Simulations of fold number selection in cross-validation of linear and LASSO regression

Angelos Vasilopoulos   Gregory J. Matthews

Abstract

k-fold cross-validation is a popular method of error estimation for model selection in computational research. However, there is limited focus in the literature on the question of what fold number k is appropriate for cross-validation with a dataset of a given size. Here we review the literature on this topic and present a simulation of linear and least absolute shrinkage and selection operator (LASSO) regression prediction error estimation at various values of k and sample size n. In agreement with a growing body of literature, we find that contrary to a persisting understanding, there is no bias-variance trade-off in selection of k. | *Keywords*: cross-validation, optimization, fold number

1   Introduction

k-fold cross-validation is a popular method of error estimation for model selection in computational research. n observations are divided into k groups. In a first iteration, $i = k - 1$ groups are used as a training set. The remaining group is used as a test set to calculate estimated prediction error ($\widehat{\text{PE}}$). This process is repeated i times, with a different test set in each iteration. The average of k $\widehat{\text{PE}}$

$$\hat{\theta} = CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \widehat{\text{PE}}_i$$

is meant to estimate the true model error θ, i.e., the error of the model tested on the population.

Despite the popularity of cross-validation, there is limited focus in the literature on the question of what fold number k is appropriate for cross-validation with a dataset of a given size. One popular idea is that selection of k comes with a bias-variance trade-off—specifically, that as k increases, the bias $Bias(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta$ and variance $Var(\hat{\theta}) = \text{E}(\hat{\theta}^2) - \text{E}(\hat{\theta})^2$ of k-fold error estimation, as in $Err(\hat{\theta}) = Bias^2(\hat{\theta}) + Var(\hat{\theta})$, decrease and increase, respectively. This idea appears in early literature (Efron (1983); Kohavi (2001)).

Subsequent, albeit limited, literature argues differently. In the case of leave-one-out cross-validation (LOOCV), i.e., with $k = n$, some authors suggest that asymptotically both bias and variance of error estimation decrease as k increases (Burman (1989)) and that bias and variance of error estimation are uniformly low (Breiman and Spector (1992)).

More recently have been proposed ways to quantify the variance reduction achieved by cross-validation when the true prediction error (PE) is not known, e.g., as mean-square stability (Kale, Kumar, and Vassilvitskii (2011)), where

$$\underset{S,z,z'}{\mathrm{E}}[(\ell_{\mathcal{A}(S)}(z) - \ell_{\mathcal{A}(S^{i,z'})}(z))^2] \leq \beta,$$

i.e., the expectation of the squared difference in loss ℓ of an algorithm $\mathcal{A}$ over a dataset S and over a dataset S' such that one element $z' \sim Z$ has been added to S is less than β, which depends on the algorithm; or as loss stability (Kumar et al. (2013)), where

$$\underset{S,z,z'}{\mathrm{E}}[(\ell'_{\mathcal{A}(S)}(z) - \ell'_{\mathcal{A}(S^{i,z'})}(z))^2] \leq \gamma,$$

$\ell'_{\mathcal{A}(S)}(z) = \ell_{\mathcal{A}(S)}(z) - \bar{\ell}_{\mathcal{A}(S)}(z)$, and γ depends on the algorithm. However, it has also been demonstrated that, due to overlap between training and test sets in cross-validation, there is no universal (i.e., valid under all distributions) unbiased estimator of the variance of k-fold cross-validation (Bengio and Grandvalet (2004)).

In addition to theoretical analyses of variance reduction by cross-validation, there are some simulation results showing this phenomenon. However, simulations currently in the literature provide limited insight into the dependence of optimal fold number k_{optimal} on sample size n (Zhang and Yang (2015)) or involve biased variance calculations (Marcot and Hanea (2021)). Here we present a simulation of linear regression and least absolute shrinkage and selection operator (LASSO) regression to observe the relationship of cross-validation fold number k to model selection accuracy for various samples of size n of a known population.

