Cross-Validation for Correlated Data

Assaf Rabinowicz [*]
Department of Statistics, Tel-Aviv University, Tel-Aviv, Israel, 69978

Saharon Rosset
Department of Statistics, Tel-Aviv University, Tel-Aviv, Israel, 69978

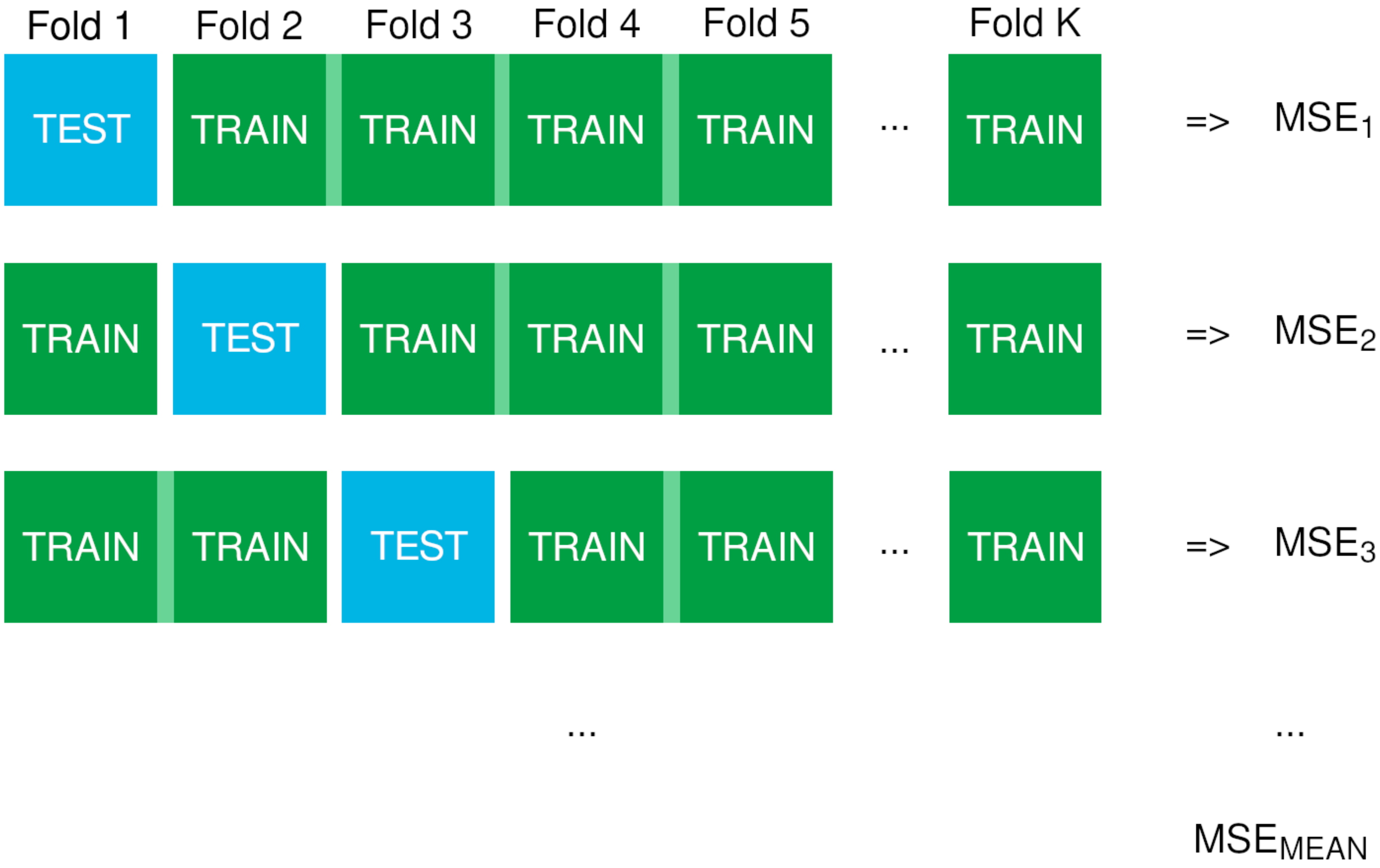
April 05, 2020

Abstract

K-fold cross-validation (CV) with squared error loss is widely used for evaluating predictive models, especially when strong distributional assumptions cannot be taken. However, CV with squared error loss is not free from distributional assumptions, in particular in cases involving non-i.i.d. data. This paper analyzes CV for correlated data. We present a criterion for suitability of standard CV in presence of correlations. When this criterion does not hold, we introduce a bias corrected cross-validation estimator which we term CV_c, that yields an unbiased estimate of prediction error in many settings where standard CV is invalid. We also demonstrate our results numerically, and find that introducing our correction substantially improves both, model evaluation and model selection in simulations and real data studies.

