Stratified Over Representative k-folds Cross-Validation

William Franz Lamberti
George Mason University
Computational and Data Sciences

1. Abstract

A method for handling small sample cases using a variation of stratified k-folds cross-validation is presented. The key difference between traditional stratified k-folds cross-validation and the sampling approach presented here is over representing the smaller strata. Further, the specific cases to utilize the new approach is when stratified k-folds cross-validation cannot be used. An intuitive explanation is provided alongside simulations using synthetic data and the famous Fisher or Anderson iris dataset.

2. Introduction

Cross-validation (CV) has become an integral part of many fields such as statistics, machine learning, and artificial intelligence. CV's roots can be traced back to Mosteller and Wallce [14]. However, the first true definition of cross-validation seems to have been provided by Mosteller and Tukey [13]. Stone provided an overall summary of cross-validation at the time[16]. Breiman and Spector and Kohavi showcased the power and efficiency of k-folds cross validation during their simulations to be used in more modern applications[2, 11]. Kohavi continues on to mention that stratified k-folds cross validation may achieve better results in regards to the bias and variance of the estimators [11].

2.1 Recent Work

Others are working on variations of CV that are tailored to specific types of modeling problems. For instance, work has already been done on a CV algorithm that is tailored for network analysis [3]. Others have developed a modeling averaging approach utilizing k-folds CV [10]. Some have investigated the relationship between CV and tuning parameters [5]. Investigations have also been conducted on studying heterogeneity and the ensembling of data sets using Bayesian methods [18]. STORKC is very general and can be applied in a variety of settings. STORKC can be thought of as a practical substitution for stratified CV when it is not feasible to utilize.

2.2 Outlined Approach

There are many variations of CV [8]. However, a critical assumption for stratified CV is that the data is sufficiently large to support such a schema that requires stratification. It is desired to develop an approach that can handle smaller samples of specific cases. Murphy has already shown on data that some of these small sample scenarios can render CV unhelpful[15]. An approach that can handle some of these scenarios will be referenced as stratified over-representative k-folds cross-validation, also known as STORKC (which is pronounced like the bird, 'Stork'). Further, the simulations will show that traditional non-stratified k-folds CV will fail to provide prediction error values at a consistent level regardless of the amount of noise in the data. This points to the need for a more robust sampling method that is more resistant the composition of the data.

The outline for the paper is as follows: the definition of terms, the intuition and simple examples, the formal setup, simulations on synthetic data and the runs on the Fisher or Anderson iris data set from base R using simple linear regression, discussion and conclusion.

3. Background

3.1 Defining Terms

Before the main idea is discussed in detail, it is desired to define some basic terminology. Different communities define the same words to mean different things, or merely swap around definitions and words. Thus, for the sake of absolute clarity, three terms will be defined and used in this paper as follows:

- Validation Data = Data never used to create the model, but used to check the model's performance.
- Training Data = Data used to create the model, but not used to check the model's performance.
- Testing Data = Data not used to create the model, but used to check the model's performance.

Note that the training and testing data can be interchanged or varied. An example of this occurs during k-folds CV. The totality of the training and testing data will be referred to as $\mathcal{T}$, or the 'T-set'.

3.2 Intuition for STORKC

The problem that STORKC attempts to solve is the following: assume you have an adequately size data set composed of two subpopulations to perform k-folds CV for a given k. Such a setup implies that stratified CV would be the optimal choice [11]. Presume, however, that one of the subpopulations is grossly larger than the other to the point where the proportions cannot be made for a given number of folds. Under this circumstance, stratified cross-validation cannot be utilized.

Recall that stratified CV calls for the k-folds to preserve the proportions of the subpopulations in the $\mathcal{T}$[11]. For example, if the $\mathcal{T}$ had 96 from subpopulation one and 4 from subpopulation two, then the k-folds must have 0.96 of each fold to be from subpopulation one and 0.04 from subpopulation two. In the case of k=10, stratified CV cannot be achieved. Note that each of the folds will have 10 observations. If there is one observation from subpopulation two, then the proportion for subpopulation two in that given fold will be $\frac{1}{10}$ = 0.10 which is greater than 0.04. On the other hand, if there is zero from subpopulation two, then the given fold will have a corresponding proportion of 0 which is less than 0.04. Thusly, stratified cross validation cannot be utilized.

