A Unified Statistical Framework for Evaluating Predictive Methods

Completed Research Paper

Patrick Urbanke
University of Göttingen
Platz der Göttinger Sieben 5
37073 Göttingen
Germany
patrick-axel.urbanke@wiwi.uni-goettingen.de

Johann Kranz
University of Göttingen
Platz der Göttinger Sieben 5
37073 Göttingen
Germany
jkranz@uni-goettingen.de

Lutz Kolbe
University of Göttingen
Platz der Göttinger Sieben 5
37073 Göttingen
Germany
lkolbe@uni-goettingen.de

Abstract

Predictive analytics is an important part of the business intelligence and decision support systems literature and likely to grow in importance with the emergence of big data as a discipline. Despite their importance, the accuracy of predictive methods is often not assessed using statistical hypothesis tests. Furthermore, there is no commonly agreed upon standard as to which questions should be examined when evaluating predictive methods. We fill this gap by defining three questions that involve the overall and comparative predictive accuracy of the new method. We then present a unified statistical framework for evaluating predictive methods that can be used to address all three of these questions. The framework is particularly versatile and can be applied to most problems and datasets. In addition to these practical advantages over hypotheses tests used in previous literature, the framework has the theoretical advantage that it is not necessary to assume a normal distribution.

Keywords: Machine learning, statistical methods, predictive modelling, business intelligence, decision support systems

Introduction

Predictive analytics is an important part of information systems research and likely to grow in importance with the emergence of big data as a discipline. It involves the use of quantitative techniques, generally machine learning algorithms, to build predictive models (Shmueli and Koppius 2011). It plays an important role in the business intelligence (Watson and Wixom 2007) and decision support systems literature and is highly relevant to both researchers and practitioners.

Despite the importance of predictive analytics, we notice that the accuracy of predictive methods is often not assessed using

proper statistical hypothesis tests and that the literature lacks a commonly agreed upon standard as to which questions should be examined when evaluating predictive methods. The purpose of this paper is to fill this gap by introducing three questions to consider when evaluating a predictive method. In addition, we present a unified statistical framework that can be used to address all three of these questions and is applicable to a wide variety of problems and datasets.

Shmueli and Koppius (2011) emphasise the importance of predictive analytics in information systems research and contrast predictive analytics with explanatory statistical modelling. They argue that predictive accuracy should be evaluated on the basis of out-of-sample predictions using measures such as root mean squared error (RMSE) or mean absolute percentage error (MAPE). They advocate the use of predictive analytics not only for the development of decision support systems, but also for theory building or the evaluation of theoretical models. See Shmueli (2010) for a thorough discussion of philosophical issues related to this approach.

The framework proposed in this paper is based on their findings, especially their emphasis on the importance of an out-of-sample evaluation. We extend the approach by introducing a unified statistical framework for evaluating out-of-sample predictions and precisely defining the questions to be examined when evaluating predictive methods. Our framework is meant to be used both in the more traditional context of evaluating decision support systems as well as the more modern context of evaluating competing theories.

In our view there are three questions that are of interest when evaluating a predictive method:

1. Does the predictive method generate statistically significant out-of-sample predictions?

2. Do the out-of-sample predictions outperform out-of-sample predictions generated by alternative methods in a statistically significant manner?

3. Are the out-of-sample predictions generated by the predictive method statistically significant *corrected for* the predictions generated by alternative methods?

We argue that the examination of all of these three questions is necessary for a thorough and rigorous evaluation of a predictive method.

Some might argue that the first question is already implied by the second and does not require explicit examination. We disagree for two reasons: First, we might be comparing the new predictive method to a method that is so poor that it does worse than a random walk. In that event, the new method might outperform the old one, even though it does not outperform a random walk. Second, it is in itself interesting to know which of the methods evaluated is actually useful. If the best method cannot be used (maybe because it is computationally too expensive), we would like to know which of the alternative methods evaluated might be an appropriate substitute.

The second question is the one question most papers focus on. In predictive analytics, it is common practise to compare the out-of-sample predictions of a newly introduced predictive method with the out-of-sample predictions of state-of-the-art approaches. We agree with the importance of doing so. However, we posit that such comparisons should be supported by statistical hypothesis testing to attain scientific rigour.

To see the importance of the third question, suppose that method A is outperformed by method B. This does not automatically imply that method A is useless. If it can be shown that method A can make correct predictions when method B fails to do so and thus compensates its weaknesses, the two methods can be combined to an ensemble that generates more accurate predictions than each of the individual methods. However, if the correct predictions generated by method A are mainly a subset of the correct predictions generated by method B, then method A is less useful.

