September Modeling Forum

Cross Validation in Time Series Analysis

Chenxi Zhou
Enterprise Analytics Office, Nationwide
Department of Statistics, The Ohio State University

September 10, 2021

Outline

1️⃣ Cross validation (CV) and its applications in statistics

2️⃣ Traditional CV may fail in time series analysis 😢

3️⃣ Remedies 😃

- ➡️ Walk forward CV
- ➡️ Purged CV
- ➡️ Embargo CV
- ➡️ Combinatorial Purged CV

4️⃣ Concluding Remarks

CV and its applications in statistics

Setup

- Let $\{(X_i, Y_i)\}_{i=1}^T$ denote a stationary process[1], where Y_i is the response/target variable and X_i is the vector of feature variables, for all $i = 1, \cdots, T$.
- *h-uncorrelated Assumption*: the correlation between (X_i, Y_i) and (X_j, Y_j) is (nearly) zero if $|i - j| \geq h$ for some $h \in \mathbb{N}$.
- Let $\{(x_i, y_i)\}_{i=1}^T$ be the observations (a realization) from the process $\{(X_i, Y_i)\}_{i=1}^T$.

[1]: A process $\{Z_i\}_i$ is said to be *stationary* if $\mathbb{E}[Z_i]$ is constant for all i and $\mathrm{Cov}(Z_i, Z_j)$ depends on $|i - j|$ only.

Goal

- Build statistical/ML/DL models to
  - learn the relationship between X_i's and Y_i's, and
  - generate forecasts
- Hope the model can forecast well on the *future unseen* data

Motivations for Cross Validation

- Need tools to estimate the generalization/prediction/forecast error in order to
  - optimize the hyperparameters of a model,
  - evaluate the performance of a model on an independent dataset, and
  - select the *best* model.
- One such tool is the Cross Validation (CV)

A Review of Traditional CV

- The purpose of CV is to estimate the generalization error, so as to
  - prevent overfitting to the training data, and
  - provide insights on how the model will generalize to unseen data.
- Many variants of CV exist.

A Review of *K*-fold CV

Procedures

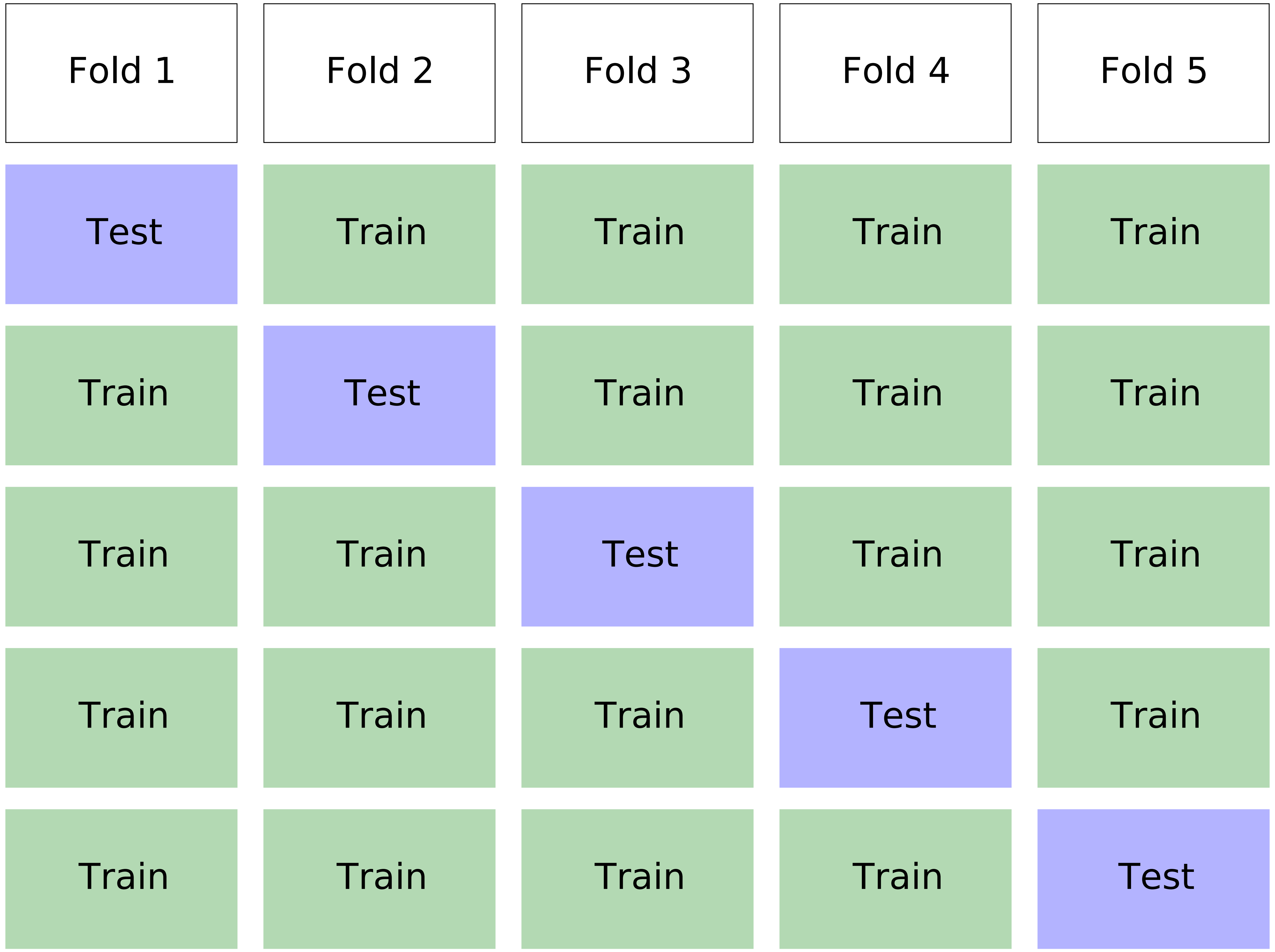
Assume (for this slide and the next two *only*) that $\{(X_i\,,\ Y_i)\}_{i=1}^{T}$ are independent and identically distributed (i.i.d).

K-fold CV works as below:

- Shuffle the entire dataset and divide it into K non-overlapping folds of roughly equal size;
- for $j = 1, 2,\ \cdots, K$:
  - Train the model on all folds excluding the j-th one;
  - Use the j-th fold as the test set and compute the generalization error of the fitted model on it.
- Average generalization errors from all K folds as the estimate of the overall generalization error.

A Review of *K*-fold CV

Illustration



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

Illustration of 5-fold CV when data are i.i.d. Each row corresponds to the entire dataset after shuffling.