What is proposed instead is to over-represent the smaller subpopulation. In essence, we ensure that in every fold of the $\mathcal{T}$, the proportion of observations from the smaller subpopulation is larger than what stratified CV demands. For instance, in the previous example explained above, the proportion from subpopulation two would be 0.10, which implies that the number of observations from subpopulation two is 1. This implies that subpopulation one would contain the remaining proportion. Note that this approach can be applied in cases where the number of subpopulatoins is greater than 2. We would simply continue to round up in the proportions for each subpopulation until the smaller strata have at least one observation in each fold. This requires that the number of subpopulations is less than the sample size in each fold.

This schema requires careful consideration in how the smaller subpopulation observations are sampled. Most importantly, should the algorithm sample with or without replacement for the smaller subpopulation? For example, in Efron's paper on the bootstrap, he sampled with replacement [4]. There he treated the observed data as the best estimate of the distribution of the population's distribution. Thus, each observation has a unique chance of being pulled $\frac{1}{N}$, where N is the total sample size. Further, one will be pulling observations for a large number of times from the data. This provides some nice sound statistical properties.

However, the problem proposed has a very important difference. Here, there are only a few observations and only a few pulls. For instance, in the previously constructed example, we only have four observations from subpopulation two and only 10 folds. Thus, the data to represent the population is only of size 4 and we will only have 4 runs. This does not allow for a large sample properties and theoretical implications to come into effect. Further, the sample size of 4 may in fact look very different from the actual true subpopulation.

Due to these complications, the following sampling schema is proposed: for the larger subpopulation, assign observations from $\mathcal{T}$as normal. For the smaller subpopulation, randomly pull without replacement. If it is the case the smaller subpopulation sample size, n_2, is smaller than k, then replace all of n_2, when all have been assigned to different folds, and repeat the sampling without replacement. Continue this process until every fold has the specified number of observations. In a way, one can think of STORKC as a variation of CV utilizing a combination of sampling with and without replacement for the creation of the k-folds.

For instance, if n_2 = 4 and the observations are labeled A, B, C, D, a possible ordering of the observations for 10-folds is:

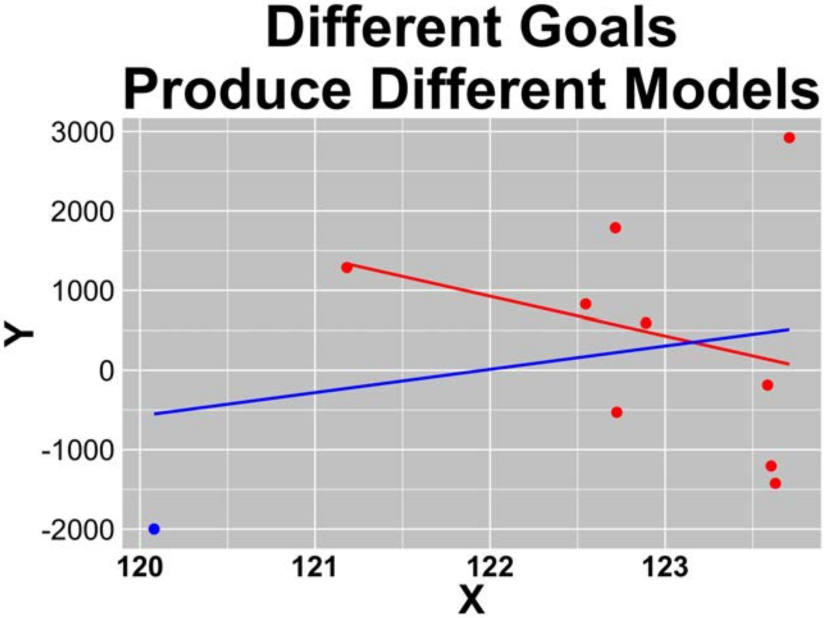
$$D, B, A, C, C, B, A, D, B, A$$

where the first D belongs to the first fold, the first B belongs to the second fold and so on.

Why perform this more complicated sampling schema? Since there are so few observations with a high chance of being noisy imperfect data, each observation will provide a different kind of insight about what the true smaller subpopulation behaves. Thus, each model will account for a different "scenario" of the smaller subpopulation. When all k models are created, those models will be better able to handle those cases from the smaller subpopulation than those models never trained on them on the first place. Thus, ensuring that all observations are represented as equally as possible is important.

Figure 1 showcases a toy example of 10 simulated data points. The red and blue lines are two simple linear regression lines of best fit. The red (or gray) line accounts for only the

Figure 1: 10 simulated observations where X is the independent variable and Y is the dependent variable. The red (or gray) and blue (or dark gray) lines are two simple linear regression lines of best fit. The red line accounts for only the red points while the blue line accounts for the red and blue points. One can think of the red dots as observations from subpopulation 1 and the blue dot as an observation from subpopulation 2.



red points while the blue (or dark gray) line accounts for the red and blue points. STORKC is intended to create lines similar to the blue line while k-fold CV will most likely create lines like the red line. Thus, it is question of modeling intention. The modeler must decide if the final model is meant to describe the typical observation from the population or all observations from the population. While the red line accounts for the typical observation better, the blue line will better predict the population more accurately.