In our view, anyone introducing a new predictive method should demonstrate that both question 1 and either question 2 or question 3 are in the affirmative with regards to the newly introduced method. If so, the novel method has been shown to be useful and to constitute a contribution to the literature.

In this paper, we present a unified statistical framework for hypothesis testing that can be used to address all of the three question mentioned above. It can be used for regression, classification and multi-label classification problems. Predictions can be both continuous or discrete. It can be used to test individual predictive methods or several predictive methods at once. It is even applicable when the assumption of a normal distribution fails.

The framework thus has several theoretical and practical advantages over the hypotheses tests used in the current predictive

analytics literature.

The remainder of this paper is organised as follows: In the next section, we examine how the three questions have been evaluated in recent predictive analytics literature. We then introduce the mathematical concept underlying the framework. Afterwards, we evaluate the framework by comparing it to a more traditional approach of evaluating predictive methods. Finally, we discuss our findings and show how our framework can be used to address the three critical issues that need to be considered when evaluating predictive methods. We then conclude.

Literature Review

We carefully reviewed the recent literature on predictive analytics focusing on how each of the three questions listed above has been evaluated. An overview is given in Table 1.

Table 1. Evaluation of Statistical Significance in Predictive Analytics

	Method	**Author(s)**
Accuracy of predictive method	t-test	Yen and Hsu 2010
	Unclear	Schumaker 2013
	Evaluated without using a hypothesis test	David et al. 2012; Hagenau et al. 2013; Lee et al. 2011
Accuracy of predictive method in comparison to alternative methods	Analysis of variance (ANOVA)	Zhao et al. 2011
	Diebold-Mariano test	Sermpinis et al. 2012
	t-test	Abbasi et al. 2010, 2012; Cao et al. 2012; Carbonneau et al. 2011; Chan and Franklin 2011; Khansa and Liginlal 2011; Kim et al. 2011; Li et al. 2012; Li and Chen 2013; Oh and Sheng 2011; Sahoo et al. 2012; Yang et al. 2010
	Wilcoxon signed-rank rest	Kao et al. 2013; Lu et al. 2012
	Unclear	Bhattacharyya et al. 2011; Cui et al. 2012; Du Jardin and Séverin 2011; Li et al. 2014
	Evaluated without using a hypothesis test	Bai 2011; Bao et al. 2013; Choi et al. 2011, 2013; Delen 2010; Delen et al. 2013; Gerber 2014; Hagenau et al. 2013; Kisilevich et al. 2013; Lau et al. 2013; Lee et al. 2011, 2012; Li and Wu 2010; Lu et al. 2012; Olson and Chae 2012; vd Reijden and Koppius 2010; Serrano-Cinca and Gutiérrez-Nieto 2013; Shin et al. 2013; Su et al. 2012; Yolcu et al. 2013
Accuracy of predictive method corrected for alternative methods	Evaluated without using a hypothesis test	Serrano-Cinca and Gutiérrez-Nieto 2013

We find that only few papers explicitly evaluate the overall accuracy of the predictive method and even fewer employ statistical hypothesis tests to do so. This might be partly attributable to the fact that standard accuracy measures such as RMSE or MAPE, as proposed by Shmueli and Koppius (2011), are difficult to interpret in an absolute sense and are more suitable for a relative comparison of different predictive methods. Nevertheless, we believe this to be problematic as

knowing which of the evaluated methods generates statistically significant predictions is necessary in order to advance scientific knowledge on the specific problem domain. We would argue that if we cannot demonstrate that the predictive method we propose generates statistically significant out-of-sample predictions, said method may be of little practical or academic relevance.

The most common proof-of-concept in the literature is to compare the predictions generated by the proposed predictive method to existing state-of-the-art methods in order to demonstrate that it constitutes an improvement. Whereas we agree with the need to compare new methods with existing state-of-art approaches, we note that it is often unclear whether the outperformance is statistically significant. We argue that it is necessary to examine the statistical significance of an out-of-sample outperformance in order to attain scientific rigour: If we cannot demonstrate that the outperformance of our proposed method is statistically significant, we cannot be certain that this outperformance is not simply attributable to the particular settings of one experiment such as the specific dataset used.

Finally, only one of the papers we reviewed (Serrano-Cinca and Gutiérrez-Nieto 2013) explicitly examined the accuracy of the predictive methods corrected for alternative methods. However, this examination was conducted without using a statistical hypothesis test. We argue that an examination of the accuracy of the predictive methods corrected for alternative methods is a good alternative to the concept of outperformance: If a new predictive method does not demonstratively outperform existing methods, then it might still be useful and interesting if it can be shown that the out-of-sample accuracy of the predictive method corrected for existing methods is still statistically significant. Such a finding implies that the new method compensates weaknesses of existing methods and the methods could be combined to form an ensemble that generates more accurate predictions.