A Review of *K*-fold CV

Remarks

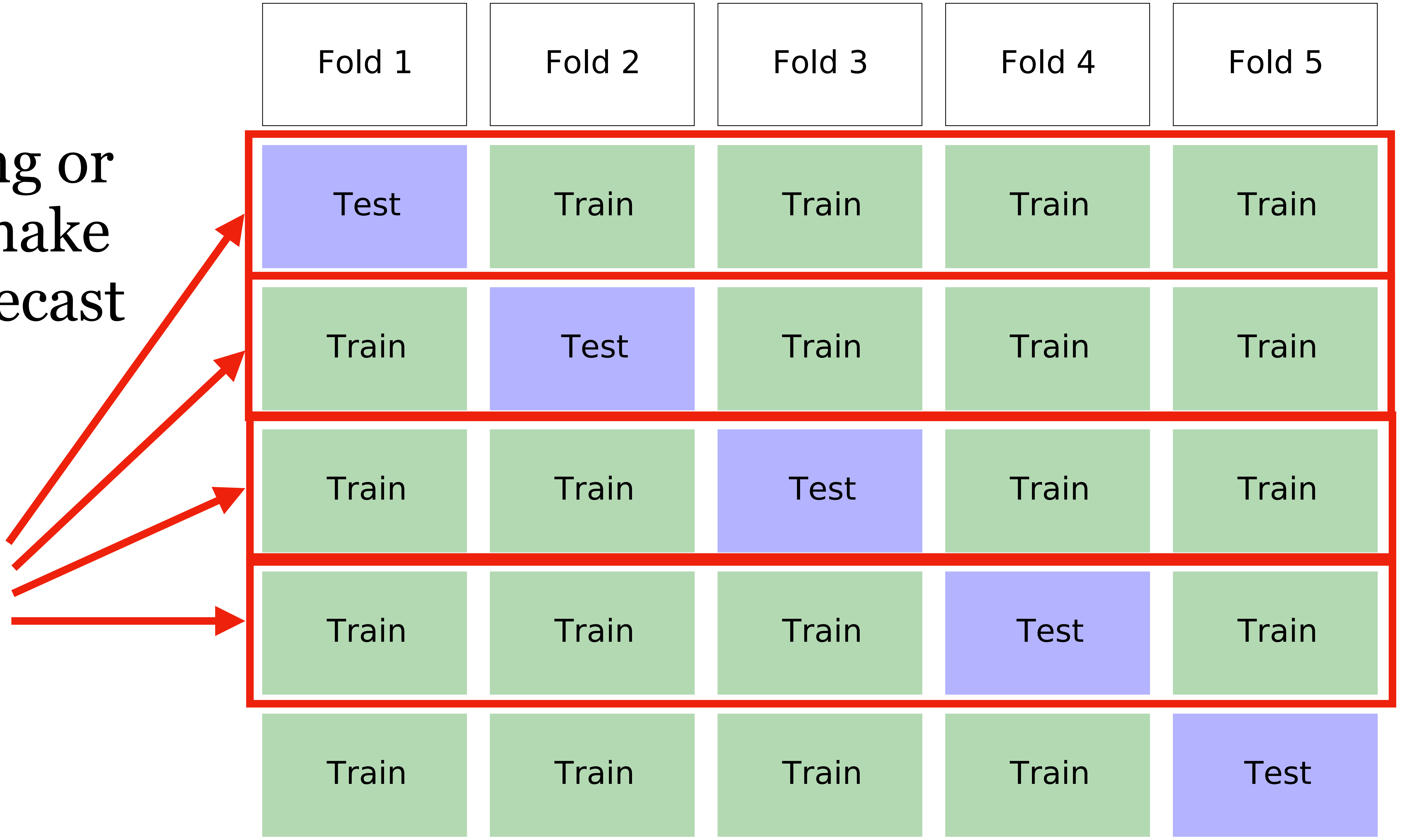
- Typical choices of *K* are 5 or 10.
- Variants of *K*-fold CV exist:
  - repeated *K*-fold CV — to remove the randomness in splitting data and reduce the variance of the estimates of the generalization error;
  - nested *K*-fold CV — to select optimal hyperparameters and evaluate the model performance on independent datasets simultaneously;
  - stratified *K*-fold CV — to ensure each fold contains approximately the same proportion of samples of each target class as the entire dataset.
  - …

Traditional CV May **Fail** in Time Series Analysis

Problem 1

Observations are *chronologically* ordered

- We cannot randomly assign observations to either training or test set as it does not really make sense to use the future to forecast the past.



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

Using the future to forecast the past

Illustration of 5-fold CV. Each row corresponds to the entire dataset where data are ordered chronologically.

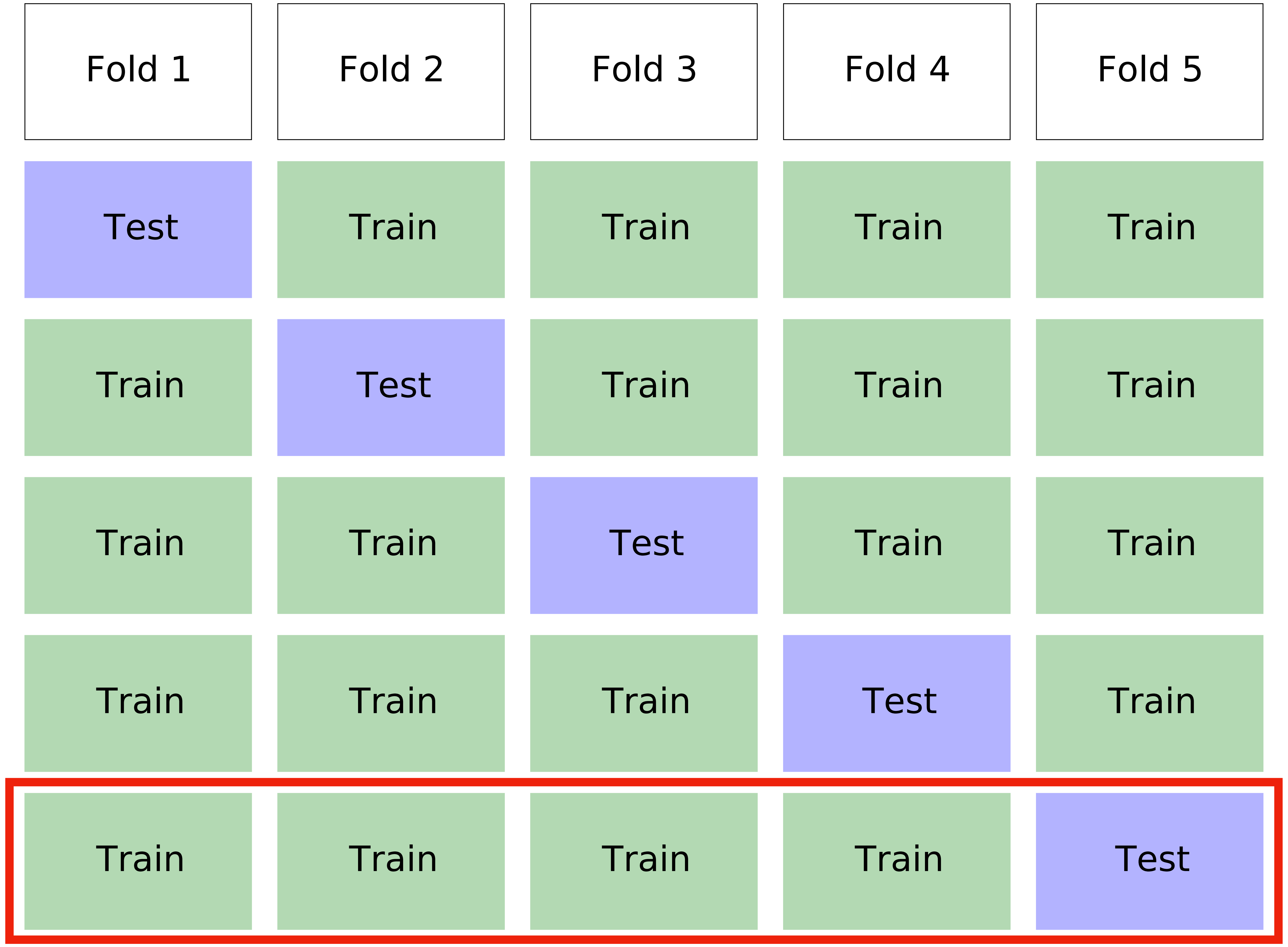
Problem 2

Existence of serial correlation

- Time series data are typically *serially correlated*, meaning (X_i, Y_i) has non-zero correlation with (X_j, Y_j), where $|i - j| < h$, under the h-uncorrelated assumption.

- This violates the i.i.d assumption in traditional CV.

If training and test sets have strong serial correlation, using traditional CV can lead to an overly optimistic estimate of the generalization error.



