint measure, variability and location are little changed as dependence increases. In contrast, the pointwise objective exhibits sharply rising variability and shifts in a negative direction as dependence gains strength. Most of the distribution changes signs entirely to favor the simpler, worse candidate M_B as α approaches 1, indicating an adverse selection rate near 100%.

Second, the different profiles exhibited by pointwise and joint CV methods (columns (b) and (c), respectively). While there are clearly differences between the specific joint and pointwise methods, whether the method is computed jointly or pointwise clearly has the greatest bearing on its behavior as dependence increases. The pointwise methods in column (b) exhibit a similar increase in variability and negative shift as the pointwise objective, while the joint methods are little changed as $\alpha \to 1$. In addition, the pointwise CV methods (joint methods too, to a lesser extent) display an interesting decrease in variability as α approaches 1. A possible explanation for this drop-off is the decrease in effective sample size (Berger et al., 2014) for autoregressive models as dependence grows, all else being equal.

Under stronger dependence, increased variability and downward movement together reduce model selection power, consistent with the known bias of CV procedures toward simpler models under small sample sizes. See, for instance Burman (1989) for an analysis of CV bias and sample size. In contrast, for the 'easy' experiment variant ($\beta_* = \beta_*^{\text{easy}}$) where the results are much clearer, cross-validation generates correct model selections for all but the most strongly dependent data (see Figure 8 in Appendix A).

4.4 The cost of an inefficient CV scheme

The distribution of the model selection statistic in Proposition 3 provides a direct method for computing the probability of adverse selection. Noting that a positive selection statistic indicates the correct model choice (corresponding to M_A), it follows that

$$\Pr(\text{adverse selection}) = \Pr\big(\omega(y) < 0\big) = F_{\omega(y)}(0). \tag{4.6}$$

The quantity $F_{\omega(y)}(0)$ can be interpreted visually in Figure 5 as the share of the $\omega(y)$ distribution that falls below the x-axis. Where the probability of adverse selection is very small we say that the models are *well-separated* under that particular CV scheme. While the infinite support of $\omega(y)$ means that (4.6) can never be zero, for practical purposes we define a small threshold γ as the cutoff for well-separatedness. For the remainder of this paper, we will use $\gamma = 0.01$.

Definition 4 (Well-separated)**.** We say the CV model selection procedure defined above is well-separated at level $\gamma \in (0, 1)$ when $F_{\omega(y)}(0) < \gamma$.

A particular CV scheme may be well-separated in one situation and poorly separated in another. Whether a model selection procedure is well-separated is determined by all aspects of that procedure, including the details of the data generating process, candidate models, any hyperparameters for the procedure, and the values of the exogenous covariates.

We have seen that an inappropriate choice of CV scheme can result in an elevated probability of adverse selection. In this section, we attempt to quantify this cost.

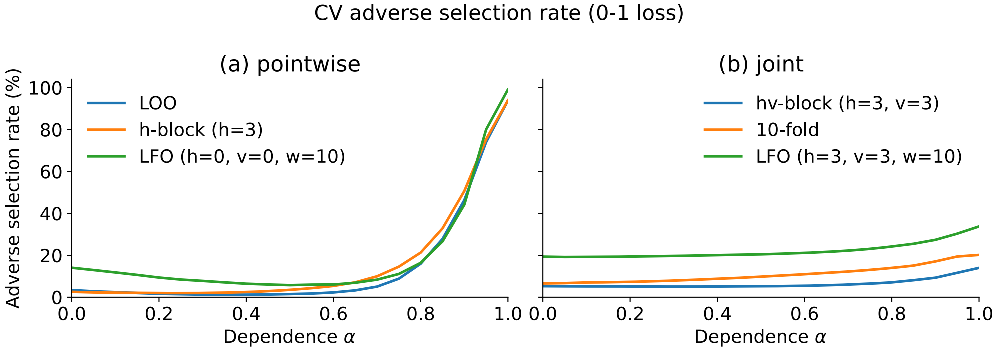


Figure 6: Adverse selection rate for pointwise and joint CV methods as data dependence increases. This experiment compares M_A : $\mathsf{ARX}(1,2)$ vs M_B : $\mathsf{ARX}(1,1)$, under an $\mathsf{ARX}(2,3)$ DGP with $\phi_* = \alpha(0.75, 0.2)$, for $\alpha \in [0,1]$ as described in Section 4. Greater values of α denote stronger serial dependence. The 'hard' case (β^{hard}) with $T = 100$ is shown.