The paired t-test is by far the most popular statistical hypothesis test in previous literature (see Table 1). Other tests used are the Wilcoxon signed-rank test, ANOVA or the Diebold-Mariano test (Diebold and Mariano 2002), which is specifically developed for time series. However, these tests have a number of practical and theoretical shortcomings that undermine their applicability and reliability: First, existing statistical hypothesis tests cannot be used for multiple label classification problems. Second, when constructing an ensemble of algorithms, we might be interested in the predictive accuracy of the combined algorithms before training any specific MetaLearner. Existing statistical hypothesis tests cannot be used for such purposes. Third, existing statistical hypothesis tests cannot be used to evaluate the accuracy of predictive methods corrected for alternative methods. This is of particular relevance for examining question 3. Fourth, existing statistical hypothesis tests rest on the assumption of a normal distribution. If that assumption fails, the tests are no longer reliable. Even though this is a well-known fact, little attention has been paid to the issue in previous literature. This implies that potentially unreliable statistical hypothesis tests have been applied.

The framework for statistical hypothesis testing proposed in this paper does not suffer from any of these shortcomings. Moreover, it can be used for regression, classification and multi-label classification problems. Predictions can be both continuous or discrete. Most importantly, it offers a unified framework to evaluate the out-of-sample performance of predictive methods along the lines of all of the three questions listed above.

Calculation

Mathematically, the statistical hypothesis test consists of three propositions, which we introduce in this section. Detailed mathematical proofs for the propositions are presented in Appendix 1.

Basic Idea of the Statistical Hypothesis Test

Suppose we have m continuous or discrete random variables X_1, X_2, X_3,..., X_m. The variables X_a, a=1,2,...m, are drawn without replacement from a dataset of out-of-sample predictions of m predictive methods. Further suppose that we have n continuous or discrete dependent random variables Y_1, Y_2, Y_3,..., Y_n. The variables Y_c, c=1,2,...n, are drawn without replacement from a dataset of the values for the testing set that the predictive methods have been trained to predict. In most

regression problems, we would find that m=n=1, but for multi-label classification problems the values for m and n are greater than one.

Let x^i_a denote the i^{th} value in the dataset associated with variable X_a. Let y^j_c denote the j^{th} value in the dataset associated with variable Y_c. Let N be the number of instances in the dataset.

Define the variable X^c_a as follows:

$$X^a_c = \sum_{i=1}^{N} x^i_a y^i_c \tag{1}$$

We interpret X^c_a as a random variable that has been constructed by randomly drawing from the two datasets without replacement. We then establish the null hypothesis that the variables X_1, X_2, X_3,..., X_m do not constitute statistically significant predictions for the variables Y_1, Y_2, Y_3,..., Y_n. In other words, our null hypothesis is that the out-of-sample predictions of the predictive methods have no predictive value:

$$E(X^a_c) = \frac{\sum_{i=1}^{N} x^i_a \sum_{j=1}^{N} y^j_c}{N} \forall a, c \tag{2}$$

Let **X** be a vector containing the X^c_a for all a and c, with one exception: Where ever a subset of the random variables X^c_a is linearly dependent, we exclude one of these variables from **X**. (See Proposition 1 for the rationale.)

We then approximate X^c_a using a normal distribution. Note that we are *not* assuming any specific underlying distribution of the variables X_a and Y_c themselves. See below for an alternative method to be used when the normal approximation is judged to be too inaccurate.

Let **V** be the variance-covariance matrix of **X**. It follows from our assumptions that we can model **X** using a chi-squared distribution:

$$(\mathbf{X} - E(\mathbf{X}))' \mathbf{V^{-1}} (\mathbf{X} - E(\mathbf{X})) \sim \chi^2(n) \tag{3}$$

Having introduced the basic idea of the statistical hypothesis test, we are left with three questions:

1. In order to ensure that **X** is not linearly dependent, we have left out some variables. Does this make our model arbitrary?

2. How do we calculate the variances and covariances in **V**?

3. How do we evaluate the statistical significance of a prediction given the predictions of other predictive methods?

These three questions are addressed in Propositions 1, 2 and 3 respectively.

Proposition 1: Proof of Non-Arbitrariness of the Model

As mentioned above, we do not include X^c_a in **X** if the inclusion would cause **X** to contain a subset of variables that are linearly dependent. The reason for this is simple: If there were a subset of variables in **X** that are linearly dependent, **V** would be a singular matrix and $\mathbf{V^{-1}}$ could not be found.

However, if there is a subset of variables that are linearly dependent, we have to eliminate one of these variables arbitrarily. This raises a very important question: If there is no justification which of these variables to eliminate, and we are thus forced to eliminate one of these variables arbitrarily, is our model outcome arbitrary?

