Approximate leave-future-block-out cross-validation for Bayesian time series models

Paul-Christian Bürkner [1*], *Jonah Gabry* [2], *& Aki Vehtari* [3]

[1] *Department of Psychology, University of Münster, Germany*

[2] *Institute for Social and Economic Research in Policy, Columbia University, USA*

[3] *Department of Computer Science, Aalto University, Finland*

[*] *Corresponding author, Email: paul.buerkner@gmail.com*

Abstract

This is an online appendix to the paper *Approximate leave-future-out cross-validation for Bayesian time series models* by Paul-Christian Bürkner, Jonah Gabry, and Aki Vehtari.

Block M-step-ahead predictions

Depending on the particular time-series data and model, the Pareto k estimates may exceed τ rather quickly (after only few observations) and thus many refits may be required even when carrying out the PSIS approximation to LFO-CV. In this case, another option is to exclude only the block of B future values that directly follow the observations to be predicted while retaining all of the more distant values $y_{(i+B):N} = (y_{i+B}, \ldots, y_N)$. This will usually result in lower Pareto k estimates and thus less refitting, but crucially alters the underlying prediction task, which we will refer to as block-M-SAP.

The block-M-SAP version closely resembles the basic M-SAP *only* if values in the distant future, $y_{(i+B):N}$, contain little information about the current observations i being predicted, apart from just increasing precision of the estimated global parameters. Whether this assumption is justified will depend on the data and model. If the time-series is non-stationary, distant future value will inform overall trends in the data and thus clearly inform predictions of the current observations being left-out. As a result, block-LFO-CV is only recommended for stationary time-series and corresponding models.

There are more complexities that arise in block-M-SAP that we did not have to care about in standard M-SAP. For example, just by removing the block, the time-series is effectively split into two parts, one before and one after the block. This poses no problem for conditionally independent time-series models, where predictions only depend on the parameters and not on the former values of the time-series itself. However, if the model's predictions are *not* conditionally independent as is the case, for instance, in autoregressive models, the observations in the left-out block have to be modeled as missing values in order to retain the integrity of the time-series' predictions after the block.

Another issue concerns the PSIS approximation of block-LFO-CV: not only does the approximating model contain more observations than the current model whose predictions we are approximating, but it also may *not* contain observations that are present in the actual model. The observations right after the left-out block are included in the current model but not in the approximating model (they were part of the block at the time the approximating model was (re-)fit). A visualisation of this situation is provided in Figure 1.

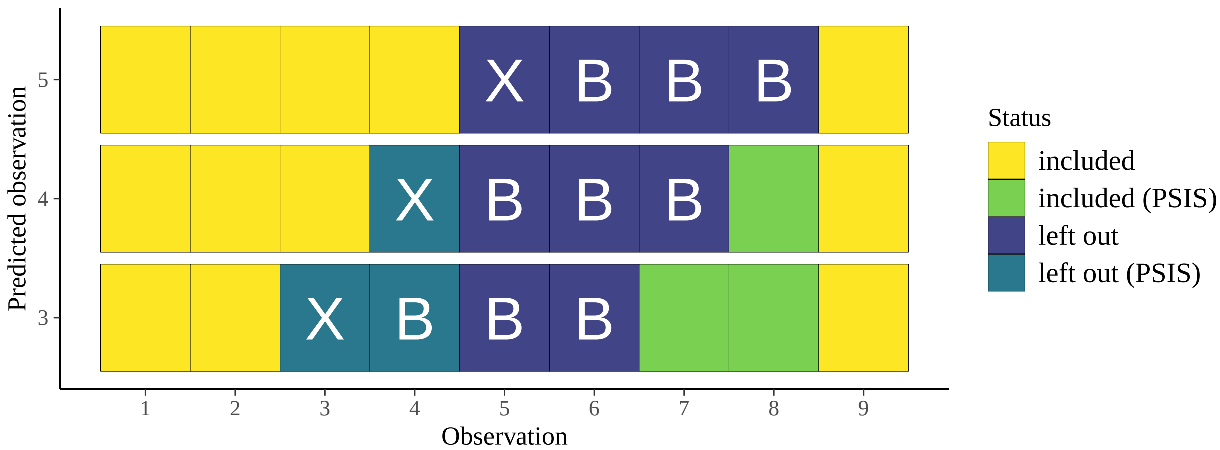

Figure 1: Visualisation of PSIS approximated one-step-ahead predictions leaving out a block of $B = 3$ future values. Predicted observations are indicated by **X**. Observation in the left out block are indicated by **B**. In the shown example, the model was last refit at the $i^\star = 4$th observation.