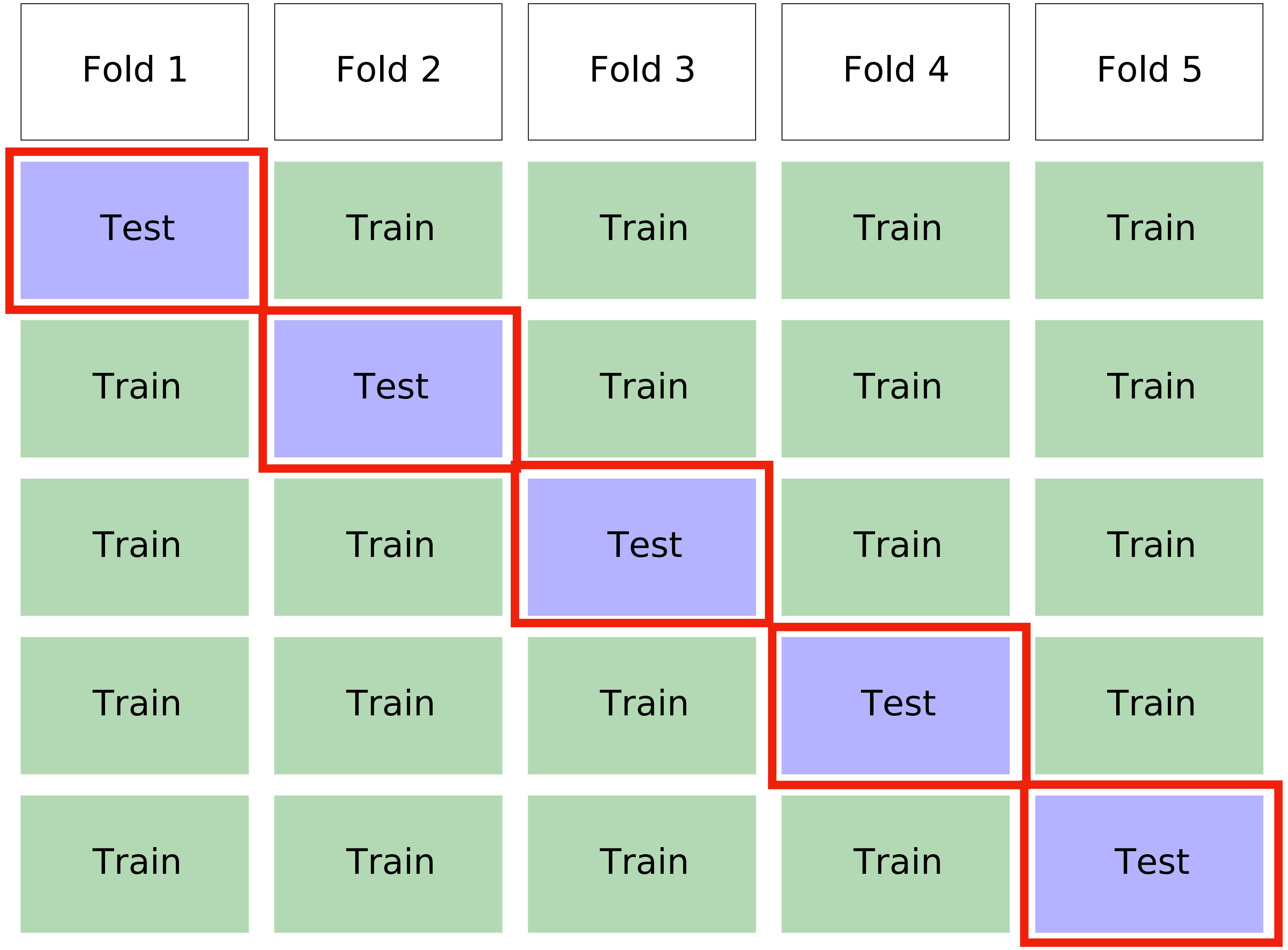
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

Illustration of 5-fold CV. Each row corresponds to the entire dataset where data are ordered chronologically.

Problem 3

Testing only one path

- Time series at hand are just **one** realization of the underlying stationary process.
- Using *K*-fold CV, we are only testing a single path, i.e., there is one and only one forecast generated for each observation.
- The future unseen data may or may not replicate the past.



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

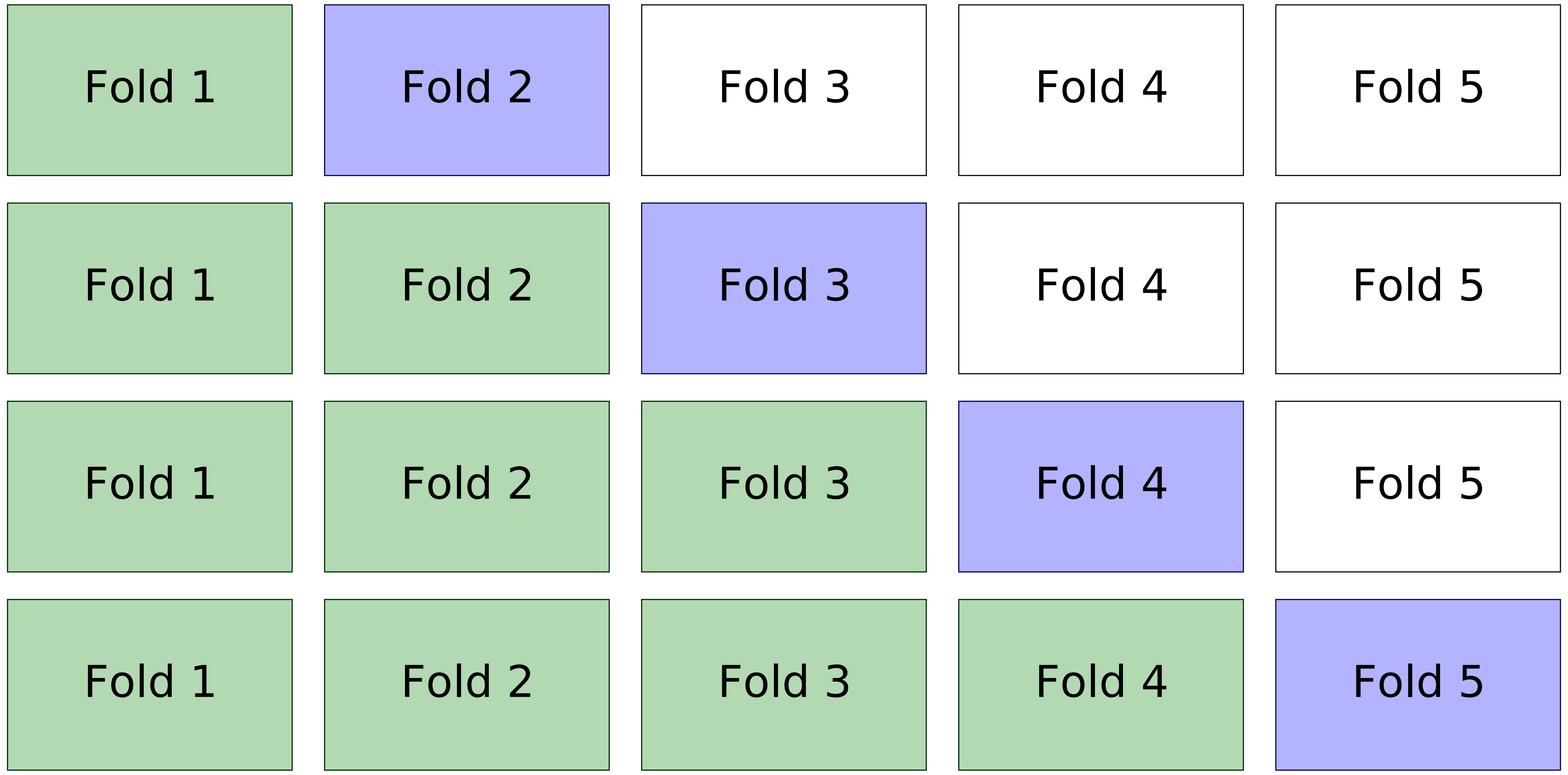
Illustration of 5-fold CV. Each row corresponds to the entire dataset where data are ordered chronologically.

Remedies

Walk Forward CV

Procedures

- Keep chronological order of observations and divide them into K non-overlapping folds of roughly equal size;
- for $j = 1, 2, \cdots, K-1$:
    - Train the model using the first j folds;
    - Use the $(j+1)$-st fold as the test set and compute the generalization error of the fitted model on it.
- Average the generalization errors in $K-1$ folds as the estimate of the overall generalization error.



🟩 Fold 1	🟦 Fold 2	Fold 3	Fold 4	Fold 5
🟩 Fold 1	🟩 Fold 2	🟦 Fold 3	Fold 4	Fold 5
🟩 Fold 1	🟩 Fold 2	🟩 Fold 3	🟦 Fold 4	Fold 5
🟩 Fold 1	🟩 Fold 2	🟩 Fold 3	🟩 Fold 4	🟦 Fold 5