Perhaps the simplest way to measure the cost of any model selection procedure is the rate of adverse model selection, also known as the '0-1 loss' because it scores all errors equally. The adverse selection rate is the probability (with respect to repeated samples) that the selection procedure will select the wrong model.

Framing the loss as a probability over all realizations of y makes good sense for this paper, since our focus is on the properties of CV methods for ARX models *in general*, without reference to a particular data realization y. Panels (a) and (b) of Figure 6 compare the adverse selection rate for joint and pointwise CV methods for the 'hard' variant of our model selection experiment (losses for the 'easy' variant are negligible, and are not shown). The adverse selection rate picks up as $\alpha \to 1$ and for pointwise procedures reaches almost 100 per cent, indicating that CV incorrectly prefers the simpler model when dependence is very strong.

The adverse selection rate overlooks a key fact, however: it does not account for the severity of the prediction error, scoring all incorrect selections equally. When there is very little difference between candidates' model predictions, it matters little which model is chosen. On the other hand, when predictions differ significantly this should be reflected in the cost of the error.

An alternative measure of the cost of adverse selection is the reduction in log utility that results from choosing the incorrect model for a given $y \sim p_{\text{true}}$. Under this formulation, we use the underlying finite-sample theoretical elpd *for each* y that CV is trying to estimate as the benchmark. This is by its nature a finite-sample loss function.

That is, in the case where CV does not select the elpd-maximizing model we can regard the CV utility cost relative to the counterfactual where the correct model was chosen as the difference between the chosen and maximal elpd:

$$\mathrm{cost}_{CV}(y) = \max_{M\in\mathcal{M}} \mathrm{elpd}(M|y) - \mathrm{elpd}\big(M^*_{CV}|y\big), \tag{4.7}$$

where $M^*_{CV} = \arg\max_{M\in\mathcal{M}} \widehat{\mathrm{elpd}}_{CV}(M|y)$.

Pointwise elppd and joint eljpd are of course not directly comparable. To put joint and pointwise CV measures on an equal footing, we specify the cost measure measured in terms of joint utility for both pointwise and joint CV procedures,

$$\widetilde{\mathrm{cost}}_{CV}(y) = \max_{M\in\mathcal{M}} \mathrm{eljpd}(M|y) - \mathrm{eljpd}\big(M^*_{CV}|y\big), \tag{4.8}$$

where $M^*_{CV} = \arg\max_{M\in\mathcal{M}} \widehat{\mathrm{elpd}}_{CV}(M|y)$ and $\widehat{\mathrm{elpd}}_{CV}(\cdot)$ can represent either the joint estimate $\widehat{\mathrm{eljpd}}_{CV}(\cdot)$ or the pointwise estimate $\widehat{\mathrm{elppd}}_{CV}(\cdot)$.

Figure 11 in Appendix A plots the log loss defined in the previous display. While excess log loss is an attractive concept, the resulting relative measures are practically indistinguishable from the adverse selection rate. For the remainder of this paper, we will use the adverse selection rate.

4.5 Joint and pointwise objectives

Our results suggest that under dependence, joint estimators usually have lower variability and consequently a lower rate of adverse selection than their pointwise counterparts. In our experiments, these differences are typically negligible for independent observations ($\alpha = 0$) and are most pronounced as α approaches 1 and the underlying data become highly persistent. This comparison is quite evident in Figure 6: when $\alpha = 0$, joint and pointwise estimators perform roughly equally. As $\alpha \to 1$, there is little change in performance for joint estimators, compared with a dramatic increase in the error rate for the pointwise estimators.