More formally, let $\tilde{J}_i$ be the index set of observations that are missing in the approximating model at the time of predicting observation i. We find

$$\tilde{J}_i = \{\max(i + B + 1, i^\star + 1), \ldots, \min(i^\star + B, N)\} \tag{1}$$

if $\max(i + B + 1, i^\star + 1) \leq \min(i^\star + B, N)$ and $\tilde{J}_i = \emptyset$ otherwise. The raw importance ratios $r_i^{(s)}$ for each posterior draw s are then computed as

$$r_i^{(s)} \propto \frac{\prod_{j \in \tilde{J}_i} p(y_j \mid \theta^{(s)})}{\prod_{j \in J_i} p(y_j \mid \theta^{(s)})} \tag{2}$$

before they are stabilized and further processed using PSIS (see Section **??**).

Simulations

In the simulation of block-M-SAP, we use the same conditions as for ordinary M-SAP, but instead of leaving out all future values, we left out a block of only $B = 10$ future values.

Results of the block-1-SAP simulations are shown in Figure 2. PSIS-LFO-CV provides an almost unbiased estimate of the corresponding exact LFO-CV for all investigated conditions, regardless of the threshold τ or the data generating model. The number of required refits was not only much smaller than when leaving out all future values, but practically approached zero for most conditions (see Table 1). PSIS-LOO-CV also has small bias, but the variance is larger than for PSIS-LFO-CV. This is plausible given that LOO-CV and LFO-CV of block-1-SAP only differ in whether they include the relatively few observations in the block when fitting the approximating model.

Results of the block-4-SAP simulations (see Figure 3) are mostly similar to the corresponding block-1-SAP simulations. In particular, PSIS-LFO-CV has small bias compared to the exact LFO-CV. However, the