Here is a practical example: suppose it is desired to model the lung capacity of runners where Y = lung capacity and X = average pace in a 10 kilometers run. Presume that the data collected looks similar to Figure 1. Suppose that the blue observation has asthma while the red observation are non-asthmatic runners. Only considering simple linear regression as a modeling technique, should the analyst jackknife the blue observation or not? If the purpose of the analysis is to model the typical observation, then yes the analyst should. However, if the purpose of the model is to describe the relationship of all runners, then the modeler should not jackknife the blue runner. The approach is the same while considering STORKC and traditional CV. One needs to verify the modeling intent is unified before deciding on which method to use. Thus, careful considerations must be made in regards to the choice of using traditional k-folds CV or STORKC.

Others have encountered different modeling situations were the need for a method like STORKC could have been implemented. For example, work has been done one building models to classify pill shapes [12]. However, some less common pill shapes, like hexagons and octagons, have less observations from which to build a model. For instance, these shapes were collected less than 10 times combined. This is dramatically smaller than the 979 round pills [12]. Thus, STORKC would provide a common methodology to build models in these types of scenarios.

3.3 Modeling Considerations

What has been said thus far has been primarily concerned with the case where the response variable is some type of measure like amount of money made, a continuous random variable, or number of complete boxes crushed, a discrete random variable. No considerations have been made concerning classification problems such as cancer or no cancer. Special care must also be made for these cases. However, STORKC can still be applied without loss of generality. Furthermore, STORKC is applicable in a variety of different modeling techniques such as SVM or neural networks.

3.4 Formal Setup

Let K = the number of roughly equal-sized parts from the $\mathcal{T}$. During the kth trial, the remaining $k - 1$ parts will be utilized as the training set. The kth part will be utilized as the testing set [8]. This is performed K times. We then combine the predictions to obtain the prediction error [8].

Let $\hat{f}^{-k}(x)$ = the fitted function, where the function is computed without the given kth set (the training data). Let N = the total sample size of the $\mathcal{T}$. Also let $L()$ = the loss function [8]. In this case, the residual sum of squares (RSS) will be utilized as the loss function, but without loss of generality. Then, the cross-validation estimate of the prediction error for the $\mathcal{T}$is

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-k}(x_i)) \tag{1}$$

or when utilizing RSS, the following is obtained

$$RSS(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{f}^{-k}(x_i))^2 \tag{2}$$

Afterwards, the validation set for the CV prediction error, or RSS, is computed in a similar fashion if applicable [8].

3.5 STORKC Formalization

STORKC has a similar setup. However, there are a few distinct differences. Let there be p subpopulations where l is the largest. Here, we will consider $p = 2$, without loss of generality. Let X_1 represent the larger subpopulation. Let X_2 represent the smaller subpopulation. Let N_1 and N_2 be the total size of the first and second subpopulations, respectively, where $N = N_1 + N_2$. Thus, when formulating the $\mathcal{T}$, X_1 will be partitioned in the usual way. However, X_2 will be allocated in such a way that at least one observation is made in each of the k partitions. Otherwise, the number of times an observation appears in each of the partitions should be as close the perfect stratified proportion as possible. This simplifies to

$$STORKC(\hat{f}) = \frac{1}{N}[\sum_{i=1}^{N_1} L(y_{i,1}, \hat{f}^{-k}(x_{i,1})) + \sum_{l=1}^{N_2} L(y_{l,2}, \hat{f}^{-k}(x_{l,2}))] \tag{3}$$

or when utilizing RSS, the following is obtained

$$RSS_{STORKC}(\hat{f}) = \frac{1}{N}[\sum_{i=1}^{N_1}(y_{i,1} - \hat{f}^{-k}(x_{i,1}))^2 + \sum_{l=1}^{N_2}(y_{l,2} - \hat{f}^{-k}(x_{l,2}))^2] \tag{4}$$

X_2 will be sampled without replacement until all of the observations from X_2 are utilized. What is possible is there may not be enough observations from subpopulation X_2 to allocate to all the different K partitions. In this case, once all of the observations from X_2 have been partitioned, repartition the observations from the X_2 until all the K folds have been created. Note that this may in turn have that not all the observations from X_1 utilized.