Proposition 1 states that the model outcome is not arbitrary, because *it does not matter which of these variables is eliminated, the model outcome will always be the same*.

Intuitively we can interpret this proposition as follows: We know that a set of variables is linearly dependent, then taking away one variable does not eliminate any information, because the last variable is automatically implied by all the others.

More formally, we can express our proposition as follows: Consider n+1 random variables, X_1, X_2, X_3,..., X_{n+1} for which $X_1 + X_2 + X_3 +...+ X_{n+1} = 0$. Consider an additional set of random variables, Y_1, Y_2, Y_3,..., Y_m that are not linearly dependent. Let $\mathbf{X_i}$ be a vector containing all X with the exception of X_i . Let $\mathbf{Y}$ be a vector containing all Y. Let $\mathbf{V_i}$ and $\mathbf{V_Y}$ be the variance-covariance matrix of $\mathbf{X_i}$ and $\mathbf{Y}$ respectively and let $\boldsymbol{\sigma_i}$ be the matrix containing their covariances. Proposition 1 states the following:

$$\begin{bmatrix} \mathbf{X_i'} & \mathbf{Y'} \end{bmatrix} \begin{bmatrix} \mathbf{V_i} & \sigma_\mathbf{i} \\ \sigma_\mathbf{i}' & \mathbf{V_Y} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_i} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X_j'} & \mathbf{Y'} \end{bmatrix} \begin{bmatrix} \mathbf{V_j} & \sigma_\mathbf{j} \\ \sigma_\mathbf{j}' & \mathbf{V_Y} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_j} \\ \mathbf{Y} \end{bmatrix} \forall i, j \tag{4}$$

Proof: See Appendix 1.

Proposition 2: Calculating the Elements of V

In order to calculate the variance-covariance matrix $\mathbf{V}$ of $\mathbf{X}$, we need to calculate the variances under the null hypothesis for every element X^c_a that is part of $\mathbf{X}$ and their respective covariances.

Recall that we have interpreted X^c_a as a random variable that has been constructed by randomly drawing from the dataset without replacement. We can therefore interpret this problem as a generalisation of the hypergeometric distribution. Suppose that there is another variable in $\mathbf{X}$ that we denote as X^d_b.

Proposition 2 states that $\text{cov}(X^c_a, X^d_b)$ can be calculated as follows:

$$cov(X_c^a, X_d^b) = \frac{(N\sum_{i=1}^{N} x_a^i x_b^i - \sum_{i=1}^{N} x_a^i \sum_{k=1}^{N} x_b^k)(N\sum_{j=1}^{N} y_c^j y_d^j - \sum_{j=1}^{N} y_c^j \sum_{l=1}^{N} y_d^l)}{N^2(N-1)} \forall a, b, c, d \tag{5}$$

Proof: See Appendix 1.

Note that this formula is also applicable to cases where a=b and/or c=d. When a=b and c=d the above formula is equivalent to calculating the variance of X^c_a.

Note further that in the special case in which x^i_a and can y^j_c only assume the values 0 or 1, a equals b *and* c equals d, this formula reduces to the standard formula for the variance of the hypergeometric distribution. When a equals b *or* c equals d, it reduces to the standard formula for the covariance of the hypergeometric distribution. This is consistent with our interpretation of the problem as a generalisation of the hypergeometric distribution.

Finally note that this is not an estimate of the variance-covariance matrix under the null hypothesis, but that it actually *is* the variance-covariance matrix under the null hypothesis.

Proposition 3: Evaluating Statistical Significance Corrected for Other Predictive Methods

Suppose $\mathbf{X_1}$ and $\mathbf{X_2}$ are two random variables of length n_1 and n_2 respectively. Further suppose that their expected values is $\mathbf{0}$, their variance-covariance-matrices are $\mathbf{V_1}$ and $\mathbf{V_2}$ respectively and their mutual covariance matrix is $\boldsymbol{\sigma_{12}}$.

$$\begin{bmatrix} \mathbf{X_1'} & \mathbf{X_2'} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \boldsymbol{\sigma_{12}} \\ \sigma_{\mathbf{12}}' & \mathbf{V_2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} - \mathbf{X_1'V_1^{-1}X_1} \tag{6}$$

Proposition 3 states that (6) is a squared Mahalanobis distance. Furthermore, assume that $\mathbf{X_1}$ and $\mathbf{X_2}$ are multivariate normal. Then:

$$\begin{bmatrix} \mathbf{X_1'} & \mathbf{X_2'} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \boldsymbol{\sigma_{12}} \\ \sigma_{\mathbf{12}}' & \mathbf{V_2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} - \mathbf{X_1'V_1^{-1}X_1} \sim \chi^2(n_2) \tag{7}$$

Proof: See Appendix 1.