To construct an apples-to-apples comparison between joint and pointwise CV methods, abstracting from other details of the CV scheme design, we construct pointwise analogs to joint designs. For the present experiment this amounts to diagonalizing the covariance matrix of the Gaussian predictive distribution, setting $\sigma^2_\ell V^{pw}_\ell = \sigma^2_\ell(I_T \odot V_\ell)$, for $\odot$ the elementwise product operator. The results confirm a sharp increase in the error rate for pointwise CV methods under strong dependence ($\alpha = 1$), compared with almost no difference in the independent case ($\alpha = 0$). (See Figure 9 in Appendix A.)

4.6 Specific CV scheme design considerations

We have seen that CV design parameters can have a significant bearing on the overall efficiency, and therefore performance, of CV model selection. In this section we look more closely at hyperparameter choices for specific CV schemes for time-series data.

Note that hv-block CV schemes require the choice of a validation block size v (the total validation set dimension is $2v + 1$) and halo size h. Consistent with earlier results, for dependent data we find that by far the most important choice is to ensure that v is large enough to capture the dynamic behavior of the model. Either parameter can harm CV performance if it is too large, since both reduce training set size, which imposes a cost on statistical efficiency, resulting in a U-shaped relationship between the adverse selection rate and both h and v when dependence is present. (Figure 13 in Appendix A compares error rates with various choices of these parameters.)

The importance of preserving the size of the training set is evident in the under-performance of LFO, which represents an extreme case for dealing with contamination by discarding the entire future sample. Several authors have pointed out that this is unnecessarily conservative (Bergmeir et al., 2018). Moreover, the analyst should carefully weigh the tradeoff between the bias resulting from the use of future data and the benefit of increasing the training set size. See Figure 12 in Appendix A for a comparison between LFO and methods that use future data. In each case, the adverse selection rate for LFO is substantially higher, a consequence of LFO's reduced training set size.

4.7 Required sample size

One important practical consequence of an inefficient CV scheme design is that a larger sample size is needed for the models under consideration to be well-separated. This either requires the experimenter to collect more data than is necessary or for the adverse selection rate to be higher than it otherwise would be. In this section we demonstrate that the required sample size increases with dependence, all else being equal. When the underlying process is highly persistent, the required sample size can be significantly larger than for the independent case.

The required sample size for a model to be well-separated goes beyond the well-known principle of the effective sample size (ESS) for a time series model (see e.g. Berger et al., 2014). In this section we demonstrate that properties of the CV procedure—especially whether the scoring rule is evaluated joint or pointwise—strongly determines the data length required.

Figure 7 compares the minimum sample size required for several more joint and pointwise CV methods. Consistent with earlier results, there is little difference required sample size between pointwise and joint methods in the independent case ($\alpha = 0$). Under stronger dependence, however, the greater variability of pointwise methods leads to a requirement for larger sample sizes for the two candidate models to be well-separated, all else being equal. These results underscore the importance of using efficient CV designs—especially the use of joint scoring rules—when strong dependence is present. See also Figure 10 in Appendix A.

As we might expect, the benefit of using more efficient CV methods is greatest when candidate models are more challenging to separate. In Figure 7, required sample size is greatest for the 'hard' variant, especially under a high degree of dependence. In comparison, under the 'easy' variant the differences between joint and pointwise methods are smaller, although there is nonetheless a pickup in the sample size required for pointwise CV methods.