4. Evaluation

4.1 Simulated Experiment 1

The simulated data will have 100 observations for the $\mathcal{T}$ and 100 observations for the validation set. The $\mathcal{T}$ will be composed of 96 random observations from a $N(100,1)$ and 4 random observations from a $N(-1880,1)$. The number of folds, k, will be fixed to 10. Thus, using traditional stratified cross-validation is unobtainable. STORKC will be utilized instead where 9 observations will be from the first subpopulation and 1 observation from the second subpopulation in each of the folds. The true relationship is defined to be a simple linear regression relationship that is specifically

$$Y = 13X$$

This will be performed 10,000 per each of the different levels of noise, j, where

$$Y = 13X + \epsilon$$

where ϵ is distributed as $N(0,j^2)$ and $j \in \{1,2,...,10\}$ is the standard deviation. Thus, there will be a total of 100,000 simulations performed for tradition k-folds cross validation and STORKC. At each of the m simulations, the RSS is recorded. Table 1 and Figure 2 shows the means of the RSS prediction error for the simulations. Table 2 and Figure 3 shows the variances of the RSS prediction error for the simulations alongside the expected value for RSS. Base R was utilized to perform the simulations, and ggplot2 was used to create the plots in conjunction with reshape, xtable, scales, and GGally.

4.2 Simulated Experiment 2

The second simulated data will be similar to the first. It will have 100 observations for the $\mathcal{T}$and will be composed of 96 random observations from a $N(100,1)$ and 4 random observations from a $N(-1880,1)$. The number of folds, k, will be fixed to 10. Thus, using traditional stratified cross-validation is unobtainable. STORKC will be utilized instead where 9 observations will be from the first subpopulation and 1 observation from the second subpopulation in each of the folds. The true relationship is defined to be a simple linear regression relationship that is specifically

$$Y = 13X - 1000Z$$

where

$$Z = \begin{cases} 1 & ; X \in X_2 \\ 0 & ; else \end{cases}$$

Figure 2: Figure summarizing Table 1 for the first experiment. Notice that traditional CV is unstable as the noise level increases. However, STORKC remains constant at a fairly predictable rate.

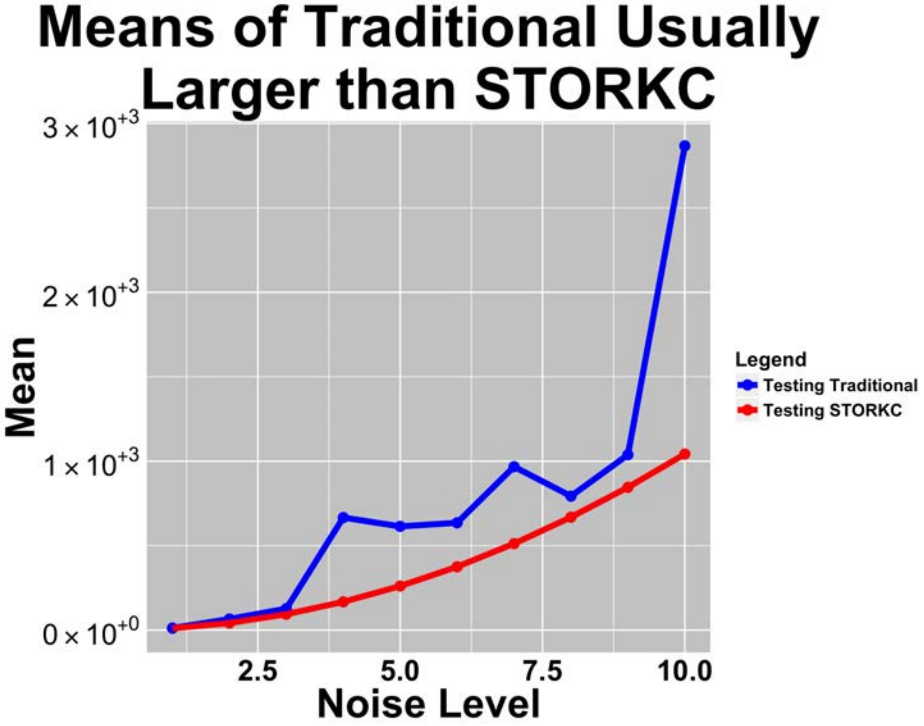


Figure 3: Figure summarizing Table 2 for the first experiment. Note that traditional CV's variance is chaotic and unpredictable while STORKC is fairly constant and consistent.