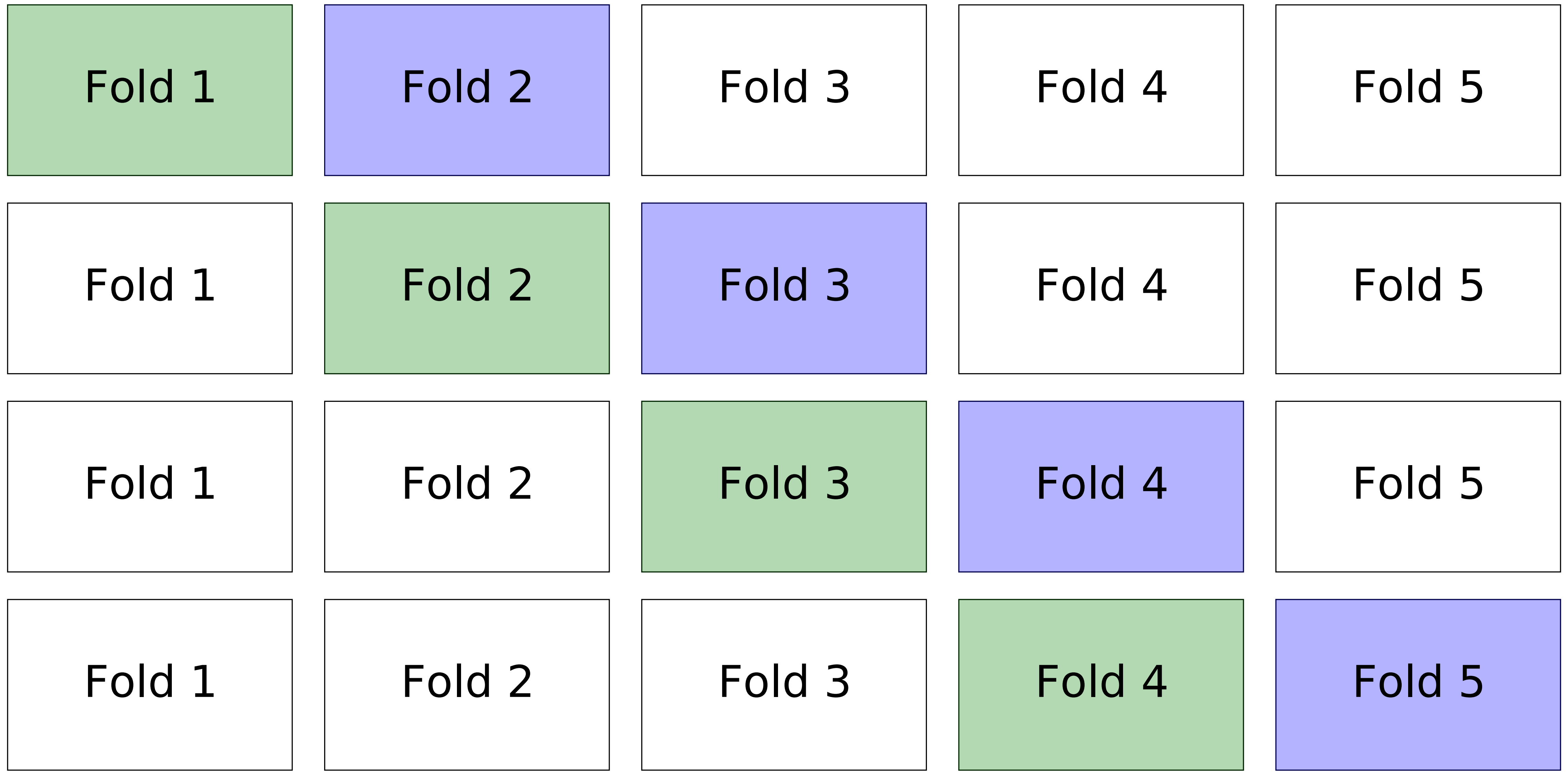
5-fold walk forward CV with an expanding window. Green boxes denote the training set, and blue boxes denote the test set.

Walk Forward CV

Procedures

Remark: The procedures in the preceding slide use an *expanding* window, i.e., the training set keeps expanding with j.

A variation is to use a *sliding* window, where the numbers of observations in the training set does *not* expand with j and is of roughly equal size for all j.



🟩 Fold 1	🟦 Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	🟩 Fold 2	🟦 Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	🟩 Fold 3	🟦 Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	🟩 Fold 4	🟦 Fold 5

5-fold walk forward CV with a sliding window. Green boxes denote the training set, and blue boxes denote the test set.

Walk Forward CV

Comments

😃 Advantage

- Problem 1 is solved: Guarantees each test set follows the training set and there is no information leakage from the future.

😟 Disadvantages

- Problem 2 persists: No adjustment for the serial correlation. The last few observations in the training set may be highly serially correlated with the first few observations in the test set.
- Problem 3 persists: Only tests a single path. Future unseen data may *not* replicate the past.
- Inefficient use of data: In the expanding window when j is small and in the sliding window, forecasts are made only based on a small set of samples, leading to biased models.

Purged CV (υ-block CV)

Motivation, Procedure and Outcome

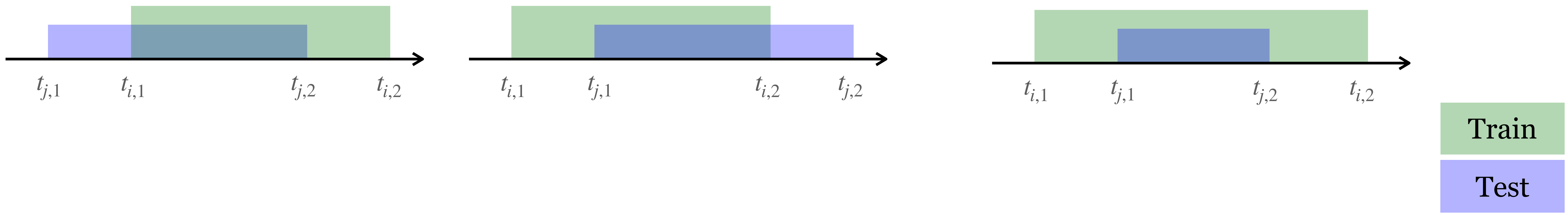
Motivation: Want to remove serial correlation between training and test sets.

Procedure: Remove υ observations in the *training set* preceding and following the test set.

Outcome: There is no *informational overlapping* between two sets.

- **Informational Overlapping:** Let $(x_i\,,\ y_i)$ belong to the training set, and $(x_j\,,\ y_j)$ belong to the test set. Suppose y_i is a function of observations between $t_{i,1}$ and $t_{i,2}$, and y_j is a function of observations between $t_{j,1}$ and $t_{j,2}$. We say there exists *informational overlapping* between y_j and y_i if one the following occurs

$t_{j,1} \leq t_{i,1} \leq t_{j,2}$, $t_{j,1} \leq t_{i,2} \leq t_{j,2}$, and $t_{i,1} \leq t_{j,1} \leq t_{j,2} \leq t_{i,2}$.



Purged CV (υ-block CV)

Illustration

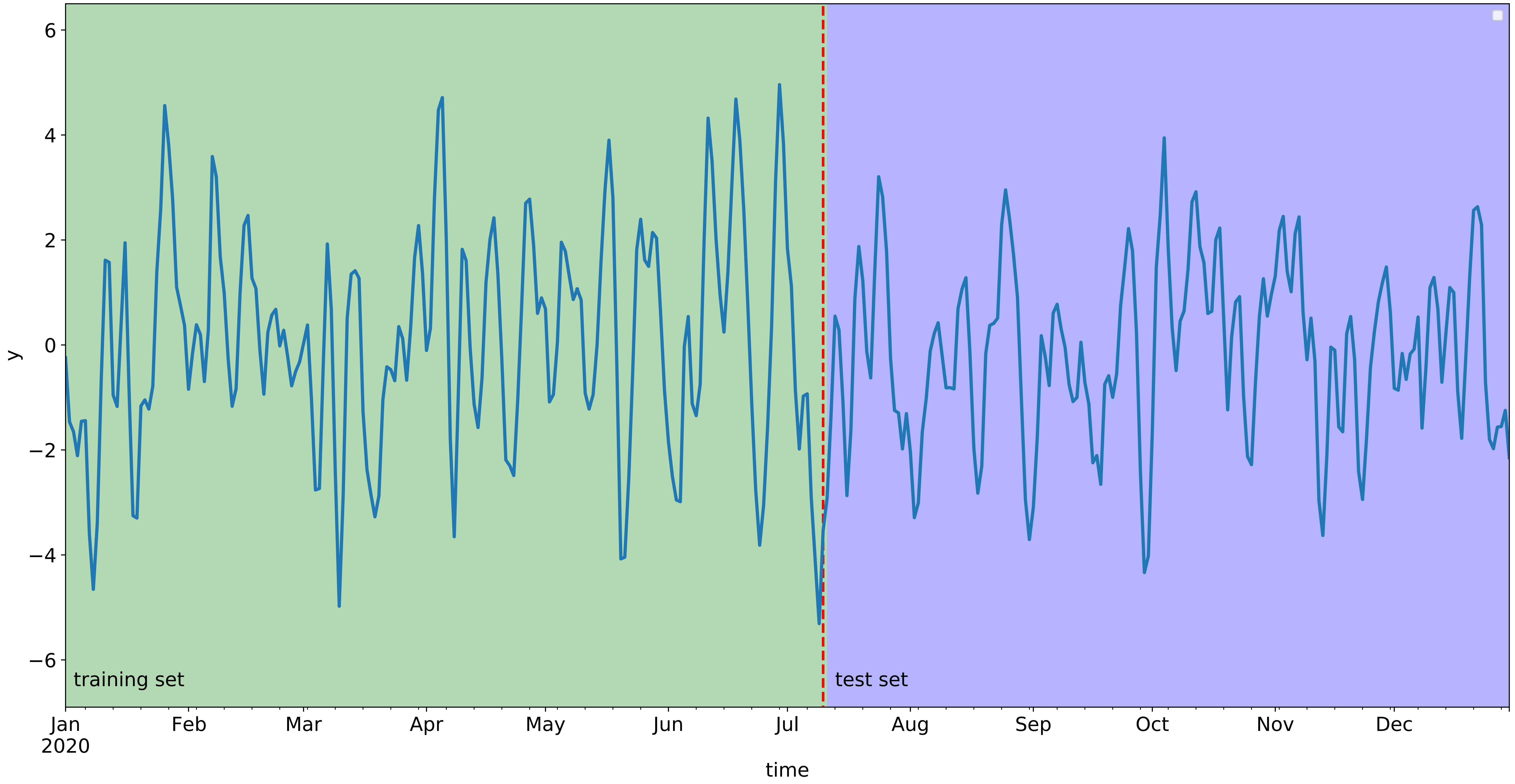


Illustration of purged CV **before** purging.