2 Simulation

Consider a population of size N = 500,000 with five of $P = 100$ features $X_1...X_{100} \sim N(0,1)$ a linear combination of feature $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5 + \epsilon$ where $\beta_1 = \cdots = \beta_5 = 1$ and $\epsilon \sim N(0,10)$. The true model of Y is $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5$ and the mean squared error (MSE) of the true model is

$$\theta = \mathrm{MSE}[\mathrm{E}(Y)] = \frac{\sum_{i=1}^N [Y_i - \mathrm{E}(Y_i)]^2}{N}.$$

From the population we take a sample of size n. We estimate Y as $\hat{Y}$ by regression on a subset of $X_1...X_{100}$, regression coefficients estimated by the least squares method, and

compute $\mathrm{MSE}(\hat{Y})$ by k-fold cross-validation as

$$\hat{\theta} = \mathrm{MSE}(\hat{Y}) = \frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{i=1}^{n/k} (Y_{j_i} - \hat{Y}_{j_i})^2}{n/k}$$

where j_i is the index of the ith element of the jth fold.

In addition to regression with $X_1...X_5$, the true model, we under-fit with $X_1...X_3$, over-fit with $X_1...X_{20}$ and $X_1...X_{100}$, and do regression with noise $X_6...X_{100}$ only for each $k \in \{2, 10, 20, 30, \ldots, n \in \{100, 200, 300, \ldots, 1000\}\}$. We perform this simulation 1000 times and then calculate simulation-wise $\mathrm{MSE}(\hat{\theta})$ for each k, for each n.

We perform an additional 1000 simulations predicting Y as $\hat{Y}$ by LASSO regression instead of linear regression for each $k \in \{2, 10, 20, 30, \ldots, n \in \{100, 250, 500, 750, 1000\}\}$, due to computational limitations. To avoid data leakage, we introduce an inner 5-fold cross-validation loop for the selection of parameter λ as in the minimization of

$$\sum_{i=j}^{p} (Y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

For each k, for each n, we count the number of times C that the true model is selected and consider as optimal fold number for each n the value k_{optimal}, at which C is the highest. This is because in practice the model with the lowest $\widehat{\mathrm{PE}}$ is more likely to be selected. We refer to this as lowest-error model selection. As our simulation results show, however, it is possible for k-fold cross-validation to result in calculations of $\mathrm{MSE}(\hat{\theta})$ such that a competing model has a lower $\widehat{\mathrm{PE}}$ than the true model. This false model may have the lowest $\widehat{\mathrm{PE}}$ but its generalization error will be higher in the long-run (i.e., when tested on a large part of the population) than the generalization error of the true model. Thus, it is preferable for the value of k selected to result in the true model having the lowest $\widehat{\mathrm{PE}}$.

Having considered the effect of k on model selection accuracy for a given n, we may also consider the relationship of k_{optimal} and n. At this point it is important to take into account the possibility of diminishing returns after a certain value of k: that C appears to increase until $k = n$ for certain n does not mean that for those n LOOCV results in significantly higher C. In other words, after a certain point, improvements in C become negligible and resource-expensive increases of k become unnecessary. We refer to this "point of diminishing returns" as $k_{optimal}^{*}$.

To estimate $k_{optimal}^{*}$ for each n, we first fit to the dataset $D(k, C)$ for each sample slize n a non-linear least squares model of the form

$$\hat{C}(k) = \frac{A}{k} + B$$

with constants A and B specific to each n. We then perform $n/10 - 1$ Mann-Kendall tests of

trend significance on the values fitted at $k_i...n \in k$, $i = [1..(n/10 - 1)]$, as the Mann-Kendall test requires at least three data points. We apply a Bonferroni correction to account for the problem of multiple comparisons, with $\alpha = 0.05$ and a corrected cut-off of $p* = \alpha/(n/10-1)$. Then

$$k^*_{optimal} = \max_{p<p^*} k$$

since $\hat{C}(k)$ increases monotonically.