Keywords: *Prediction Error Estimation; Model Selection; Dependent Data; Linear Mixed Model; Gaussian Process Regression*

K-fold Cross Validation



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5		Fold K		
TEST	TRAIN	TRAIN	TRAIN	TRAIN	…	TRAIN	=>	MSE$_1$
TRAIN	TEST	TRAIN	TRAIN	TRAIN	…	TRAIN	=>	MSE$_2$
TRAIN	TRAIN	TEST	TRAIN	TRAIN	…	TRAIN	=>	MSE$_3$
			…					…
								MSE$_{MEAN}$

K-fold Cross Validation

$X = \{\boldsymbol{x}_i\}_{i=1}^{n}$ − data matrix, where $\boldsymbol{x}_i$ comes from distrubution P_X

$\boldsymbol{y} = \{y_i\}_{i=1}^{n}$ − response vector, in some relation with X

$T = \{\boldsymbol{y}, X\}$ − dataset

T_k − $k-$th fold, for $k = 1, 2, \dots K$

$T_{-k} = T \backslash T_k$

$\hat{y}(\boldsymbol{x}_i, T_{-k})$ − prediction on $\boldsymbol{x}_i$ based on model build on T_{-k}

$$CV = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in k-\text{th f.}} MSE(y_i, \hat{y}(\boldsymbol{x}_i, T_{-k}))$$

K-fold Cross Validation

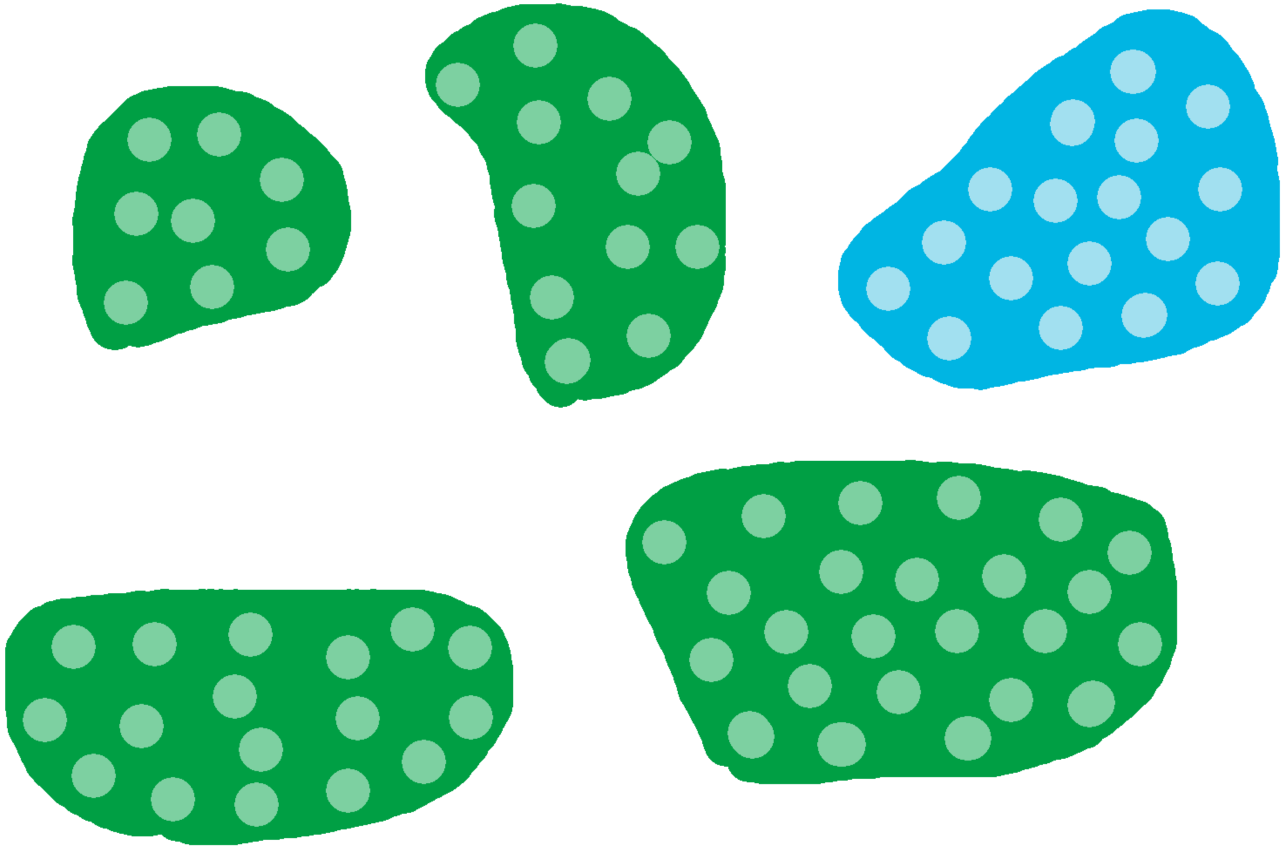
CV is an estimator of **generalization error**