Purged CV (v-block CV)

Illustration

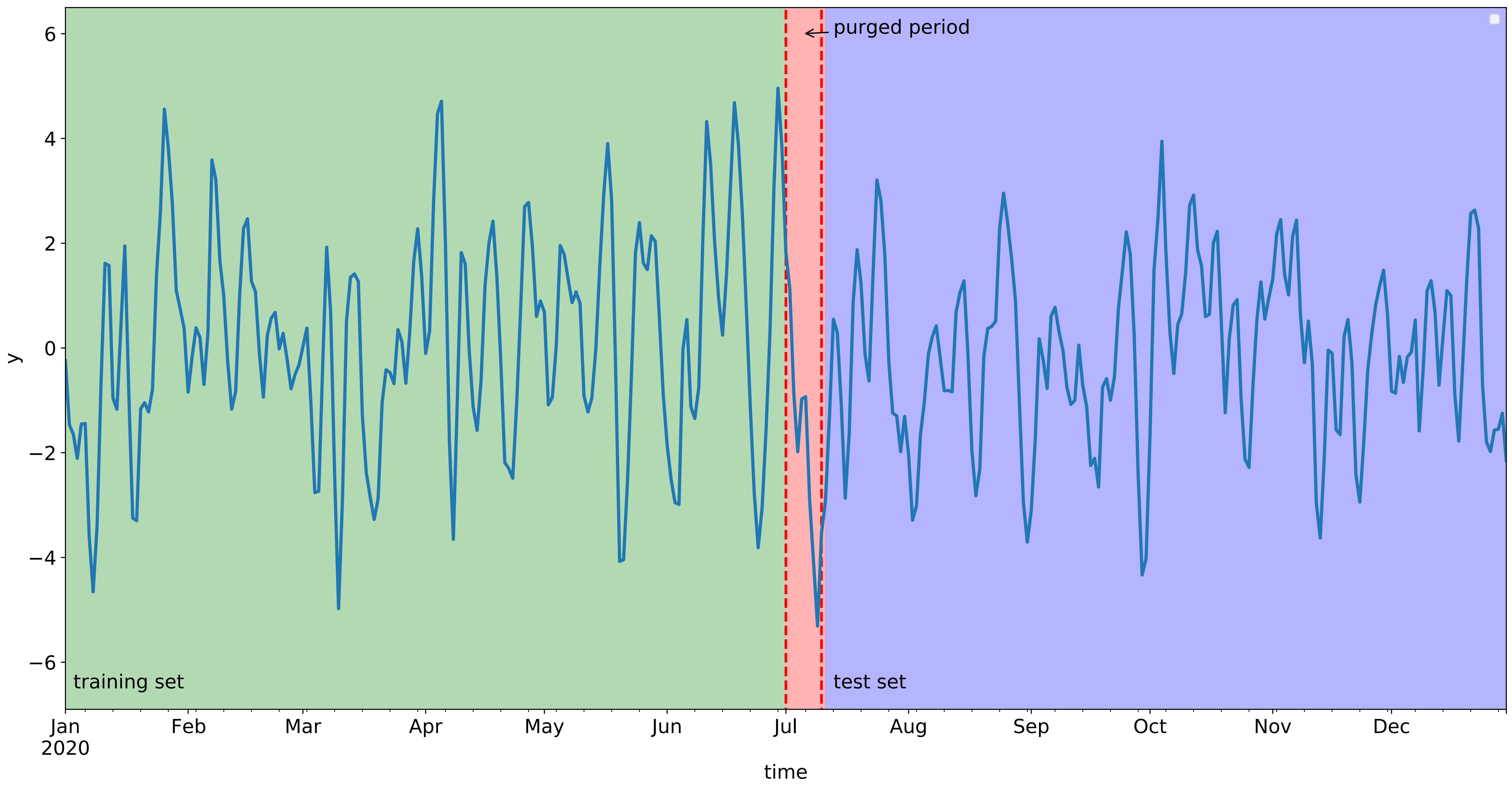


Illustration of purged CV after purging.

Purged CV (v-block CV)

How much to purge?

- The length of the purged period (i.e., the choice of v) depends on:
  - how long the feature variables span in the model assumption, and
  - the auto- and cross-correlations of feature variables and target variable.

Purged CV (*v*-block CV)

Remarks

- Using the h-uncorrelated assumption, with $v \geq h$, one can ensure that training and test sets are (nearly) uncorrelated.
- Note that the test set is not changed but the training set is shortened.

K-fold Purged CV
Motivation and Procedures

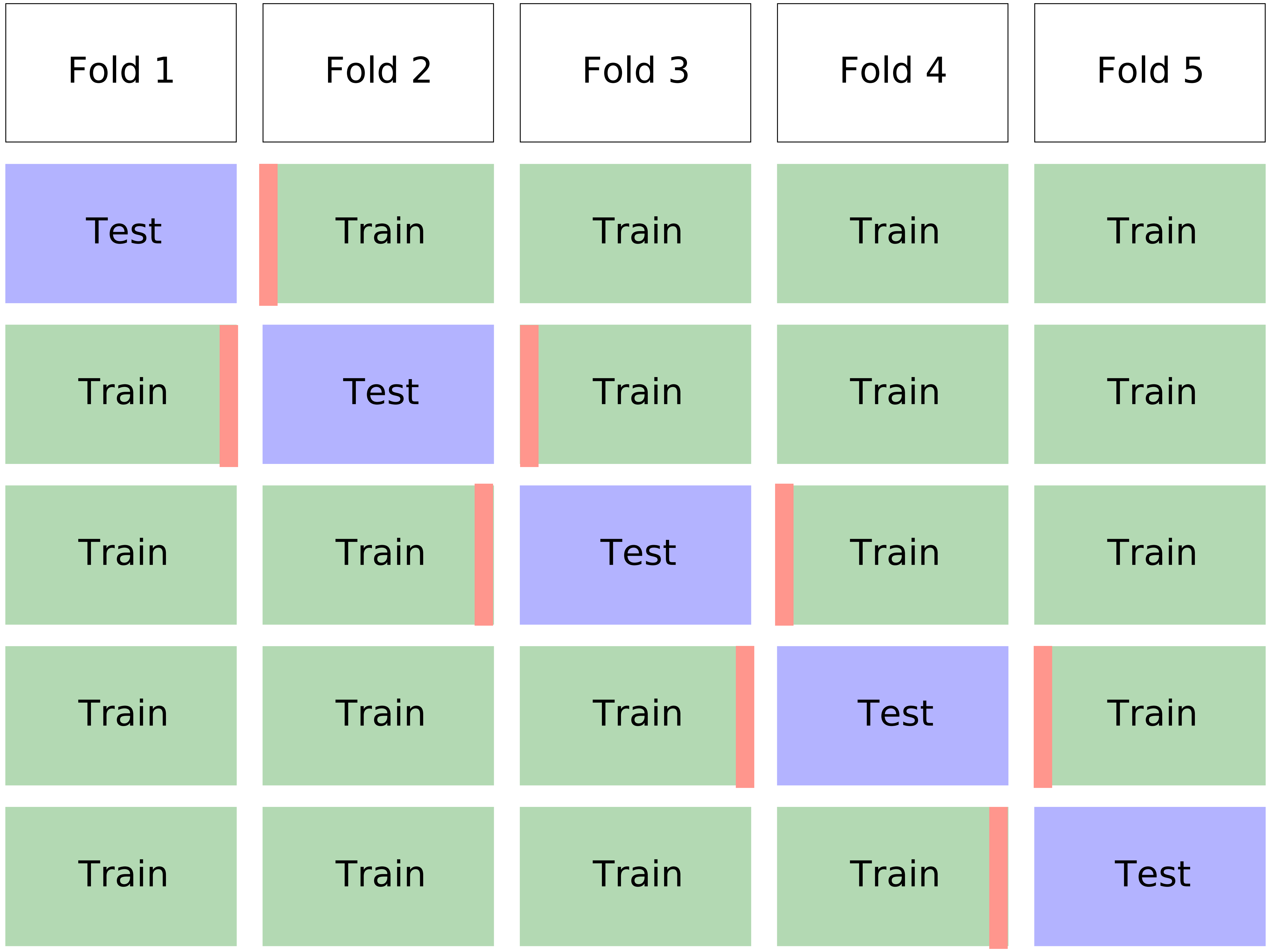
Motivation: Purged CV described so far creates only *one* split of the training and test sets. We can devise the *K*-fold version of the purged CV.