3   Results and discussion

Interpretations of early literature have resulted in lasting misconceptions about the use of cross-validation. Such misconceptions include the idea that there is a bias-variance trade-off

$$Bias^2(\hat{\theta}) \propto \frac{1}{Var(\hat{\theta})}$$

associated with selection of k and that $k = 10$ is the best value to use in k-fold cross-validation (Kohavi (2001)).

In agreement with a small but growing body of literature, our simulation results suggest that neither of these ideas are necessarily correct. Instead, in the context of linear and LASSO regression with normally distributed data, we find that for various n both bias and variance decrease as k increases (Figures 1 - 3), i.e., $Bias^2(\hat{\theta}) \propto Var(\hat{\theta})$, and although 10-fold cross-validation seems to be superior for large n, for smaller samples other values of k are better for model selection (Figure 4).

As previous authors have noted, it is important to distinguish the possible goals of cross-validation: estimation of PE and model selection (Zhang and Yang (2015)), which previously have been conflated (Kohavi (2001)). In this study, as described in the previous section, we define k_{optimal} in terms of its utility in the latter. For LASSO regression, we find that for larger n, e.g., $n = 1000$, as k_{optimal} increases past a certain value lowest-error model selection becomes less reliable, i.e., C decreases, i.e., the true model has the lowest MSE less frequently (Figure 4).

The reason for this is related to what is known as the cross-validation paradox (Yang (2006)): greater quantity of data results in more accurate estimation of PE; occasionally, this makes model-wise differences in $\widehat{PE} = CV_{(k)}$ less exaggerated, making it more difficult to distinguish between models and resulting in less accurate model selection.

However, we do not observe the same for linear regression in our simulations. Although cross-validation results in similar bias-variance reduction in both cases, in the case of linear regression C increases with increasing k, so that LOOCV appears to result in the most accurate model selection. However, this discrepancy is not surprising, as our linear regression simulation does not involve variable selection, so that the differences in the models are

more pronounced. To observe the cross-validation paradox in linear regression would require comparison of more similar linear models predisposed to similarity in $\widehat{\text{PE}}$, as in the simulation of LASSO regression.

The results dispalyed in Figure 5 suggest an increase in k_{optimal}/n with increasing n (Figure 5). In other words, as n increases, the training set size required for optimal model selection accuracy increases.

4   Conclusion

Early literature suggests that increasing cross-validation fold number is related to decreasing bias and increasing variance of error estimation. However, more recent work suggests that this is not the case. Instead, increasing k results in bias and variance reduction. This phenomenon is observable in our simulation results, although it is also clear from the results that the reductions are not limitless, meaning that there is a point of diminishing returns in increasing k and computational power allocation, as error estimation does not improve, at least not substantially, after a certain value of k.

Our results also indicate a relationship between k_{optimal} and sample size n, although more data is necessary to elucidate a trend. This is interesting as a future direction, as modeling the relationship between n and k_{optimal} would have practical utility, potentially improving the selection of k from a largely arbitrary decision between 5 and 10.