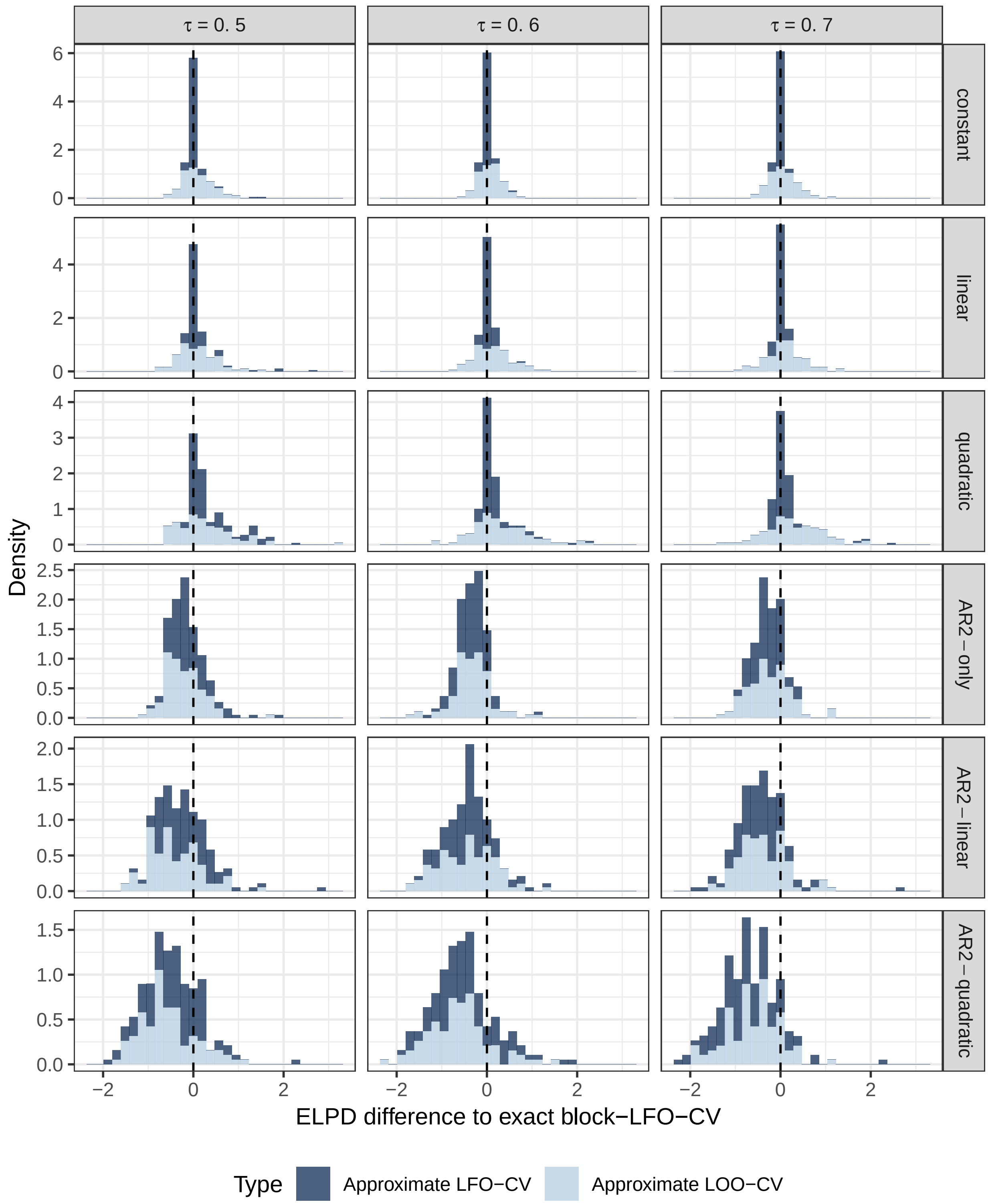

Figure 2: Simulation results of block 1-step-ahead predictions. Histograms are based on 100 simulation trials of time-series with $N = 200$ observations requiring at least $L = 25$ observations to make predictions. The number of left-out future observations was set to $B = 10$.

Table 1: Mean proportions of required refits for block-M-SAP.

M	τ	constant	linear	quadratic	AR2-only	AR2-linear	AR2-quadratic
1	0.5	0	0	0.01	0.01	0.01	0.02
	0.6	0	0	0.00	0.00	0.00	0.01
	0.7	0	0	0.00	0.00	0.00	0.00
4	0.5	0	0	0.01	0.01	0.01	0.02
	0.6	0	0	0.00	0.00	0.00	0.01
	0.7	0	0	0.00	0.00	0.00	0.00

Note: Results are based on 100 simulation trials of time-series with $N = 200$ observations requiring at least $L = 25$ observations to make predictions. The number of left-out future observations was set to $B = 10$. Abbreviations: τ = threshold of the Pareto k estimates; M = number of predicted future observations.

accuracy of PSIS-LFO-CV for block-4-SAP is highly variable when applied to autoregressive models (see Figure 3), something that is also visible in the block-1-SAP results, although to a lesser degree. This may seems to be a counter-intuitive result since the predictions should have less uncertainty in the block version, which uses more observations to inform the model. However, it can be explained as follows. In autoregressive models, predictions of future observations directly depend on past observations (they are not conditionally independent). This becomes a problem when dealing with observations that are missing in the approximating model right after the block of left out observations because the immediately preceding observations are part of the block and are thus treated as missing values (for details see Section). This implies a disproportionally high variability in the predictions for observations right after the block in autoregressive models, which then naturally propagates into the higher variability we see in the PSIS-LFO-CV approximations.