Procedures:

- Keep the chronological order of data and divide them into K non-overlapping folds of roughly equal size;
- for $j = 1, 2, \cdots, K$:
  - Choose the data on all folds excluding the j-th one as the training set, and purge it;
  - Train the model on the purged training set;
  - Use the j-th fold as the test set and compute the generalization error of the fitted model on it.
- Average the generalization errors from all K folds as the estimate of the overall generalization error.

K-fold Purged CV

Illustration



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test	🟥 Train	Train	Train	Train
Train 🟥	Test	🟥 Train	Train	Train
Train	Train 🟥	Test	🟥 Train	Train
Train	Train	Train 🟥	Test	🟥 Train
Train	Train	Train	Train 🟥	Test

Illustration of 5-fold purged CV. The green boxes correspond to the training sets, and the blue boxes correspond to the test sets. The red regions correspond to the purged data.

Embargo CV

Motivation and Procedure

- **Motivation:** In the case where the training set may come *after* a test set, want to avoid peeking ahead into the future. For example,
  - in the purged *K*-fold CV, and
  - in the scenario analysis or sensitivity analysis (thinking about using the model constructed from the data in the normal period to forecast the target variable in a highly volatile period, and the normal period may come after the volatile period).
- **Procedure:** Set an embargo period and remove all observations belonging to this period.

Embargo CV

Illustration

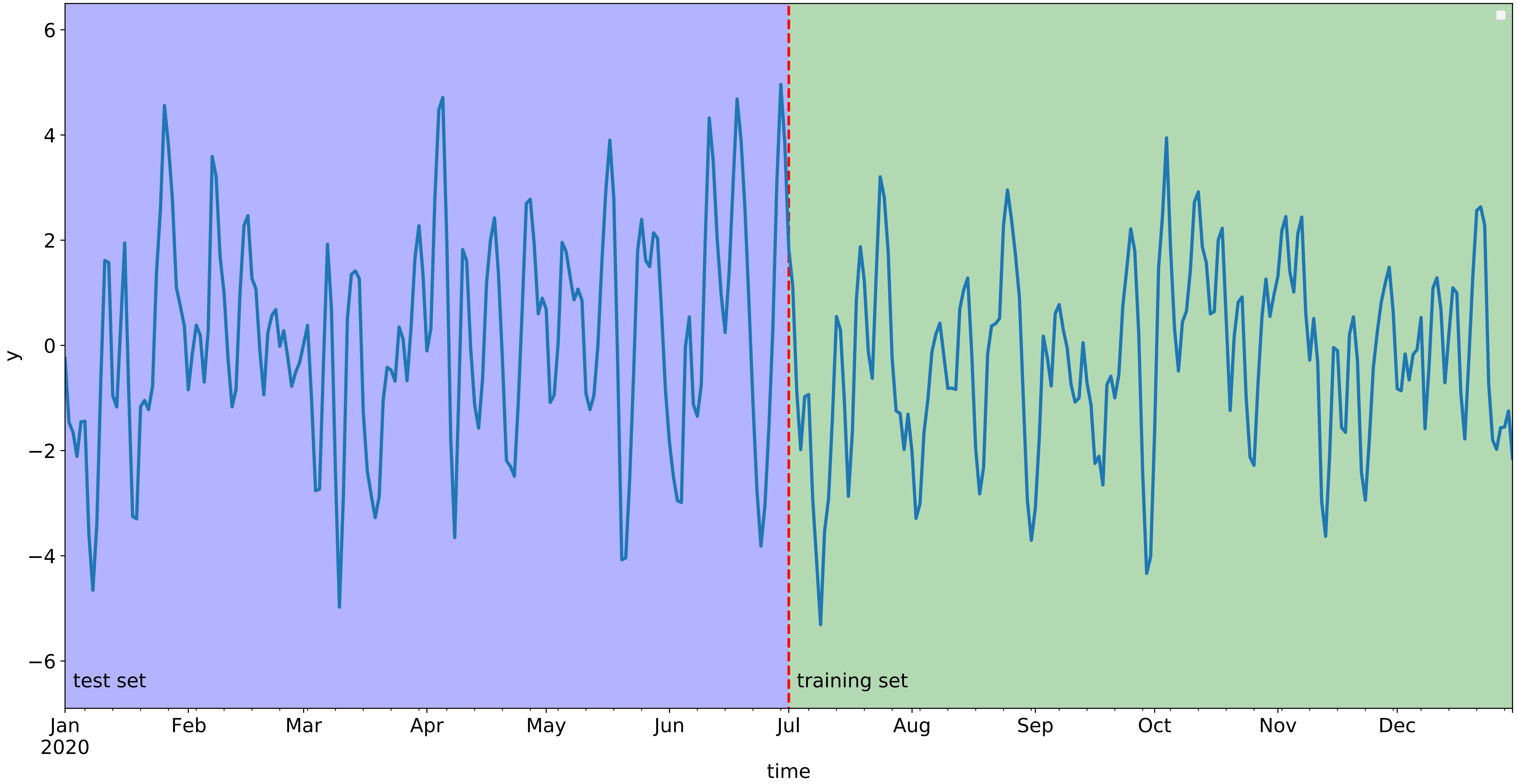


Illustration of embargo CV **before** embargoing.

Embargo CV

Illustration

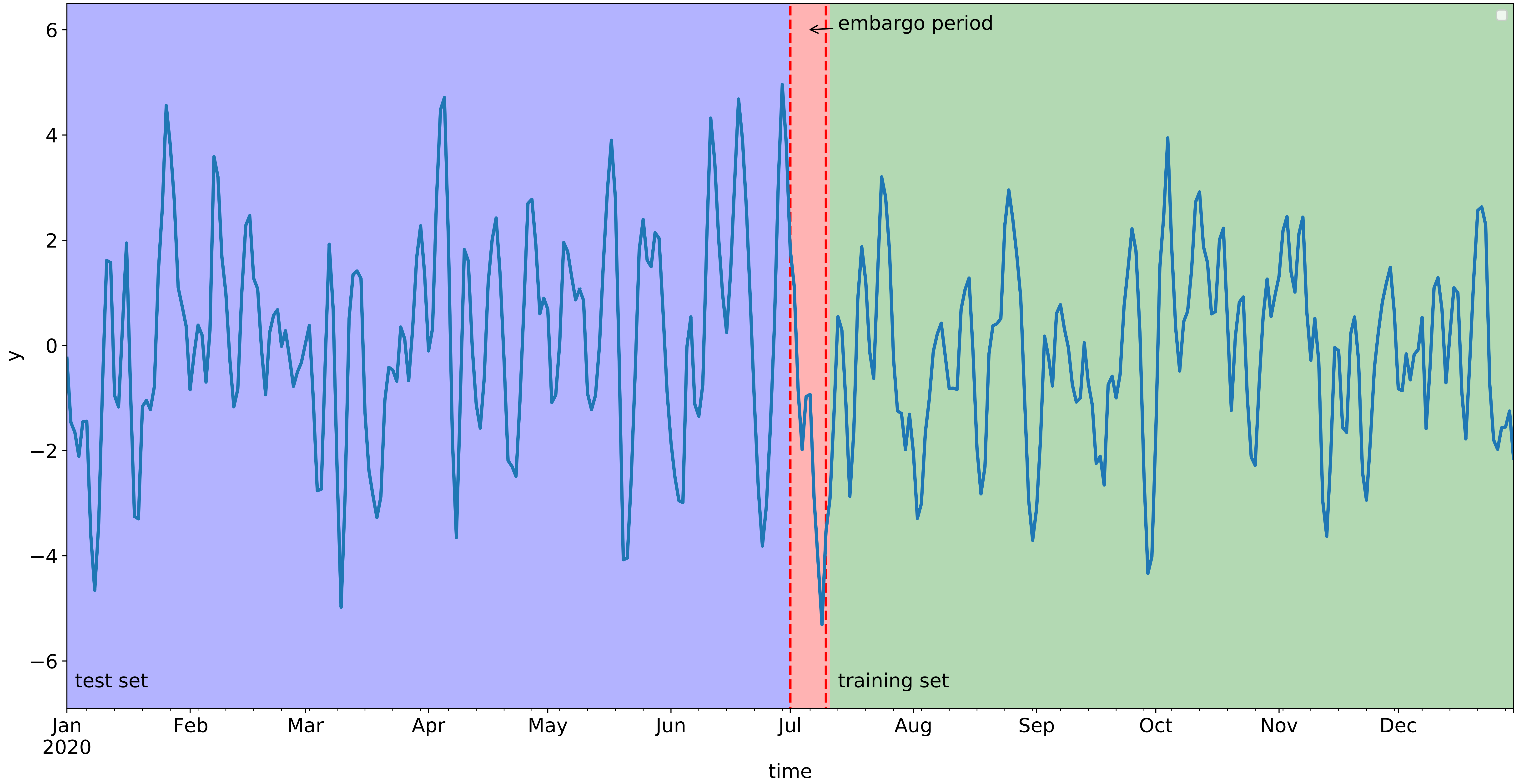


Illustration of embargo CV **after** embargoing.