Future research may also focus on error estimation of other models, including models capturing non-linear relationships or involving tuning of multiple hyper-parameters, e.g., random forest or gradient boosting. It may also be interesting to study the value of k_{optimal} in repeated cross-validation or nested cross-validation, with the value of k variable in the inner loop, outer loop, or both.
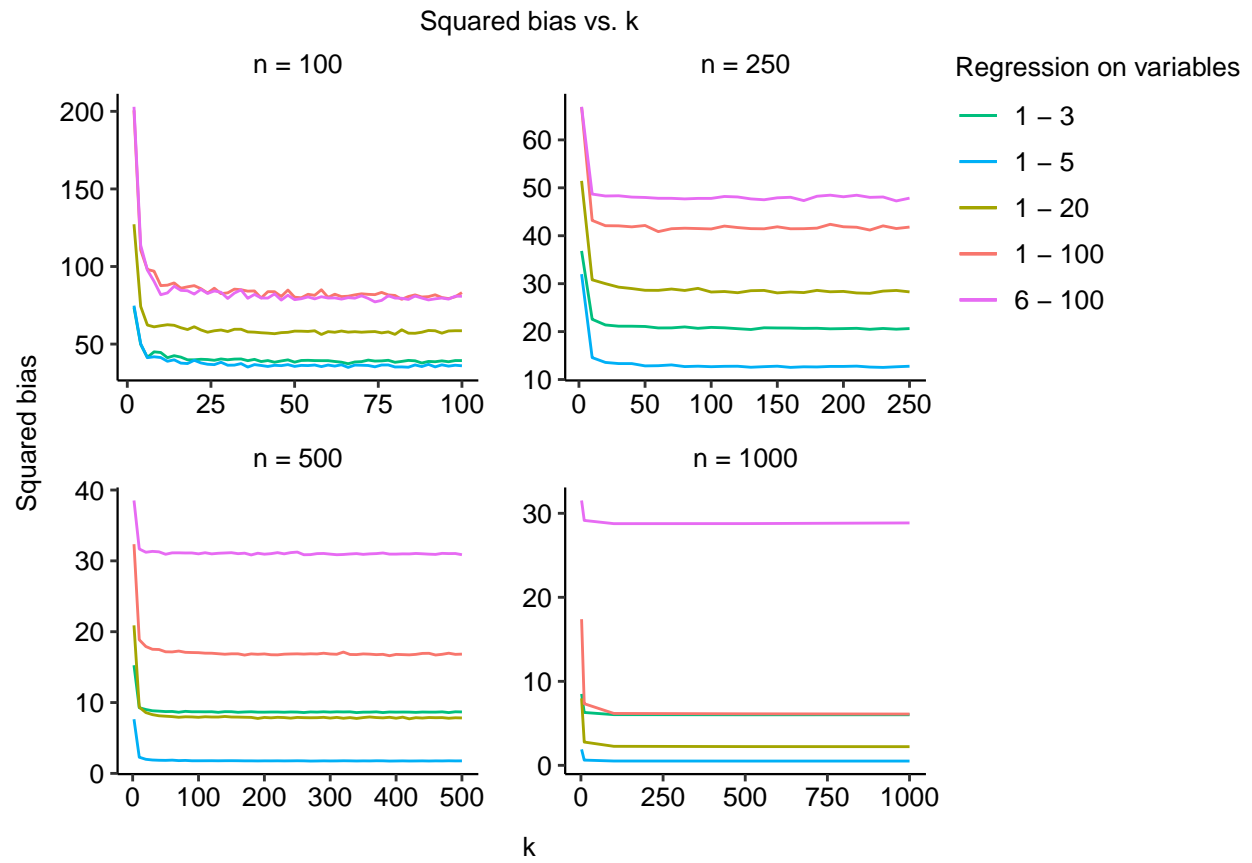
5 Appendix