$$\boldsymbol{E}_{T_{tr},T_{te}} \frac{1}{n_{(te)}} MSE\left(y_{(te)i}, \hat{y}\left(\boldsymbol{x}_{(te)i}, T_{tr}\right)\right)$$

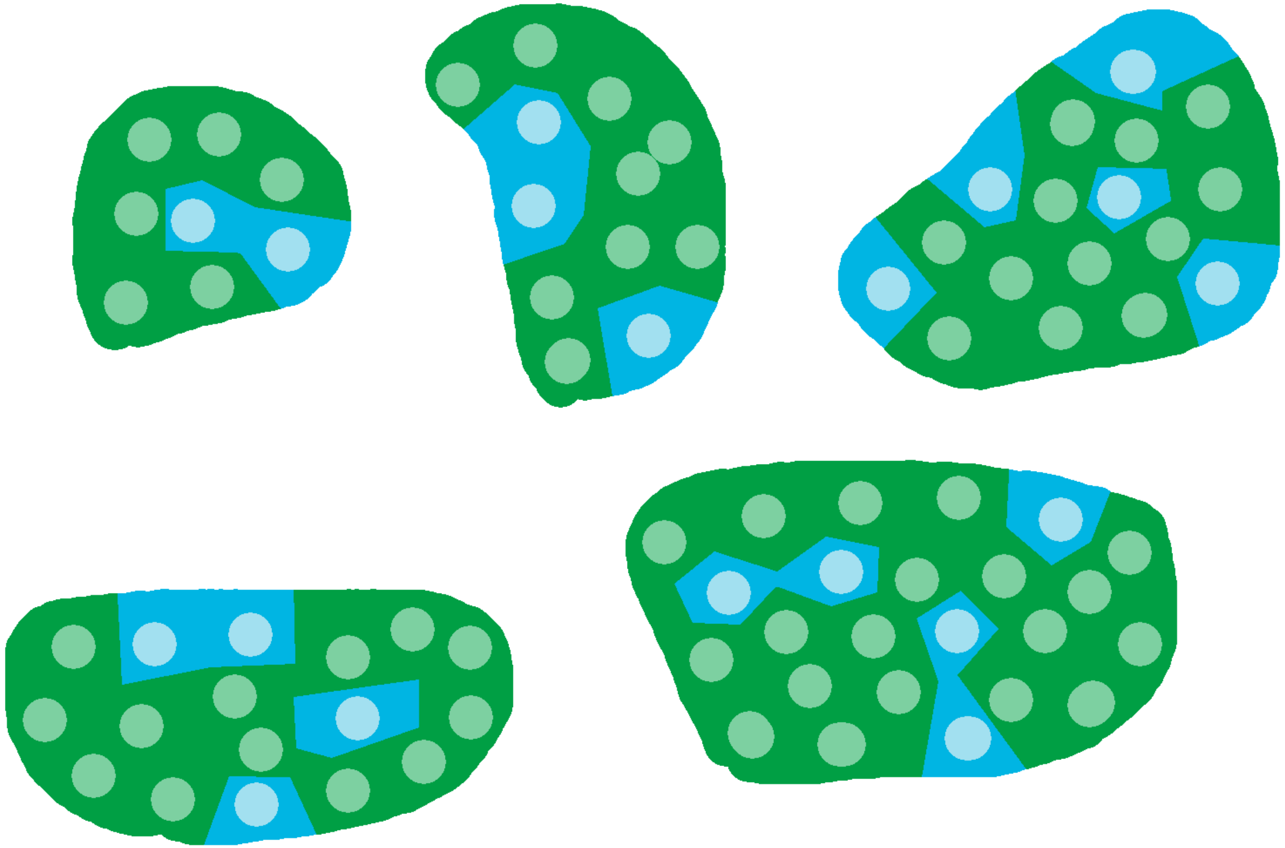
where

T_{tr}, T_{te} $-$ training and test set (random variables in this context)