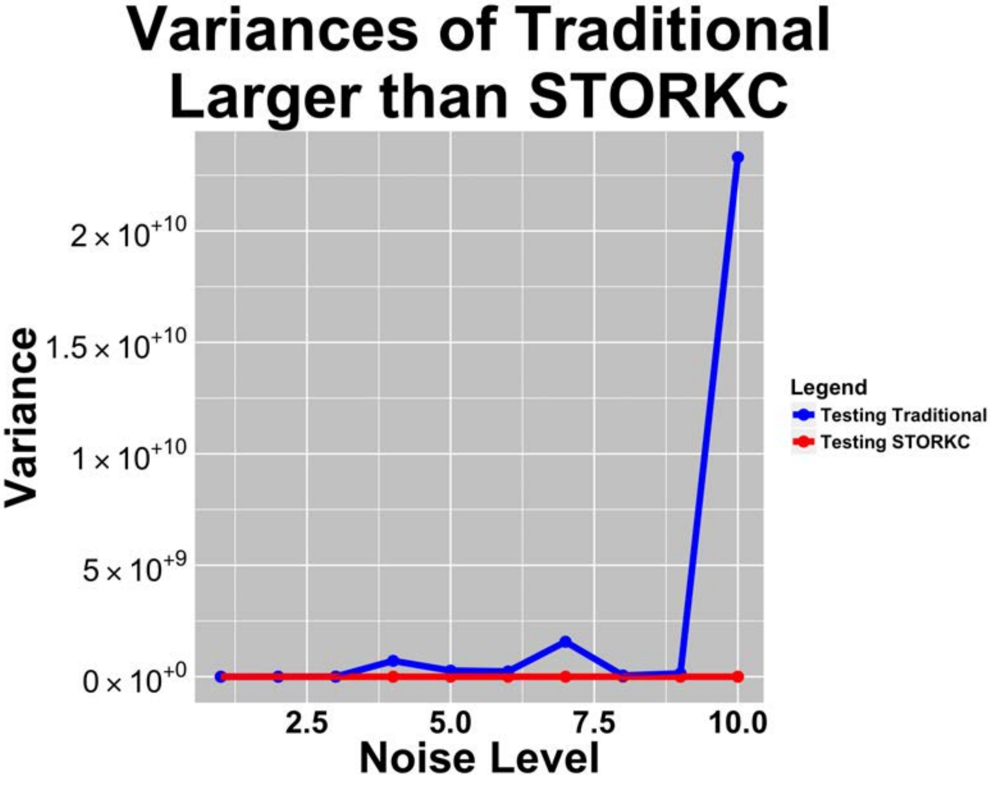


Table 1: Table including the means rounded to 2 significant digits of the RSS prediction errors of each of the 10,000 simulations at each of the 10 noise levels for the first experiment. The means are reported for both the $\mathcal{T}$sets for both traditional cross-validation (CV) and STORKC. The exact expected value for RSS, E[RSS], is given in the last column.

Noise Level	STORKC	Traditional CV	E[RSS]
1	1.31×10^{01}	2.51×10^{02}	9.8×10^{1}
2	4.44×10^{01}	2.33×10^{02}	3.92×10^{2}
3	9.62×10^{01}	2.75×10^{02}	8.82×10^{2}
4	1.69×10^{02}	3.27×10^{02}	1.568×10^{3}
5	2.63×10^{02}	3.89×10^{02}	2.450×10^{3}
6	3.79×10^{02}	5.46×10^{02}	3.528×10^{3}
7	5.12×10^{02}	1.11×10^{03}	4.802×10^{3}
8	6.69×10^{02}	7.44×10^{02}	6.272×10^{3}
9	8.50×10^{02}	2.20×10^{03}	7.938×10^{3}
10	1.05×10^{03}	1.46×10^{03}	9.800×10^{3}

Table 2: Table including the variances of the RSS prediction error of each of the 100,000 simulations at each of the noise levels for the first experiment. The variances are reported for the testings sets for both traditional cross-validation (CV) and STORKC.

Noise Level	STORKC	Traditional CV
1	4.91×10^{00}	9.65×10^{07}
2	5.46×10^{01}	2.50×10^{08}
3	2.65×10^{02}	2.62×10^{08}
4	7.90×10^{02}	2.15×10^{08}
5	1.95×10^{03}	1.39×10^{08}
6	4.07×10^{03}	1.62×10^{08}
7	7.29×10^{03}	1.50×10^{09}
8	1.28×10^{04}	6.51×10^{07}
9	2.07×10^{04}	3.00×10^{09}
10	3.13×10^{04}	7.40×10^{08}

Thus, the observed model will be

$$Y = 13X - 1000Z + \epsilon$$

However, we will have our estimated model be $Y = 13X + \epsilon$. We run the model in this manner as some of the folds will not be containing any observations from X_2 in the traditional CV models. Thus, Z would not be able to be estimated in those cases. Thusly, Z is removed from the model in this simulation.