Embargo CV

How much to embargo?

- The length of the embargo period depends on:
    - how long the feature variables span in the model assumption, and
    - the auto- and cross-correlations of feature variables and target variable.

Embargo CV

Remarks

- Embargo CV is *only* used when the training set comes *after* the test set.
- The test set is *not* changed but the training set is shortened.

Combining Purged CV and Embargo CV

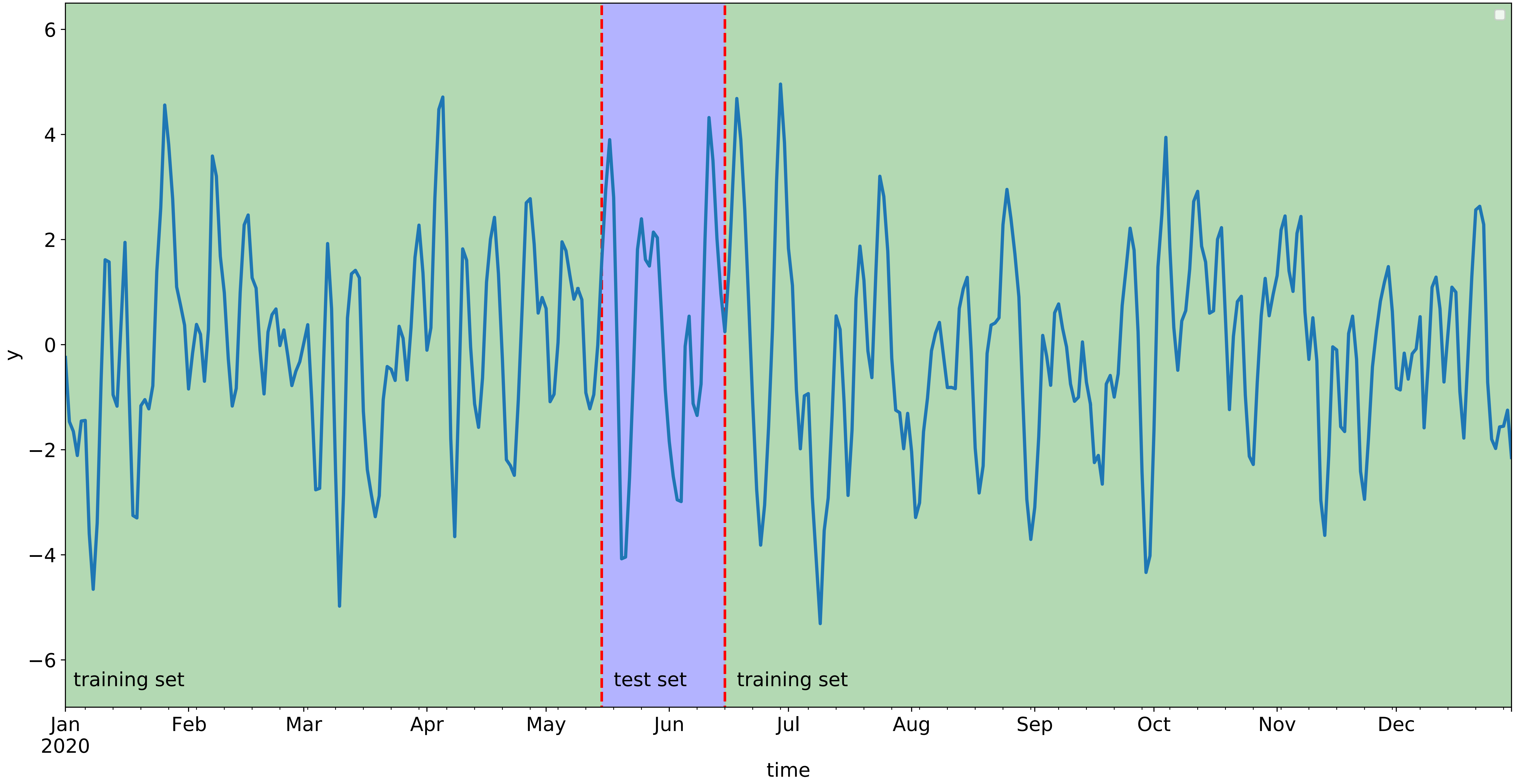


Illustration of combining purged CV and embargo CV.

Combining Purged CV and Embargo CV

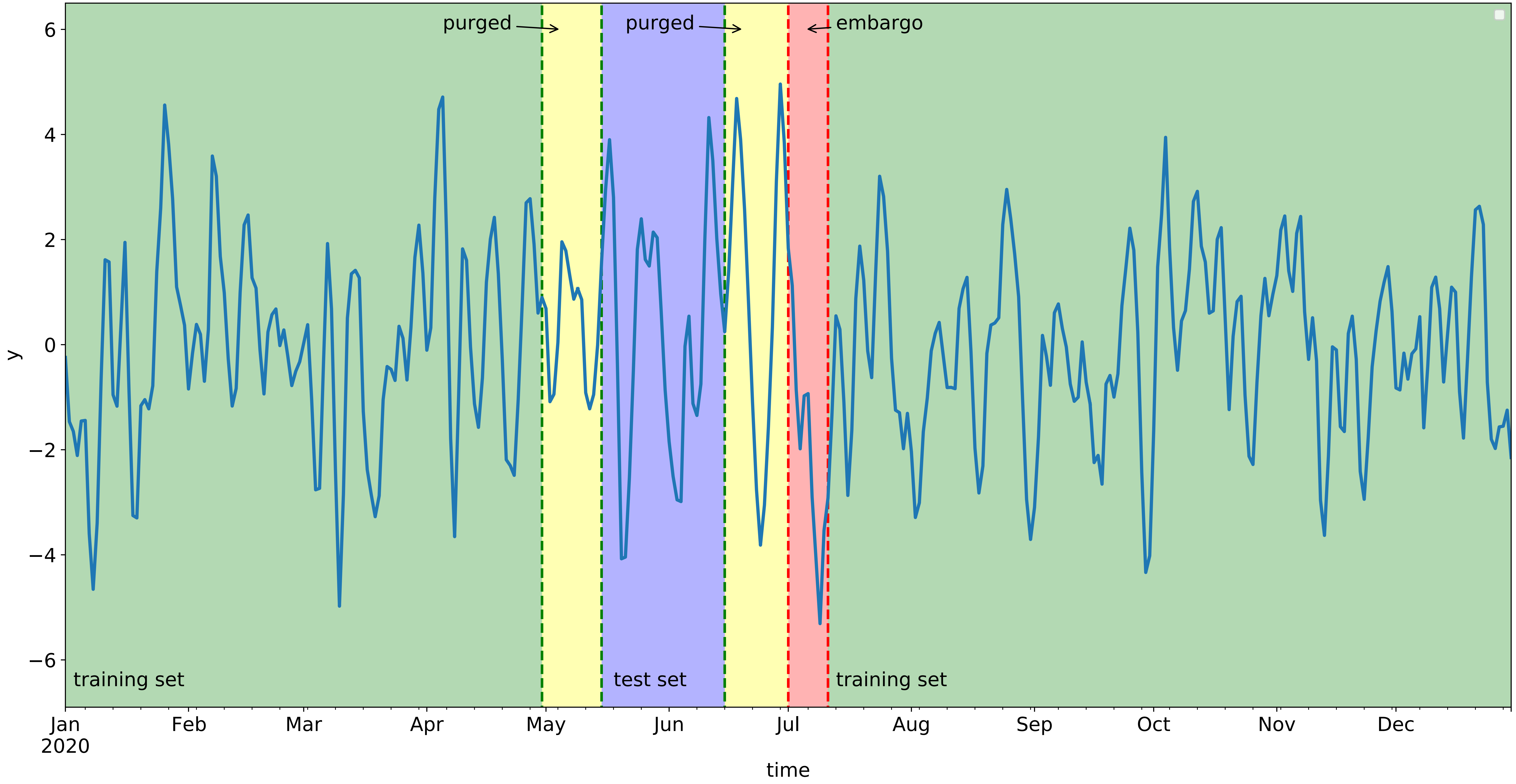


Illustration of combining purged CV and embargo CV.