Annual measurements of the level of Lake Huron

In the following, we discuss block-LFO-CV in the context of our case study on annual measurements of the water level in Lake Huron (see Section **??**). It is not entirely clear how stationary the time-series is as it may have a slight negative trend across time (see Figure **??**). However, the AR(4) model we are using assumes stationarity and it is appropriate to also use block-LFO-CV for this example, at least for illustration. We choose to leave out a block of $B = 10$ future values as the dependency of an AR(4) model will not reach that far into the future. That is, we will include all observations after this block when re-fitting the model.

Approximate LFO-CV of block-1-SAP reveals $\mathrm{ELPD}_{\mathrm{exact}}$ = -88.54 and $\mathrm{ELPD}_{\mathrm{approx}}$ = -88.48, which are almost identical. Plotting the Pareto k estimates reveals that the model had to be refit 2 times, out of a total of $N - L = 78$ predicted observations (see Figure 4). On average, this means one refit every 39.0 observations, which again implies a drastic speed increase compared to exact LFO-CV. What is more, as expected based on our simulation results in Section we needed even fewer refits than in non-block LFO-CV. Performing LFO-CV of block-4-SAP, we compute $\mathrm{ELPD}_{\mathrm{exact}}$ = -489.01 and $\mathrm{ELPD}_{\mathrm{approx}}$ = -484.34, which are similar but not quite a close as in the 1-SAP case. Since AR-models fall in the class of conditionally *dependent* models, predicting observations right after the left-out block may be quite difficult as shown in Section . However, for this data set, the PSIS approximations of block-LFO-CV seem to have worked reasonably well.

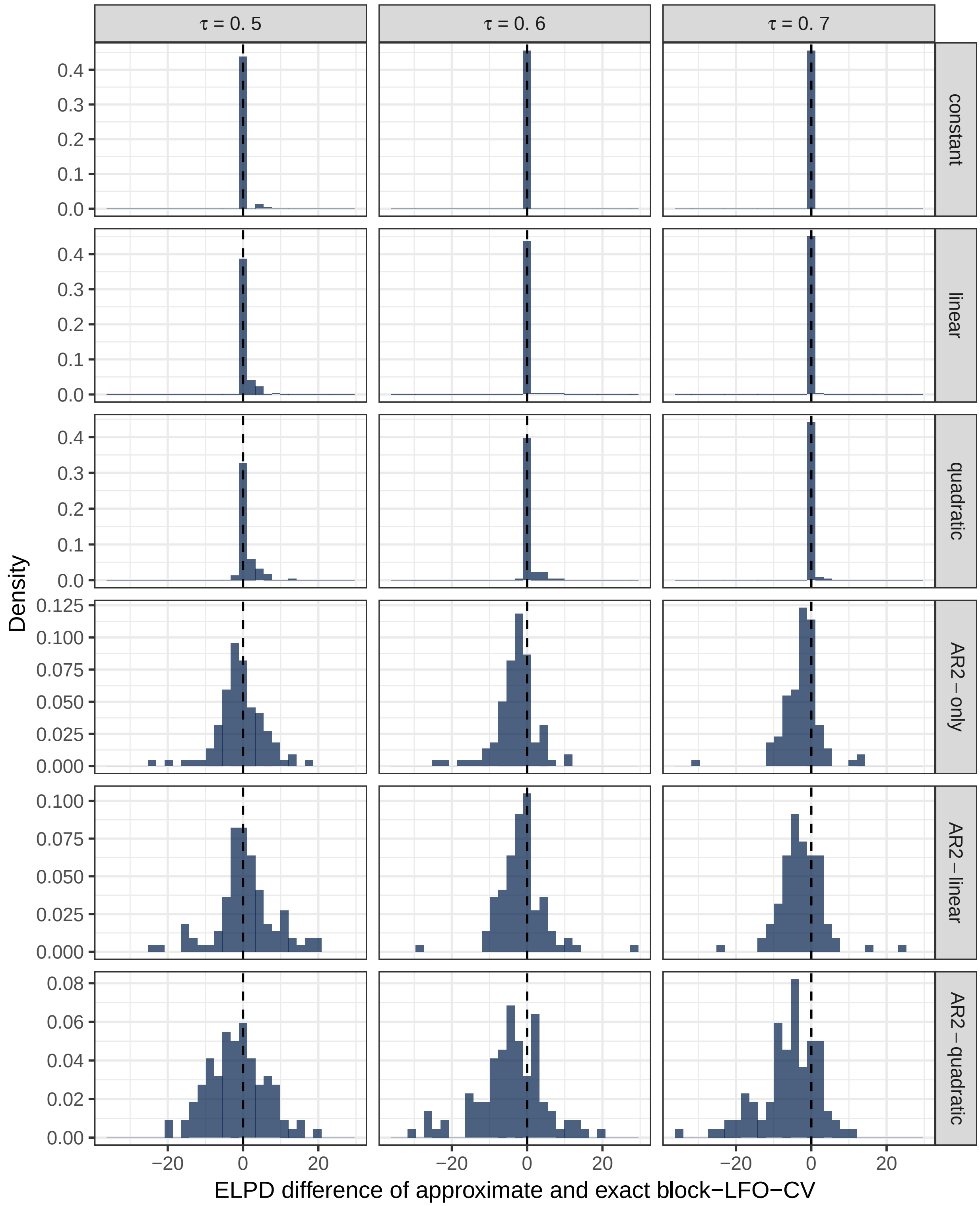

Figure 3: Simulation results of block 4-step-ahead predictions. Histograms are based on 100 simulation trials of time-series with $N = 200$ observations requiring at least $L = 25$ observations to make predictions. The number of left-out future observations was set to $B = 10$.

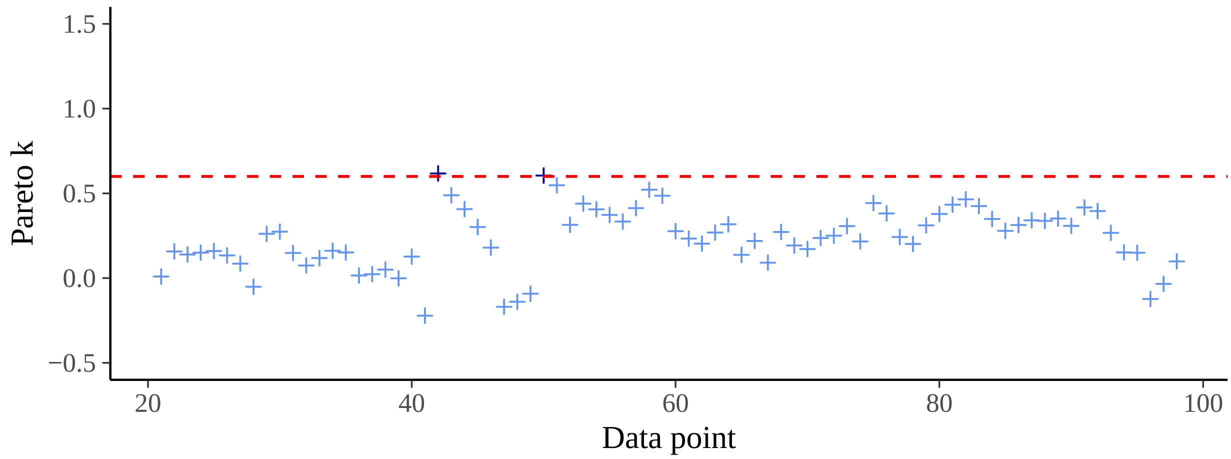

Figure 4: Pareto k estimates for PSIS-LFO-CV of the Lake Huron model leaving out a block of 10 future values. The dotted red line indicates the threshold at which the refitting was necessary.

Conclusion

Among other things, our simulations indicate that the accuracy of PSIS approximated block-M-SAP is highly variable for conditionally dependent models such as autoregressive models. Together with the fact that block-M-SAP is only theoretically reasonable for stationary time series, this leaves PSIS approximated block-M-SAP in a difficult spot. It appears to be a theoretically reasonable and empirically accurate choice only for conditionally independent models fit to stationary time-series. If the time-series is not too long and the corresponding model not too complex, so that a few more refits are acceptable, it may be more consistent and safe to just use PSIS-LFO-CV of M-SAP instead of trying approximate block-M-SAP.