The simulation will be performed 10,000 per each of the different levels of noise, j, where

$$Y = 13X + \epsilon$$

where ϵ is distributed as $N(0, j^2)$ and $j \in \{1, 2, ..., 10\}$ and is the standard deviation. Thus, there will be a total of 100,000 simulations performed for tradition k-folds cross validation and STORKC. At each of the m simulation, the RSS is recorded. Table 3 and Figure 4 shows the means of the RSS prediction error for the simulations. Table 4 and Figure 5 shows the variances of the RSS prediction error for the simulations alongside the expected value for RSS.

Table 3: Table including the means rounded to 2 significant digits of the RSS prediction errors of each of the 10,000 simulations at each of the 10 noise levels for the second experiment. The means are reported for both the $\mathcal{T}$sets for both traditional cross-validation (CV) and STORKC. The exact expected value for RSS, E[RSS], is given in the last column.

Noise Level	STORKC	Traditional CV	E[RSS]
1	1.31×10^{01}	2.51×10^{02}	9.8×10^{1}
2	4.44×10^{01}	2.33×10^{02}	3.92×10^{2}
3	9.62×10^{01}	2.75×10^{02}	8.82×10^{2}
4	1.69×10^{02}	3.27×10^{02}	1.568×10^{3}
5	2.63×10^{02}	3.89×10^{02}	2.450×10^{3}
6	3.79×10^{02}	5.46×10^{02}	3.528×10^{3}
7	5.12×10^{02}	1.11×10^{03}	4.802×10^{3}
8	6.69×10^{02}	7.44×10^{02}	6.272×10^{3}
9	8.50×10^{02}	2.20×10^{03}	7.938×10^{3}
10	1.05×10^{03}	1.46×10^{03}	9.800×10^{3}

Table 4: Table including the variances of the RSS prediction error of each of the 100,000 simulations at each of the noise levels for the second experiment. The variances are reported for the testings sets for both traditional cross-validation (CV) and STORKC.

Noise Level	STORKC	Traditional CV
1	4.91×10^{00}	9.65×10^{07}
2	5.46×10^{01}	2.50×10^{08}
3	2.65×10^{02}	2.62×10^{08}
4	7.90×10^{02}	2.15×10^{08}
5	1.95×10^{03}	1.39×10^{08}
6	4.07×10^{03}	1.62×10^{08}
7	7.29×10^{03}	1.50×10^{09}
8	1.28×10^{04}	6.51×10^{07}
9	2.07×10^{04}	3.00×10^{09}
10	3.13×10^{04}	7.40×10^{08}

Figure 4: Figure summarizing Table 3 for the second experiment. Notice that traditional CV is unstable as the noise level increases. However, STORKC remains constant at a fairly predictable rate.