A Short Summary

- CV strategies so far solve the two main problems:
  - the test set may precede the training set, and
  - the test set may have high serial correlation with the training set.
- Problem 3 has not been addressed: we are effectively looking at *one test path* and there is one and only one forecast generated for each observation.
  - This can be solved by CPCV (see the next slide).

Combinatorial Purged CV (CPCV)

Introduction

- Keep the chronological order of T observations, and partition them into N non-overlapping groups of roughly equal size.

- From N groups, choose k of them to be in the test set.
    - The number of possible training and test splits is $\binom{N}{k}$.
    - Since each combination has k test groups, the total number of groups in the test set is $k\binom{N}{k}$.

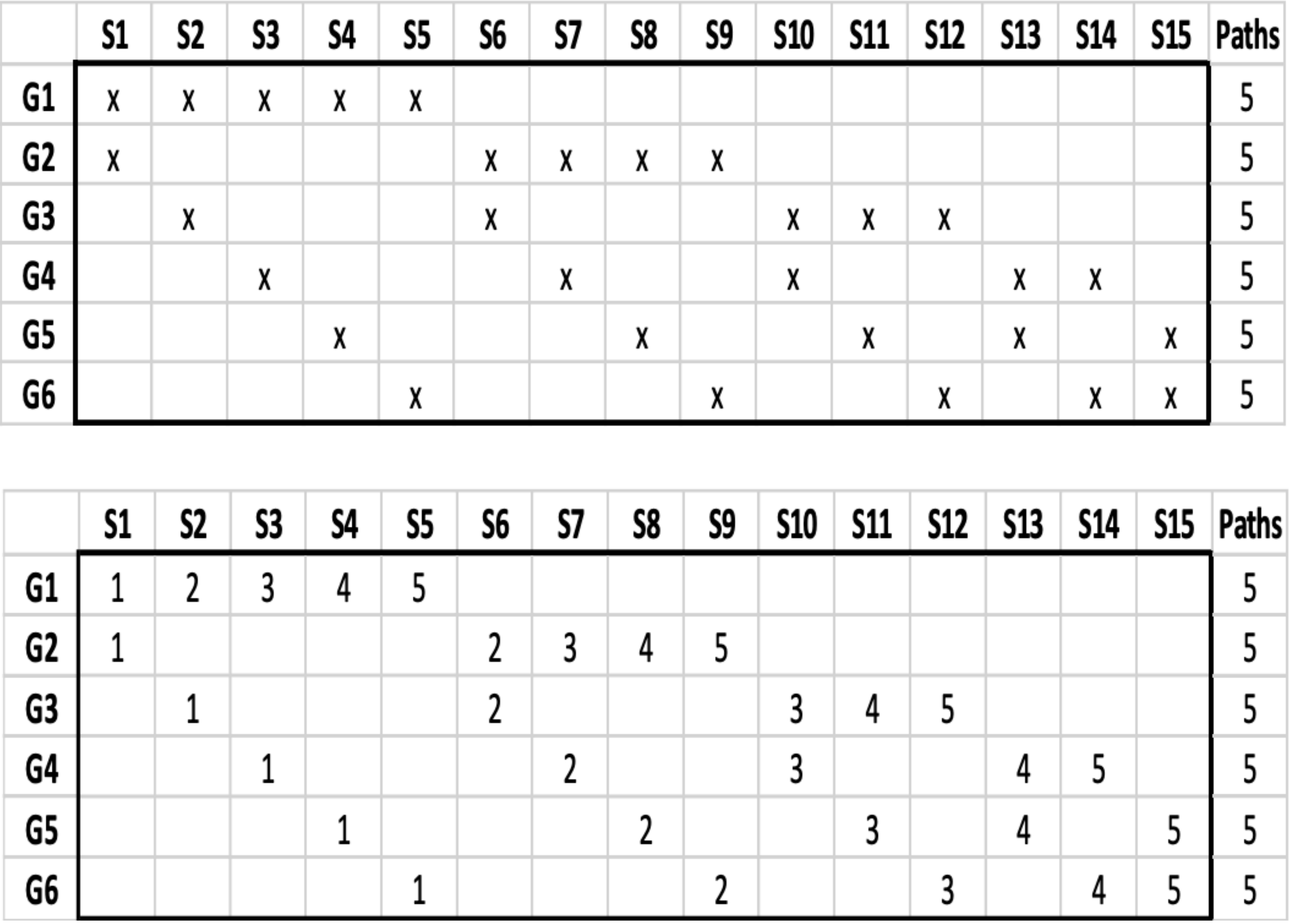
- We can test a total number of $\varphi(N, k) := \frac{k}{N}\binom{N}{k}$ paths.

Combinatorial Purged CV (CPCV)

Example

- Divide all observations into $N = 6$ groups and use $k = 2$ groups as the test set. There are $\binom{6}{2} = 15$ splits, indexed by $S1, \cdots, S15$. In the upper panel, for each split, groups marked by x belong to the test set, and unmarked groups belong to training set.

- We have $\varphi(6{,}2) = 5$ test paths. For example, Path 1 consists of forecasts from (G1, S1), (G2, S1), (G3, S2), (G4, S3), (G5, S4) and (G6, S5).



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	Paths
G1	x	x	x	x	x											5
G2	x					x	x	x	x							5
G3		x				x				x	x	x				5
G4			x				x			x			x	x		5
G5				x				x			x		x		x	5
G6					x				x			x		x	x	5

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	Paths
G1	1	2	3	4	5											5
G2	1					2	3	4	5							5
G3		1				2				3	4	5				5
G4			1				2			3			4	5		5
G5				1				2			3		4		5	5
G6					1				2			3		4	5	5

Example of training and test splitting in the CPCV.

Combinatorial Purged CV (CPCV)

Procedures

- Keep the chronological order of T observations, and partition them into N groups;
- Compute all possible training and test splits with $N - k$ groups in training set and k groups in the test set;
- Use the purged CV and the embargo CV (when necessary) to fit models on the training sets, and produce forecasts on the test sets;
- Compute the mean generalization error for each of the $\varphi(N, k)$ paths.

Combinatorial Purged CV (CPCV)

Comments and Remarks

- We can now examine multiple test paths.

- We have $\varphi(N, k)$ mean generalization errors and can examine the *distribution* of generalization errors.

- When $k = 1$, we have $\varphi(N, 1) = 1$ path, and CPCV reduces to CV.

- A typical choice of k is 2. As a consequence, we obtain $\varphi(N, 2) = N - 1$ paths.

Concluding Remarks

- CV in time series analysis (or more broadly, CV in analysis of dependent/correlated data) is still an active research area.

- CV techniques introduced here can be used as building blocks of fancier variants, for example,

  - nested *K*-fold purged CV / CPCV,

  - stratified *K*-fold purged CV / CPCV,

  - ...

- In practice, one can adopt different CV techniques and compare their empirical performances to avoid overfitting to the training data.

- Prior to any modeling, please
  - ensure the time series is stationary; otherwise, do appropriate transformation or smoothing or estimation.
  - understand the auto- and cross-correlations among all variables.
- Besides using CV to avoid overfitting, try strategies in model fitting as well, including
  - early stopping in the base estimators, and
  - using ensemble methods such as bagging and stacking.

Thank You
Asante
Merci
Welalin
Kiitos
Vinaka
Kia Ora
Spasibo
Salamat
Multumesc
Mochchakkeram
Ua Tsaug Rau Koj
Dank Je
Terma Kasih
Cam on ban
Mamana
Obrigado
Matondo
Raibh Maith Agat
Grazie
Arigato
Juspaxar
Maake
Dankon
Chokrane
Matur Nuwun
Nirringrazzjak

Questions?

References

- Prado, Marcos Lopez de. 2018. Advances in Financial Machine Learning. John Wiley & Sons.
- Burman, Prabir, Edmond Chow, and Deborah Nolan. 1994. “A Cross-Validatory Method for Dependent Data.” Biometrika 81 (2): 351–58.
- Racine, Jeff. 2000. “Consistent Cross-Validatory Model-Selection for Dependent Data: hv-Block Cross-Validation.” Journal of Econometrics 99 (1): 39–61.