Problem – when CV is unbiased?



Problem – when CV is unbiased?



The Unbiasedness Criterion

If $P_{T_{tr},T_{te}} = P_{T_k,T_{-k}}$ for each k in $\{1, 2, \ldots, K\}$, then CV is an unbiased estimator of the generalization error.

In other words: when the CV partitioning preserves the distributional relation between the prediction set to the training set, then CV is unbiased.

It is highly dependent on correlation of data, however, correlated data does not imply biasedness automatically.

Bias correction

Under some conditions, we are able to calculate corrected CV (CV_c). This method may be more profitable than expected optimism or other bias-correction methods.

Results

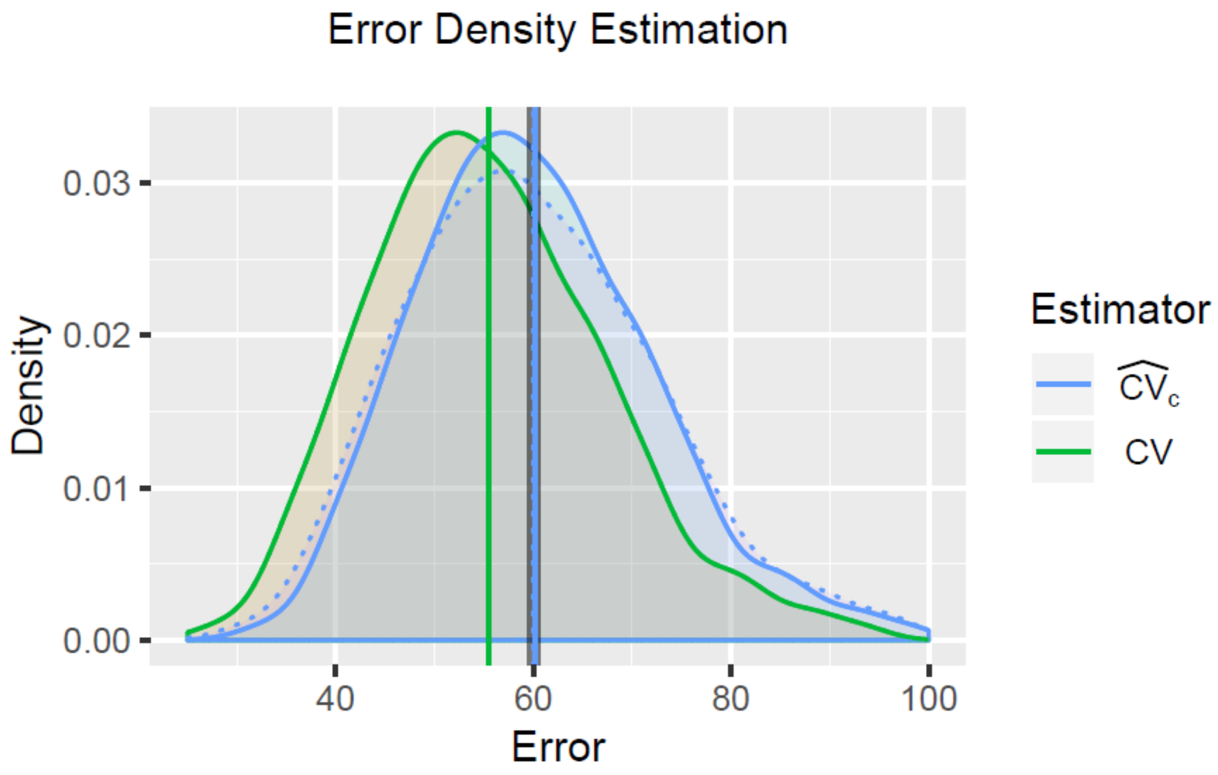


Figure 1: Densities of CV and $\widehat{CV}_c$ and their means compared to the generalization error (in black). Two scenarios are presented: when the variance parameters are known (solid line) and when they are estimated (dashed line).