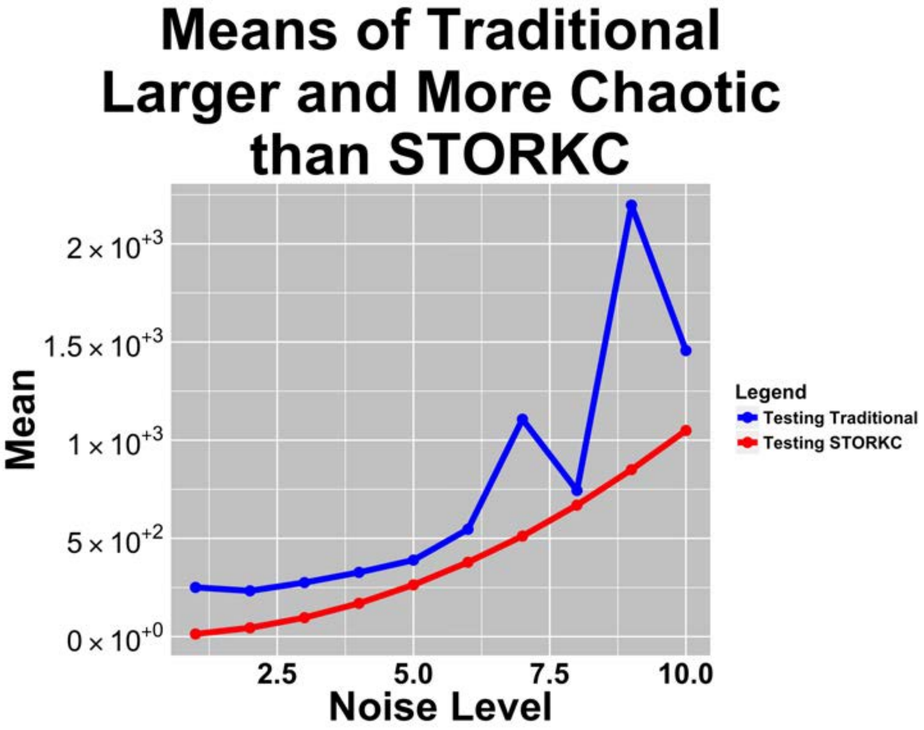
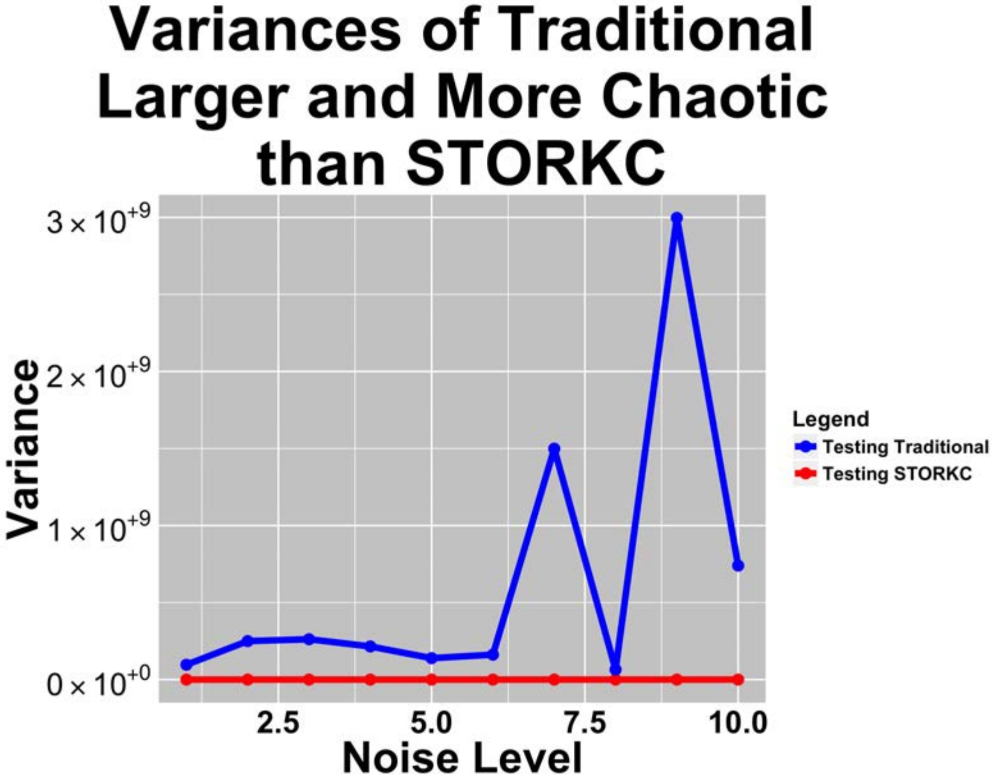


Figure 5: Figure summarizing Table 4 for the second experiment. Note that traditional CV's variance is chaotic and unpredictable while STORKC is fairly constant and consistent.



4.3 Experiment on the Iris Data

The famous Fisher Iris data set provided in base R was utilized to compare the two methods [6]. A similar approach is utilized as the simulation setup. Simple linear regression was still utilized, where a subset of the virginica and setosa species were used. 25 observations from the setosa subpopulation and 5 observations from the virginica subpopulation are utilized in the $\mathcal{T}$. Only the petal width and petal length variables are used where

$$\text{Petal Width} = \beta_0 + \beta_1(\text{Petal Length})$$

This is performed 10,000 times using $k = 10$ folds utilizing both traditional cross-validation and STORKC. The only major difference between the iris data runs and the artificial data simulations is that the runs are not done over different noise levels. Table 5 summarize the results for the testing data sets.

Table 5: Table including the summary results for the RSS prediction error mean, variance, and Q3 for the testing data sets.

RSS Prediction Error	STORKC Testing	Regular Testing
Mean	9.549×10^{-2}	8.778×10^{-2}
Variance	1.1791×10^{-3}	1.203×10^{-3}

5. Discussion

For the first experiment, Figures 2 and 3 show that traditional CV is more chaotic than STORKC. While in many cases, the means of the Traditional CV RSS is closer to the expected values of the RSS, the variation of the actual values will greatly differ given the composition of the data. However, STORKC will given a much more consistent and reliable estimate, despite being generally overly optimistic on the estimated value of the RSS. However, the fact that a method utilizing some form of CV and providing an optimistic value for the RSS is not a new phenomena [9]. Providing a method that in fact is more consistent in the value it provides is worthwhile, which is what STORKC is able to achieve.