There is a simple interpretation to Proposition 3: The difference between the two chi-squared distributions is in itself chi-squared distributed and can be interpreted as the statistical significance of $\mathbf{X_2}$ *corrected for* $\mathbf{X_1}$. Note that this result is non-trivial as the difference between two chi-squared distribution is generally *not* chi-squared distributed.

When the Normal Approximation Fails: The Monte Carlo Method

Lahiri et al. (2007) demonstrate that the accuracy of a normal approximation to the hypergeometric distribution increases with the sample size and decreases with a larger deviation of the number of possible successes and the number of draws from 50% of the total number of samples. In addition to these factors, the accuracy of a normal approximation to the distribution proposed in this paper will depend on the prevalence of outliers.

For cases in which the approximation error is judged to be unacceptably large, we propose an alternative procedure: We reinterpret the test statistics defined in (3) and (6) as the squared Mahalanobis distance of the observation from the expected value under the null hypothesis. Essentially, a chi-squared test is designed to calculate the probability that a random procedure would produce an observation with a Mahalanobis distance that is greater or equal a given value.

Where the normal approximation and thus the chi-squared test fails, we construct a new test based on a Monte Carlo experiment: For this experiment, we rearrange values y^j_c randomly, such that $y^{j\_new}_c = y^{j\_old}_c$ and $y^{j\_new}_d = y^{j\_old}_d$ for all c, d. We now know that the null hypothesis (H_0) is true and we calculate the test statistic as defined in (3) or (6), depending on what we want to test. We repeat this procedure a large number of times and record the test statistics generated.

We then compare the randomly generated test statistics to the test statistic based on the original values. Our new p-value will be the share of randomly generated test statistics that exceed the test statistic based on the original values.

Note that $E(X^c_a)$ and $\mathbf{V}$ are unaffected by the random rearrangement of values y^j_c. We only need to recalculate X^c_a as defined in (1). Since this is computationally very cheap, the Monte Carlo method is computationally feasible even for large datasets and a large number of iterations.

Note further that the proposed methods of calculating $E(X^c_a)$ and $\mathbf{V}$ are accurate regardless of whether a normal distribution is assumed.

Using the Concept to Address the Three Questions

Having introduced the mathematical concept underlying the statistical framework, we are still left with the question how the concept can be used to address the three questions introduced in the first section of this paper.

The first question can be examined using equations (1), (2), (3) and (5). Equation (3) defines the test statistic itself. Its input variables $\mathbf{X}$, $E(\mathbf{X})$ and $\mathbf{V}$ are defined in equations (1), (2) and (5) respectively. When the normal approximation is judged to be appropriate, the test statistics can be interpreted as chi-squared distributed under the null hypothesis. Otherwise, it can be compared with the results of the Monte Carlo experiment.

The second question can be examined by slightly modifying the approach defined above. For a pairwise comparison of two predictive methods, their out-of-sample predictive accuracy should first be measured using an appropriate loss function, such as the squared prediction error. We then create a dataset of size 2n consisting of the squared prediction errors of the predictive methods we want to compare. These are our values for x^i_a as defined in equation (1). y^i_c is then a dummy variable that assumes the value of 1 if x^i_a has been generated by the new method and 0 otherwise.

The null hypothesis we examine in the second question is that there is no statistically significant difference between the predictive errors generated by the two methods. We can therefore apply equations (1), (2), (3) and (5) in a similar fashion as

above.

The third question can be examined using equations (1), (2), (3), (5) and (7). We first calculate the test statistic as defined in equation (3) for the combined predictions of the new predictive model(s) and the predictive model(s) we would like to correct for. We then do the same only for the predictive model(s) we would like to correct for. By equation (7), the difference between the two test statistics is in itself chi-squared distributed under the null hypothesis, if the normal approximation is judged to be appropriate. Otherwise, the Monte Carlo approach can be employed.

Evaluation

We evaluated the usefulness of our framework by comparing it to a more traditional approach of assessing predictive accuracy. As an illustration, we used the *digits* dataset, which is a public domain dataset commonly used in the computer science literature for evaluating machine learning algorithms. The purpose of the dataset is to train algorithms to recognise handwritten digits. The dataset contains a total of 1797 samples (about 180 for each digit) and 64 features. We selected 13 machine learning algorithms and used the machine learning library *scikit-learn*, which was developed by Pedregosa et al. (2011) for Python. All of these algorithms are standard predictive methods widely accepted and used in the recent predictive analytics literature. We created out-of-sample predictions using stratified 10-fold cross-validation.