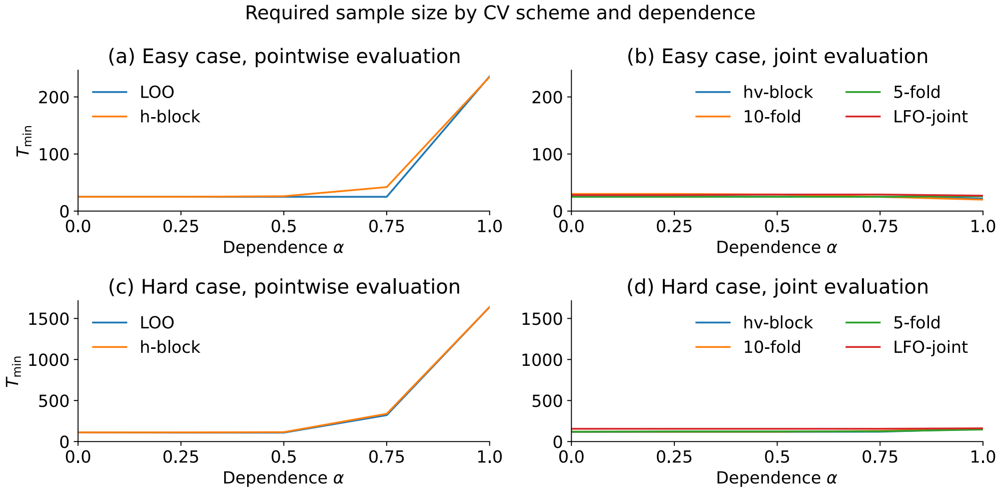


Figure 7: Minimum data length T to be well-separated at the 1% level, found by binary search over the range 10-2500. This model selection experiment compares M_A : $\mathsf{ARX}(1,2)$ vs M_B : $\mathsf{ARX}(1,1)$ under an $\mathsf{ARX}(2,3)$ DGP, as described in Section 4. The DGP autoregressive parameter is $\phi_* = \alpha(0.75, 0.2)$. See also Figure 10 in Appendix A.

Figure 7 also underscores the benefit of using as much of the available sample as possible, rather than discarding future observations as in LFO methods. As dependence increases, the additional statistical efficiency associated with allowing the model to learn from future data results in a well-separated model with shorter overall data lengths.

5 Discussion and conclusion

We have demonstrated that in settings where serial dependence is present, appropriate CV procedure design can dramatically improve model selection performance. Working with the logarithmic scoring rule and the $\mathsf{ARX}(p,q)$ class of autoregressive models, we have shown that evaluating the score pointwise can yield highly inefficient CV estimators that perform poorly when compared with procedures that target joint densities. Our experiments show that pointwise CV estimators exhibit greater variability and require larger sample sizes than joint designs.

We are not the first to compare the performance between joint and pointwise densities in predictive model assessment. Osband et al. (2022), for instance, apply model assessment with joint densities in the context of neural networks.

Our results show that the consequences of using an inefficient CV procedure can be particularly pernicious under strong serial dependence. One consequence is that CV procedures become biased toward overly simple models. In extreme cases where serial dependence is greatest, this can result in different CV procedures assigning completely different orderings to candidate models. Viewed another way, the consequence of the use of inappropriate CV procedures is the need for larger sample sizes to achieve good separation between models.

Relatively conservative methods like LFO can be a good choice in applied contexts, especially when CV is able to clearly separate models even without the use of the full sample for reducing estimator variability. Furthermore, some authors have advocated for the use of LFO (e.g., Bürkner et al., 2020) when the dependence structure is unknown. While LFO certainly eliminates the possibility that contamination will bias results, the results of Section 4.6 suggest that contamination arising from the use of future data and training set size does need to be carefully traded off against the benefit of retaining a larger training set. In general, it seems unlikely that the optimal CV design would be the corner solution that excludes all future observations. Risks associated with contamination can also be avoided by the use of specification tests on candidate models before conducting model comparison, including standard Bayesian model criticism procedures and testing for autocorrelation in the residuals (Bergmeir et al., 2018). Model assessment is especially important when the underlying process is highly persistent. Not only is the need for larger effective sample sizes greatest under persistence processes, but the adverse impact of contamination is most pernicious.

Our goal throughout this paper has been model selection. We are not claiming that exploiting future observations in CV schemes yields nearly unbiased estimators of the elpd. Instead, here we are targeting relative measures that are efficient model selection objectives.