The second experiment shows similar results as the first. In short, STORKC is more stable and predictable than traditional CV. This applies to both situations simulated. The first was the case of attempting to model the true model, and the second was the case of missing one of the actual variables in the model. There are some cases where the two methods give the same values. This is more likely to occur at lower noise levels. However, STORKC is more does not provide any surprises or oddities. It is a much more stable method for assessing model performance, which traditional k-folds CV is volatile in other circumstances.

The iris data experiment shows that the STORKC and traditional k-folds CV show similar values. In this case, the analyst to utilize what they deem as the more appropriate method. For instance, if the analyst is attempting to create a model that will be consistently reproduced for the population, STORKC may be the better choice.

Some may argue that STORKC may overfit the model to the data. This may in fact occur. However, it is still believed that given subpopulations may be known and desired to be accounted for during the model building process in some fashion. The reasons for this can vary by the different applications. For instance, a model may desire to classify pill shapes, but simply foes not have enough observations for each subpopulation [12]. Without

STORKC or some other method, models will be required to extrapolate needlessly. Thus, while overfitting may be an inherent risk, STORKC allows for models to be built that would otherwise be unable to be created.

6. Conclusion

There is promise that STORKC has the capacity to improve models more consistently. While some of the theoretical properties have be investigated for k-folds CV, this is warranted for STORKC [3, 19]. Additional investigations comparing leave on out cross validation may also be worthwhile.

Furthermore, this modeling technique favors the models built for the population, rather the those models meant for the typical observation. However, even with noisy data, STORKC performs consistently well under these conditions. Further investigations are needed to see if other sampling variations based upon the spirit of k-folds CV show improvement.

Additionally, further investigations with more variables are warranted. The performance of STORKC under these more complicated conditions is imperative to many real life applications. Further, investigating the potential of STORKC to be utilized in conjunction with the least absolute shrinkage and selection operator, the LASSO, and other methods of variable selection like ridge regression are worthwhile investigations [1, 7, 17].

Acknowledgements

We would like to thank Dr. William G. Kennedy, John S. Schuler, and Jon Murphy for their helpful comments, critiques, and suggestions.

References

[1] David M. Allen. The Relationship between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, 16(1):125–127, 1974.

[2] Leo Breiman and Philip Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, 60(3):291–319, 1992.

[3] Kehui Chen and Jing Lei. Network Cross-Validation for Determining the Number of Communities in Network Data. *Journal of the American Statistical Association*, 113(521):241–251, January 2018.

[4] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[5] Yang Feng and Yi Yu. Consistent Cross-Validation for Tuning Parameter Selection in High-Dimensinoal Variable Selection. *arXiv:1308.5390 [stat]*, August 2013. arXiv: 1308.5390.

[6] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[7] Gene H. Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, 1979.

[8] Trevor Hastie, Tibshirani Robert, and Friedman Jerome. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd (corrected 12th printing) edition, January 2017.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors. *An introduction to statistical learning: with applications in R*. Number 103 in Springer texts in statistics. Springer, New York, 2013. OCLC: ocn828488009.

[10] Yoonsuh Jung and Jianhua Hu. A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, 27(2):167–179, April 2015.

[11] Ron Kohavi. A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Articial Intelligence*, page 7, 1995.

[12] K. T. Maddala, R. H. Moss, W. V. Stoecker, J. R. Hagerty, J. G. Cole, N. K. Mishra, and R. J. Stanley. Adaptable Ring for Vision-Based Measurements and Shape Analysis. *IEEE Transactions on Instrumentation and Measurement*, 66(4):746–756, April 2017.

[13] Frederick Mosteller and J.W. Tukey. Data analysis, including statistics. 1968. OCLC: 80947523.

[14] Frederick Mosteller and David L. Wallace. Inference in an Authorship Problem. *Journal of the American Statistical Association*, 58(302):275–309, June 1963.

[15] Kevin R. Murphy. Fooling Yourself with Cross-Validation: Single Sample Designs. *Personnel Psychology*, 36(1):111–118, March 1983.

[16] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[17] Tibshirani Robert. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, April 2011.

[18] Lorenzo Trippa, Levi Waldron, Curtis Huttenhower, and Giovanni Parmigiani. Bayesian nonparametric cross-study validation of prediction methods. *The Annals of Applied Statistics*, 9(1):402–428, March 2015.

[19] Ping Zhang. Model Selection Via Multifold Cross Validation. *The Annals of Statistics*, 21(1):299–313, March 1993.