We calculated the RMSE based on the probabilistic out-of-sample predictions generated by each of the 13 algorithms. We then sorted the algorithms in order of their predictive performance and used Welch's t-test to test whether the difference of the squared prediction errors for each of these algorithms in comparison to the next algorithm is statistically significant. Results are reported in Table 2.

Table 2. Results for the digits dataset (t-test)

	RMSE	**t**	**p-value**
Support vector machine (polynomial kernel)	0.07317	-0.03047	0.97570
Support vector machine (radial basis function kernel)	0.07330	-4.71988	2.3711e-06***
AdaBoost	0.09585	-0.74261	0.45772
Gradient boosting	0.09947	-2.92146	0.00349***
Stochastic gradient descent	0.11299	-1.24922	0.21159
Logistic regression	0.11875	-0.33890	0.73469
Support vector machine (linear kernel)	0.12030	-0.32270	0.74692
Probit regression	0.12179	-1.94641	0.05162*
K-nearest neighbours	0.12939	-0.58249	0.56024
Linear discriminant analysis	0.13172	-5.40331	6.5924e-08***
Random forest	0.15304	-5.20449	1.9598e-07***
Decision tree	0.17360	-69.59584	0.00000***
Naive Bayes	0.53725	-	-

Note: *p < 0.1, **p < 0.05, ***p < 0.01

The results indicate that among the algorithms tested a support vector machine with either a polynomial kernel or a radial basis function kernel are the best algorithms for the digits dataset. Differences in performance between the two algorithms appear to be relatively small and it does seem to be important whether we use a polynomial kernel or a radial basis function kernel.

However, these results can only give us an indication of the relative performance of each of our algorithms: From our

results, we know that the outperformance of some algorithms in comparison to others is statistically highly significant, but we cannot infer statistical significance of the predictions themselves. For instance, Naive Bayes might be such a weak predictor that its out-of-sample predictions are negatively correlated with the actual class labels. If that is the case, an algorithm that does not outperform a random walk might still outperform Naive Bayes. Alternatively, all of these predictive methods might be highly informative, with some being more informative than others. This traditional approach cannot give us any indication which of these possible interpretations is true.

We then contrast these results with the insight gained from our own framework: As above, we sorted the algorithms by their RMSE and investigated the statistical significance in comparison to the next algorithm using the approach we developed for the second question (accuracy of predictive method in comparison to alternative method). For the third question (accuracy of predictive method corrected for alternative methods), we used all remaining algorithms as control variables. We also conducted a Monte Carlo simulation with 100,000 iterations for each of the algorithms and each of the three questions. Results are reported in Table 3.

Table 3. Results for the digits dataset

	Accuracy of predictive method			**Accuracy of method in comparison to next method**			**Accuracy of method corrected for other methods**	
	$X^c_a - E(X^c_a)$	χ2 (p-value)	p-value from Monte Carlo simulation	RMSE	χ2 (p-value)	p-value from Monte Carlo simulation	χ2 (p-value)	p-value from Monte Carlo simulation
SVM (polynomial kernel)	1353.12	13875 (0.00000***)	0.00000***	0.07317	0.000756965 (0.978051)	0.97739	4.5544 (0.03283**)	0.03406**
SVM (radial basis function kernel)	1353.35	13871 (0.00000***)	0.00000***	0.07330	20.2292 (6.87e-06***)	0.00000***	0.280358 (0.596467)	0.59219
AdaBoost	1305.66	13317.3 (0.00000***)	0.00000***	0.09585	0.531503 (0.465976)	0.46664	8.34505 (0.00387***)	0.00369***
Gradient boosting	1288.26	13193.5 (0.00000***)	0.00000***	0.09947	7.94789 (0.00481***)	0.00465***	1.34062 (0.246924)	0.24910
Stochastic gradient descent	1276.13	12718.7 (0.00000***)	0.00000***	0.11299	1.40713 (0.2355)	0.23399	2.53227 (0.11154)	0.11085
Logistic regression	1288.49	12531.4 (0.00000***)	0.00000***	0.11875	0.105197 (0.745679)	0.74523	0.610687 (0.434529)	0.43296
SVM (linear kernel)	1230.67	12444.8 (0.00000***)	0.00000***	0.12030	0.095851 (0.756867)	0.75844	5.02729 (0.02495**)	0.02456
Probit regression	1288.27	12438 (0.00000***)	0.00000***	0.12179	3.35216 (0.0671168*)	0.06729*	2.20726 (0.137362)	0.13607
K-nearest neighbours	1048.23	12140.1 (0.00000***)	0.00000***	0.12939	0.281373 (0.595803)	0.59749	2.97042 (0.0848*)	0.08510*
Linear discriminant analysis	1254.65	12060.4 (0.00000***)	0.00000***	0.13172	27.892 (1.28e-07***)	0.00000***	1.29648 (0.254857)	0.25317
Random forest	842.26	11835.8 (0.00000***)	0.00000***	0.15304	28.7859 (8.08e-08***)	0.00000***	1.98829 (0.15852)	0.15799
Decision tree	1057.87	10105.6 (0.00000***)	0.00000***	0.17360	1837.69 (0.00000***)	0.00000***	0.0185835 (0.891567)	0.89055
Naive Bayes	944.77	3130.61 (0.00000***)	0.00000***	0.53725	-	-	2.62881 (0.10494)	0.10000