This analysis has focused on serial dependence, which most often appears in time series models. However, we expect that similar results would apply for other forms of dependence such as spatial and spatio-temporal data, especially where the autoregressive signal is relatively strong compared with the global conditional mean. Naturally, dependence structures in more than one dimension presents additional analytical challenges, so further research is warranted for these and other dependence structures. Furthermore, many of our conclusions are not specific to the logarithmic score and would also apply under other scoring rules (Gneiting and Raftery, 2007).

From a practical standpoint, it should be noted that implementing joint CV methods tends to be computationally costly when compared with pointwise procedures like LOO. In situations where the difference between two models is absolutely clear, as for the 'easy' variant of our examples under weak serial dependence, pointwise CV estimators may be adequate for performing model selection and are far more convenient to construct. This is particularly relevant considering the availability of efficient computational shortcuts for computing pointwise CV procedures, such as PSIS-LOO (Vehtari et al., 2017; Bürkner et al., 2020), and a lack of similar shortcuts for joint procedures. Although in principle PSIS-LOO can also approximate joint CV procedures, in practice the thick tails of the weight distributions tend to cause importance sampling to fail.

With the complexity of implementing joint procedures in mind, we recommend the following workflow for model selection under serial dependence. Begin with a thorough model criticism of each of the candidate models, and iterate model specification until the candidate models are well-specified. Where inference results show that serial dependence is relatively weak, pointwise CV methods such as PSIS-LOO can be used as a first pass for model selection. If the pointwise CV results show a clear preference for one candidate, then that candidate can be selected. Otherwise, a joint CV procedure should be implemented and relied upon instead.

The present paper represents a first look at the uncertainty of CV-based model selection under serial dependence, but there is considerable work remaining. We have focused on identification of the regression parameter, leaving to one side the tasks of identifying the autoregressive component, theoretical analysis of these results, choosing suitable priors for model identification procedures, and constructing efficient computational methods for CV under serial dependence. We leave these aspects for future work.

Funding

AC's work was supported in part by an Australian Government Research Training Program Scholarship. AV acknowledges the Academy of Finland Flagship program: Finnish Center for Artificial Intelligence, and Academy of Finland project (340721). CF acknowledges financial support under National Science Foundation Grant SES-1921523. LK gratefully recognises support from the National Institutes of Health (5R01AG067149-02).

Acknowledgments

The authors are grateful for the input of two anonymous referees and an associate editor. Any remaining errors are our own.

Supplementary Material

Appendixes (DOI: 10.1214/23-BA1409SUPP; .pdf). This supplement contains additional plots and tables referenced in the main text. The supplement also contains derivations and proofs to support the experiments in the paper.

References

Arlot, S. and Celisse, A. (2010). "A survey of cross-validation procedures for model selection." *Statistics Surveys*, 4: 40–79. MR2602303. doi: `https://doi.org/10.1214/09-SS054`. 2

Bengio, Y. and Grandvalet, Y. (2004). "No unbiased estimator of the variance of K-fold cross-validation." *Journal of Machine Learning Research*, 5: 1089–1105. MR2248010. URL `https://www.semanticscholar.org/paper/17f4a82822309d4ba0e9b2840afc5dfaa499be97`. 13

Berger, J., Bayarri, M. J., and Pericchi, L. R. (2014). "The effective sample size." *Econometric Reviews*, 33(1-4): 197–217. MR3170846. doi: `https://doi.org/10.1080/07474938.2013.807157`. 17, 20

Bergmeir, C. and Benítez, J. M. (2012). "On the use of cross-validation for time series predictor evaluation." *Information Sciences*, 191: 192–213. doi: https://doi.org/10.1016{%}2Fj.ins.2011.12.028. 10

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). "A note on the validity of cross-validation for evaluating autoregressive time series prediction." *Computational Statistics and Data Analysis*, 120: 70–83. MR3742209. doi: https://doi.org/10.1016/j.csda.2017.11.003. 3, 9, 10, 20, 22

Bürkner, P. C., Gabry, J., and Vehtari, A. (2020). "Approximate leave-future-out cross-validation for Bayesian time series models." *Journal of Statistical Computation and Simulation*, 1–25. MR4145352. doi: https://doi.org/10.1080/00949655.2020.1783262. 10, 22