Figure 1: Squared bias vs. fold size k for sample sizes n. As k increases, bias decreases initially before leveling off.
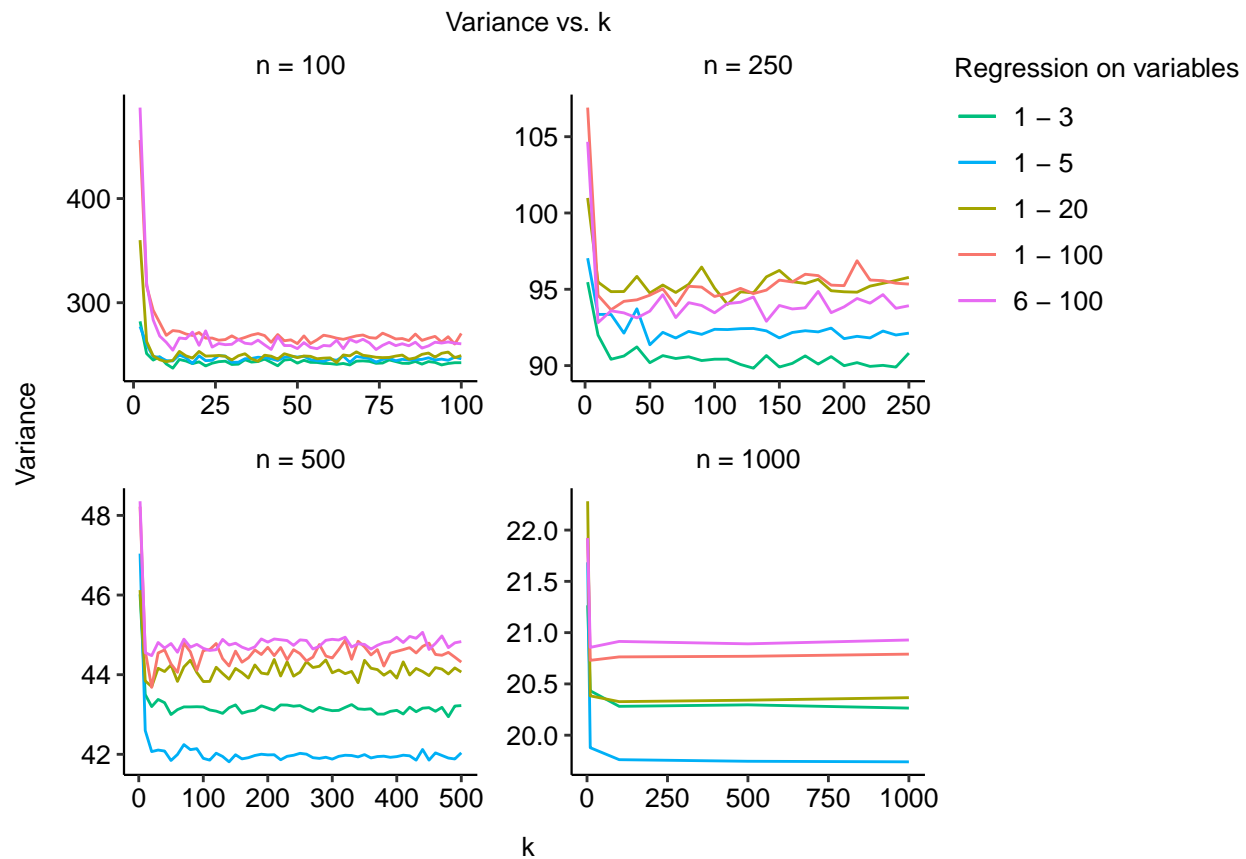
Figure 2: Variance vs. fold size k for sample sizes n. As k increases, variance decreases initially before leveling off.