Note: SVM=support vector machine, *p < 0.1, **p < 0.05, ***p < 0.01

We find that the relative performance indicated by the chi-squared-values is identical to the relative performance indicated by RMSE. Even though we would expect them to be similar, there is no mathematical reason to assume that this should always be the case.

We also find that the p-values from the chi-squared-distribution are very close to the p-values from the Monte Carlo simulations. This indicates that the normal approximation appears to be fairly accurate for this dataset.

The results demonstrate that the out-of-sample predictions generated by all of these algorithms are positively correlated with the actual class labels (as indicated by the positive values for $X^c_a - E(X^c_a)$ in the second column) and statistically highly significant. Therefore every single algorithm, including Naive Bayes, is a highly effective predictor in this problem domain.

Results also demonstrate that it *does* matter whether we use a polynomial kernel or a radial basis function kernel for a support vector machine: Even though the difference in overall performance may be insignificant, a support vector machine with a polynomial kernel is considerably less correlated with the other algorithms we investigated and is therefore able to better compensate their weaknesses. A polynomial kernel is therefore highly preferable to a radial basis function in this setting.

Finally, we are able to show that AdaBoost, even though not being the best performer among the algorithms we evaluated, still remains statistically significant even when corrected for all other algorithms. Whereas traditional approaches to evaluating predictive methods would have led us to conclude that AdaBoost is not an interesting contribution to the field, our own statistical framework for evaluating predictive methods demonstrates that it may be able to effectively compensate weaknesses of other algorithms.

Discussion

The results of our evaluation demonstrate the usefulness of the framework we developed: Whereas traditional approaches rely mainly on the evaluation of relative performance, we were able to demonstrate that the additional information generated by our own statistical framework is important: The distinction whether we are comparing mediocre predictors to bad predictors or highly effective predictors to slightly less effective predictors is an important one. In addition, to the best of our knowledge, our statistical framework is the only framework able to evaluate the predictive performance of algorithms corrected for the predictions generated by other algorithms. The results of our evaluation demonstrate that this performance measure cannot be inferred from either absolute or relative performance and is interesting in its own right.

Our framework might benefit researchers and practitioners alike: Being able to measure the intercorrelation of out-of-sample predictions, researchers could use this framework to build more effective ensembles of machine learning algorithms. Practitioners, on the other hand, often rely on several decision support systems or predictive algorithms to guide important decisions. However, such an approach is only useful if these different systems are sufficiently independent from each other. Since our framework is able to assess the statistical significance of predictive methods when corrected for other methods, it could help them choose an appropriate set of decision support systems.

Conclusion

The objective of this paper was to define three important questions to consider when evaluating predictive models and to develop a unified statistical framework to examine these questions. Our framework is applicable to a wide variety of problems, including regression, classification and multi-label classification. Predictions can be both continuous or discrete. It can be used to test individual predictive methods or several predictive methods at once. It is even applicable when the assumption of a normal distribution fails. Owing to its versatility and its practical and theoretical advantages over current statistical hypothesis tests used to evaluate predictive models, the framework could be of interest to most researchers in the field of predictive analytics or big data.

Researchers could make use of the theoretical benefit of the framework, namely that it provides a "fail-safe" option for cases in which the assumption of a normal distribution is not applicable. This issue has been largely ignored in previous literature.

Even though the framework is complete and can be readily applied to examine predictive methods, there is still potential for further research: Most importantly, there is still a need to define the exact circumstances under which the normal approximation is appropriate. Even though the normal approximation is not necessary for the framework to be applicable, it significantly reduces computational cost, particularly for large datasets. Finding these exact circumstances can only be done numerically.

Most importantly, we hope that this paper can be seen as an encouragement to use statistical hypothesis testing to evaluate predictive models thus enhancing the rigour of the predictive analytics literature.

References

Abbasi, A., Albrecht, C., Vance, A., and Hansen, J. 2012. "MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud.," *MIS Quarterly* (36:4), pp. 1293–A12.

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker, J., Jay F. 2010. "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly* (34:3), pp. 435–461.

Bai, X. 2011. "Predicting consumer sentiments from online text," *Decision Support Systems* (50:4), pp. 732–742.