Burman, P. (1989). "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods." *Biometrika*, 76(3): 503–514. MR1040644. doi: https://doi.org/10.1093/biomet/76.3.503. 17

Burman, P., Chow, E., and Nolan, D. (1994). "A cross-validatory method for dependent data." *Biometrika*, 81(2): 351–358. MR1294896. doi: https://doi.org/10.1093/biomet/81.2.351. 10

Cerqueira, V., Torgo, L., and Mozetic, I. (2020). *Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods*, volume 109. Springer US. MR4172968. doi: https://doi.org/10.1007/s10994-020-05910-7. 10

Cooper, A., Simpson, D., Kennedy, L., Forbes, C., and Vehtari, A. (2024). "Supplementary Material for "Cross-validatory model selection for Bayesian autoregressions with exogenous regressors"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1409SUPP. 12

Davies, R. B. (1973). "Numerical inversion of a characteristic function." *Biometrika*, 60(2): 415–417. MR0321152. doi: https://doi.org/10.1093/biomet/60.2.415. 15

Dawid, A. P. (1984). "Statistical theory: the prequential approach." *Journal of the Royal Statistical Society. Series A*, 147(2): 278–292. MR0763811. doi: https://doi.org/10.2307/2981683. 7, 15

Geisser, S. (1975). "The predictive sample reuse method with applications." *Journal of the American Statistical Association*, 70(350): 320–328. MR0474574. doi: https://doi.org/10.1080/01621459.1975.10479865. 1

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, 3rd edition. Boca Raton, FL, USA: Chapman & Hall/CRC. MR3235677. 7, 8, 9

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. URL MR2345548. doi: https://doi.org/10.1198/016214506000001437. 2, 6, 7, 22

Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. New York, NY: Springer. MR1027837. doi: https://doi.org/10.1007/978-1-4612-3686-3. 10

Lee, D. (2011). "A comparison of conditional autoregressive models used in Bayesian disease mapping." *Spatial and Spatio-temporal Epidemiology*, 2(2): 79–89. doi: https://doi.org/10.1016/j.sste.2011.03.001. 6

Madiman, M. and Tetali, P. (2010). "Information inequalities for joint distributions, with interpretations and applications." *IEEE Transactions on Information Theory*, 56(6): 2699–2713. MR2683430. doi: https://doi.org/10.1109/TIT.2010.2046253. 7

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Lu, X., Ibrahimi, M., Lawson, D., Hao, B., O'Donoghue, B., and Roy, B. V. (2022). "The neural testbed: evaluating joint predictions." In: Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*. URL https://openreview.net/forum?id=JyTT03dqCFD 21

Racine, J. (2000). "Consistent cross-validatory model-selection for dependent data: hv-block cross-validation." *Journal of Econometrics*, 99(1): 39–61. doi: https://doi.org/10.1016{%}2Fs0304-4076{%}2800{%}2900030-0. 3, 10

Schad, D. J., Nicenboim, B., Bürkner, P. C., Betancourt, M., and Vasishth, S. (2022). "Workflow techniques for the robust use of Bayes factors." *Psychological Methods*. 2

Sims, C. A. (1980). "Macroeconomics and reality." *Econometrica*, 48(1): 1–48. 6

Sivula, T., Magnusson, M., Matamoros, A. A., and Vehtari, A. (2020). "Uncertainty in Bayesian leave-one-out cross-validation based model comparison." ArXiv preprint. arXiv:2008.10296 2, 3, 5, 6, 14

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. MR3647105. doi: https://doi.org/10.1007/s11222-016-9696-4.; URL https://doi.org/10.1007%2Fs11222-016-9709-3 22

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(1): 142–228. MR3011074. doi: https://doi.org/10.1214/12-SS102. 1, 7, 8

Vuong, Q. H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, 57(2): 307–333. 14

Ward, E. J. (2008). "A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools." *Ecological Modelling*, 211(1-2): 1–10. 2