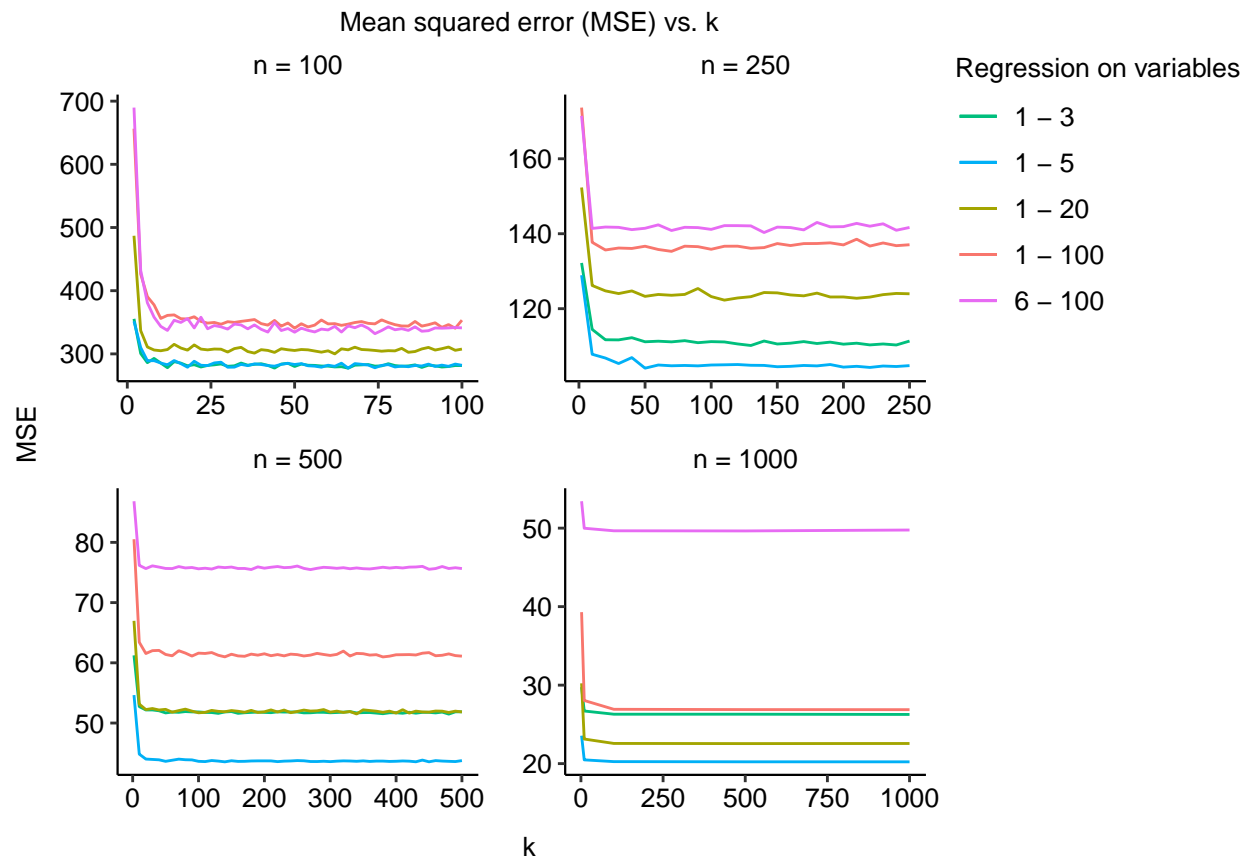
Figure 3: Mean squared error (MSE) vs. fold size k for sample sizes n. As k increases, MSE decreases initially before leveling off.
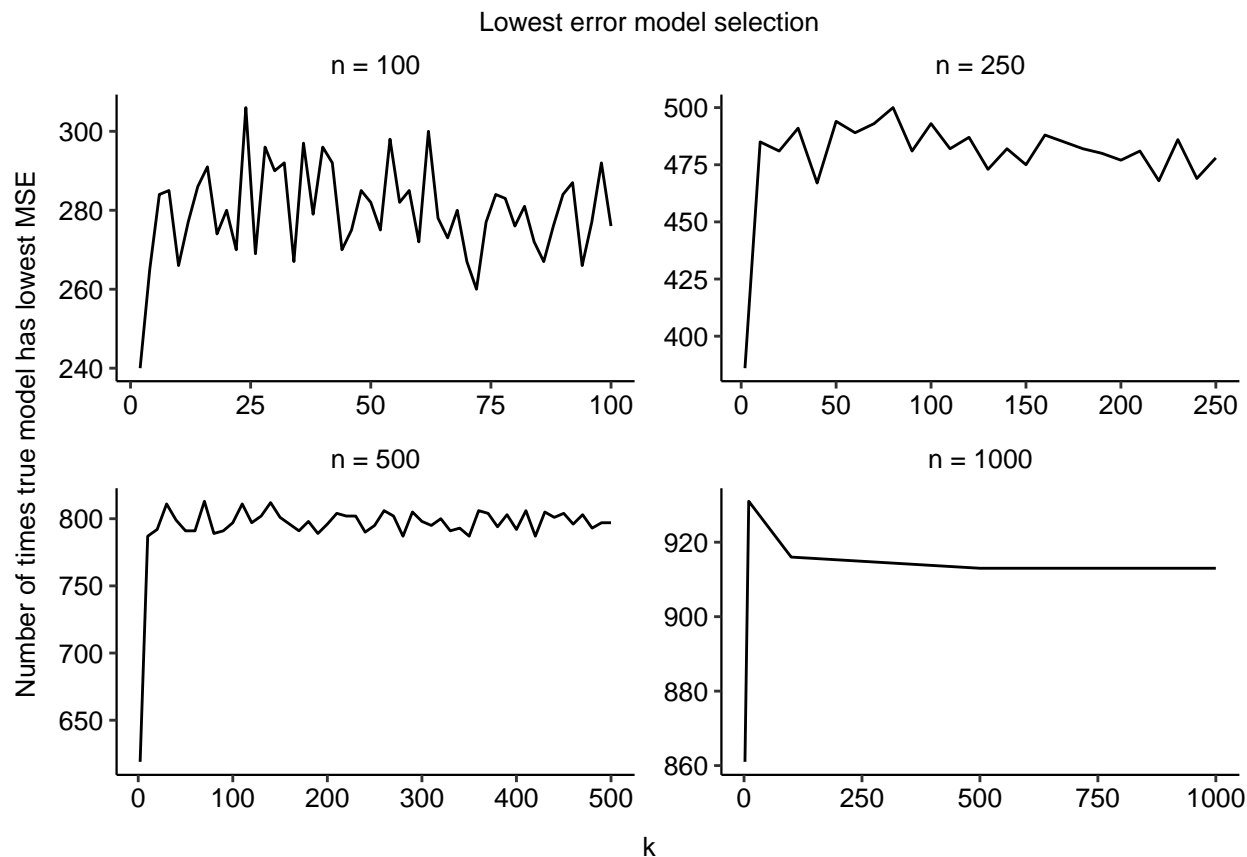
Figure 4: Count of true model selection by the lowest-error criterion vs. fold size k for sample sizes n.