Bao, H., Li, Q., Liao, S. S., Song, S., and Gao, H. 2013. "A new temporal and social PMF-based method to predict users' interests in micro-blogging," *Decision Support Systems* (55:3), pp. 698–709.

Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. 2011. "Data mining for credit card fraud: A comparative study," *Decision Support Systems* (50:3), pp. 602–613.

Cao, Q., Ewing, B. T., and Thompson, M. A. 2012. "Forecasting medical cost inflation rates: A model comparison approach," *Decision Support Systems* (53:1), pp. 154–160.

Carbonneau, R. A., Kersten, G. E., and Vahidov, R. M. 2011. "Pairwise issue modeling for negotiation counteroffer prediction using neural networks," *Decision Support Systems* (50:2), pp. 449–459.

Chan, S. W. K., and Franklin, J. 2011. "A text-based decision support system for financial sequence prediction," *Decision Support Systems* (52:1), pp. 189–198.

Choi, T.-M., Hui, C.-L., Liu, N., Ng, S.-F., and Yu, Y. 2013. "Fast fashion sales forecasting with limited data and time," *Decision Support Systems* , pp. 84–92.

Choi, T.-M., Yu, Y., and Au, K.-F. 2011. "A hybrid SARIMA wavelet transform method for sales forecasting," *Decision Support Systems* (51:1), pp. 130–140.

Cui, G., Wong, M. L., and Wan, X. 2012. "Cost-Sensitive Learning via Priority Sampling to Improve the Return on Marketing and CRM Investment," *Journal of Management Information Systems* (29:1), pp. 341–374.

David, M., Perkovič, M., Suban, V., and Gollasch, S. 2012. "A generic ballast water discharge assessment model as a decision supporting tool in ballast water management," *Decision Support Systems* (53:1), pp. 175–185.

Delen, D. 2010. "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems* (49:4), pp. 498–506.

Delen, D., Zaim, H., Kuzey, C., and Zaim, S. 2013. "A comparative analysis of machine learning systems for measuring the impact of knowledge management practices," *Decision Support Systems* (54:2), pp. 1150–1160.

Diebold, F. X., and Mariano, R. S. 2002. "Comparing predictive accuracy," *Journal of Business & economic statistics* (20:1), pp. 134–144.

Gerber, M. S. 2014. "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems* (available at http://www.sciencedirect.com/science/article/pii/S0167923614000268).

Hagenau, M., Liebmann, M., and Neumann, D. 2013. "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decision Support Systems* (55:3), pp. 685–697.

Du Jardin, P., and Séverin, E. 2011. "Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model," *Decision Support Systems* (51:3), pp. 701–711.

Kao, L.-J., Chiu, C.-C., Lu, C.-J., and Chang, C.-H. 2013. "A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting," *Decision Support Systems* (54:3), pp. 1228–1244.

Khansa, L., and Liginlal, D. 2011. "Predicting stock market returns from malicious attacks: A comparative analysis of vector autoregression and time-delayed neural networks," *Decision Support Systems* (51:4), pp. 745–759.

Kim, H.-N., El-Saddik, A., and Jo, G.-S. 2011. "Collaborative error-reflected models for cold-start recommender systems," *Decision Support Systems* (51:3), pp. 519–531.

Kisilevich, S., Keim, D., and Rokach, L. 2013. "A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers' context," *Decision Support Systems* (54:2), pp. 1119–1133.

Lahiri, S. N., Chatterjee, A., and Maiti, T. 2007. "Normal approximation to the hypergeometric distribution in nonstandard cases and a sub-Gaussian Berry–Esseen theorem," *Journal of Statistical Planning and Inference* (137:11), pp. 3570–3590.

Lau, H. C. W., Ho, G. T. S., and Zhao, Y. 2013. "A demand forecast model using a combination of surrogate data analysis and optimal neural network approach," *Decision Support Systems* (54:3), pp. 1404–1416.

Lee, H., Lee, Y., Cho, H., Im, K., and Kim, Y. S. 2011. "Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model," *Decision Support Systems* (52:1), pp. 207–216.

Lee, Y.-H., Wei, C.-P., Cheng, T.-H., and Yang, C.-T. 2012. "Nearest-neighbor-based approach to time-series classification," *Decision Support Systems* (53:1), pp. 207–217.

Li, D.-C., Chang, C.-C., and Liu, C.-W. 2012. "Using structure-based data transformation method to improve prediction accuracies for small data sets," *Decision Support Systems* (52:3), pp. 748–756.

Li, N., and Wu, D. D. 2010. "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems* (48:2), pp. 354–368.

Li, Q., Wang, T., Gong, Q., Chen, Y., