Cross Validation for Time Series

Preliminary Plan – 21/11/2025

Leonardo Cartesegna - 408405
Filippo Reina - 416318
Dario Liuzzo - 408241

Introduction and motivation

In many applications, data are collected over time or space and exhibit short-range dependence. Examples include financial time series, environmental monitoring, or spatial measurements from sensor networks. Standard cross-validation (CV) procedures such as random k-fold CV or leave-one-out CV (LOOCV) assume that the training and validation samples are independent. When this assumption is violated, the validation and training sets contain highly correlated observations, which tends to underestimate prediction error and leads to overly complex (over-smoothed or under-smoothed) fits (Chu and Maaron (1991)).

The goal of this project is to study how different CV schemes behave in the presence of temporal (and possibly spatial) autocorrelation in a nonparametric regression setting. We aim to compare "naive" CV procedures with several dependence-aware variants, and to understand in which situations each method provides reliable model selection and error estimation.

Objectives

The project has three main objectives: to illustrate why standard CV procedures fail when data exhibit temporal or spatial dependence; to compare several modified CV schemes that account for autocorrelation when tuning nonparametric regression models; and to evaluate these methods both on synthetic data with controlled correlation structure and on real-world datasets.

Methodology

To evaluate our CV schemes, we will explore three distinct non-parametric settings for estimating f:

- **Penalized Splines:** Our primary classical estimator will be a smoothing spline (or penalized spline), where the amount of smoothing is controlled by a penalty parameter λ.
- **Kernel Regression:** As a second classical method, we will consider kernel regression, where the smoothness is controlled by the bandwidth h.
- **Gradient Boosted Trees (XGBoost):** To compare with a modern machine learning approach, our third setting will be an `XGBRegressor` model. Here the model's complexity (e.g., `max_depth`) will be the key tuning parameter.

For each of the non-parametric settings, we plan to compare the following CV strategies for selecting the smoothing parameter:

- **Naive CV baseline:** Random k-fold CV or LOOCV that ignores dependence. This serves as a benchmark to show the magnitude of the bias induced by autocorrelation.
- **Block CV:** The time series is partitioned into contiguous blocks, and block-wise k-fold CV is performed by leaving out entire blocks at a time.
- **Buffered CV (leave-$(2l+1)$-out):** This method removes a buffer of length l around each validation point from the training data, reducing short-range dependence. When applying this scheme in the

penalized spline setting, we will leverage the fast computational approximations from Wood (2024) to make hyperparameter optimization feasible.

- **Walk-forward (forward-chaining) CV:** Training is performed on past data and validated on future data, mimicking a forecasting setup.

For each method we will compare (i) the estimated prediction error, (ii) the smoothing parameter chosen by CV, and (iii) the resulting out-of-sample performance on an independent test set.

Data and simulation design

We will start with synthetic data, where the true regression function f and the dependence structure of the errors ε_t are known. Examples of processes we plan to consider include:

- AR(1) errors with different autocorrelation levels (e.g. $\rho = 0.2, 0.5, 0.8$),
- MA(1) errors with various parameters,
- More complex processes ARIMA,
- Processes with seasonal components.

This controlled setting will allow us to quantify bias in CV error estimates, and to examine how frequently each method chooses oversmoothed or undersmoothed fits.

Afterwards, we will analyse several real-world datasets to assess the performance of the methods in practice. Potential examples include:

- financial time series, such as daily or intraday measures of stock volatility or returns;
- environmental or climate time series, such as temperature or air pollution measurements from monitoring stations;
- if time permits, a spatial dataset (e.g. air quality or precipitation observed at multiple locations), treated with appropriate spatial versions of the CV schemes.

For these datasets we will explore how the different CV methods affect the estimated smooth trend and predictive performance, and whether the conclusions from the simulation study carry over to real data.

Expected outcomes

We expect to observe that naive CV strongly underestimates prediction error and selects overly flexible models when autocorrelation is strong, while block-based and neighbourhood-based methods provide more reliable smoothing parameter choices.

Moreover, we hope to identify practical guidelines for choosing CV schemes based on their performance in different settings.