Results

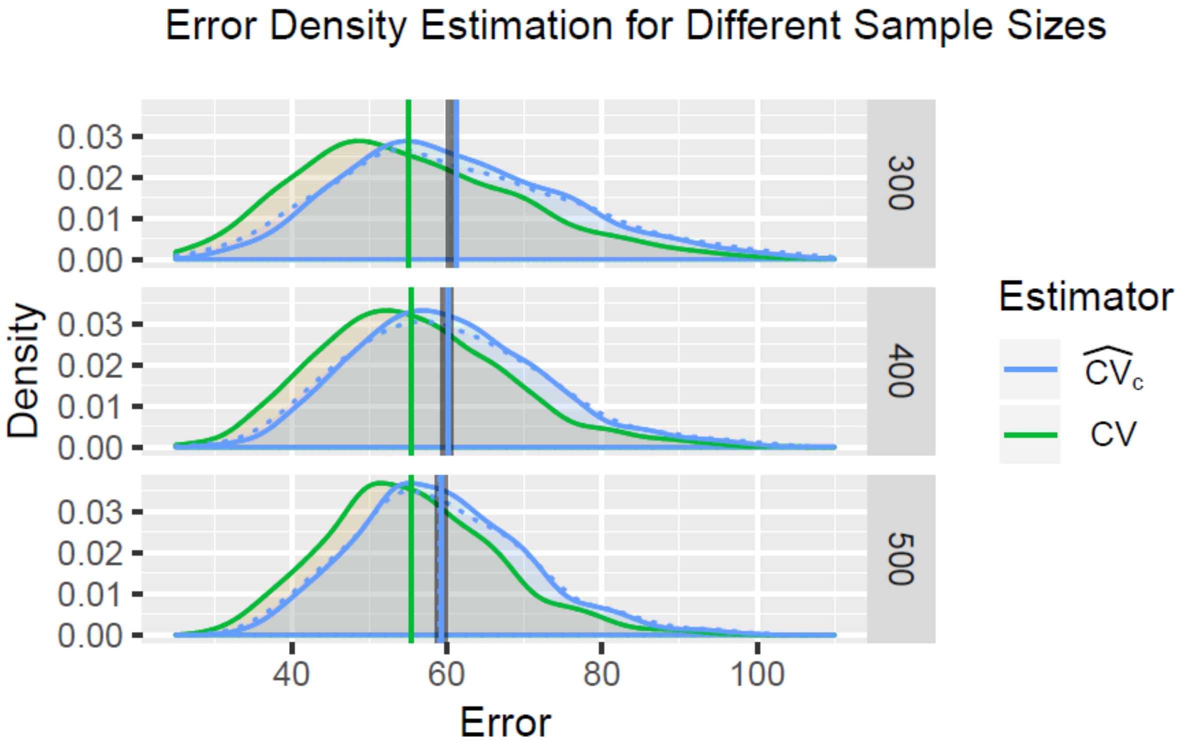
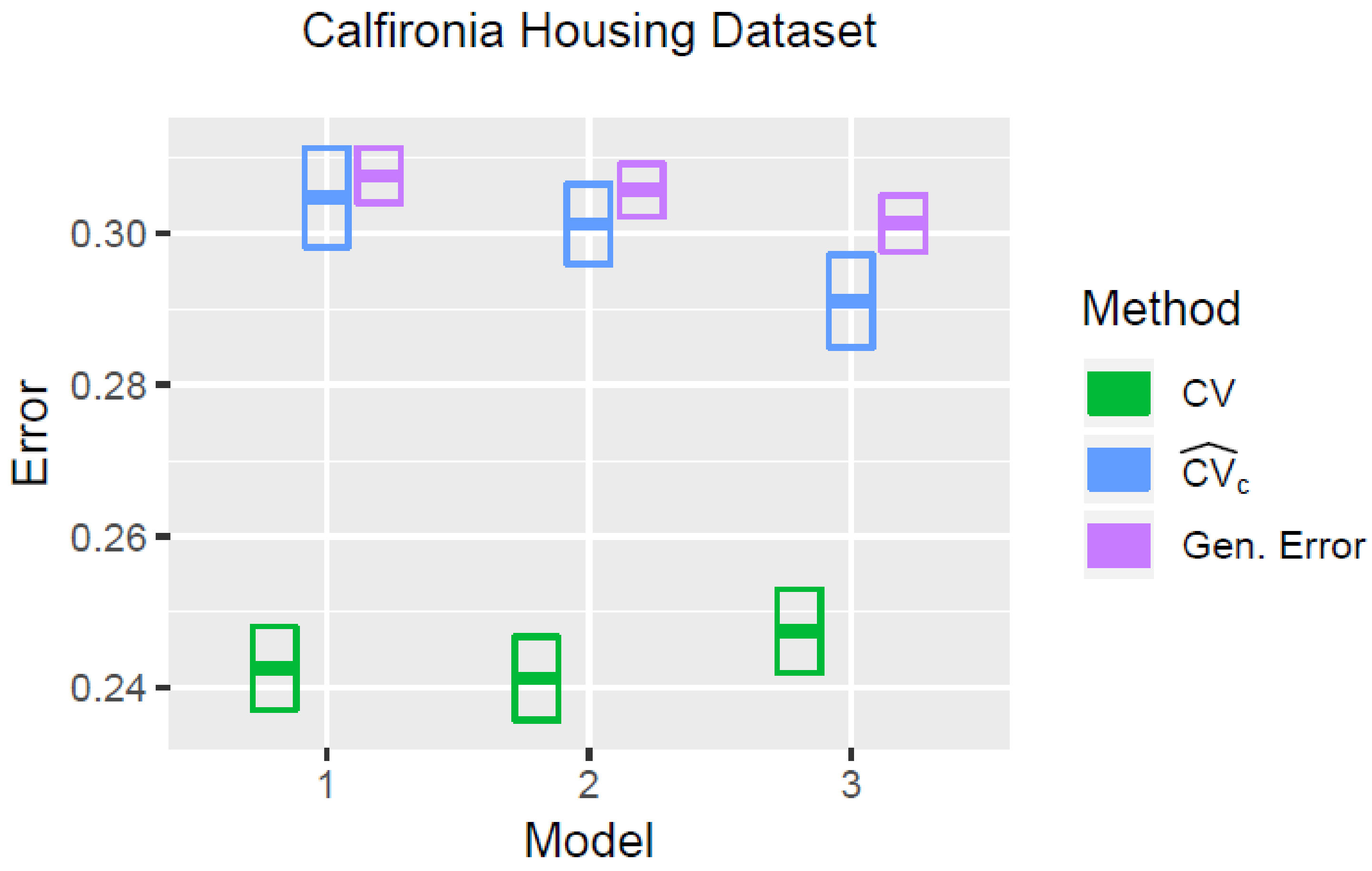
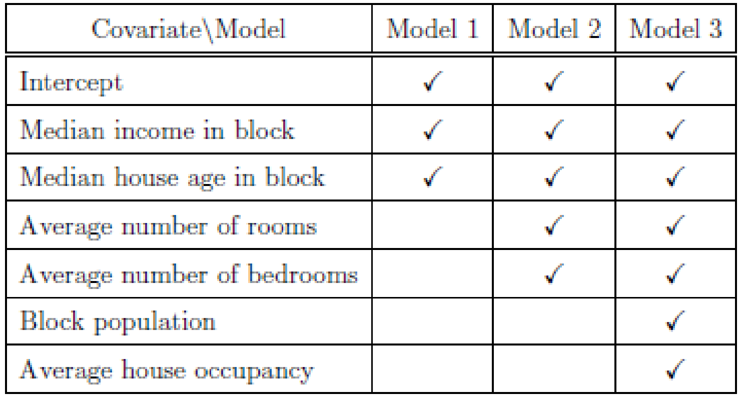


Figure 2: Densities of CV and $\widehat{CV}_c$ and their means compared to the generalization error (in black) for different sample sizes. Two scenarios are presented: when the variance parameters are known (solid line) and when they are estimated (dashed line).

Results





Covariate\Model	Model 1	Model 2	Model 3
Intercept	✓	✓	✓
Median income in block	✓	✓	✓
Median house age in block	✓	✓	✓
Average number of rooms		✓	✓
Average number of bedrooms		✓	✓
Block population			✓
Average house occupancy			✓

Important conclusions

- We should be more aware of correlations in our data
- We should be more aware of how we are using CV
- We should be more aware of which modeling approach to use