Acknowledgements

Supplementary Material

All code for reproducing the analyses in this paper is publicly available at https://github.com/gjm112/optimal_k.

References

Bengio, Yoshua, and Yves Grandvalet. 2004. "No Unbiased Estimator of the Variance of k-Fold Cross-Validation," June.

Breiman, Leo, and Philip Spector. 1992. "Submodel Selection and Evaluation in Regression. The x-Random Case." *International Statistical Review / Revue Internationale de*

Statistique 60 (3): 291. https://doi.org/10.2307/1403680.

Burman, Prabir. 1989. "A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods." *Biometrika* 76 (September): 503–14. https://doi.org/10.1093/biomet/76.3.503.

Efron, Bradley. 1983. "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation." *Journal of the American Statistical Association* 78 (382): 316–31. https://doi.org/10.1080/01621459.1983.10477973.

Kale, Satyen, Ravi Kumar, and Sergei Vassilvitskii. 2011. "Cross-Validation and Mean-Square Stability." In, 487–95.

Kohavi, Ron. 2001. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection" 14 (March).

Kumar, Ravi, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. 2013. "Near-Optimal Bounds for Cross-Validation via Loss Stability." In *Proceedings of the 30th International Conference on Machine Learning*, edited by Sanjoy Dasgupta and David McAllester, 28:27–35. Proceedings of Machine Learning Research 1. Atlanta, Georgia, USA: PMLR. https://proceedings.mlr.press/v28/kumar13a.html.

Marcot, Bruce, and Anca Hanea. 2021. "What Is an Optimal Value of k in k-Fold Cross-Validation in Discrete Bayesian Network Analysis?" *Computational Statistics* 36 (September). https://doi.org/10.1007/s00180-020-00999-9.

Yang, Yuhong. 2006. "Comparing Learning Methods for Classification." *Statistica Sinica* 16 (April).

Zhang, Yongli, and Yuhong Yang. 2015. "Cross-Validation for Selecting a Model Selection Procedure." *Journal of Econometrics* 187 (1): 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006.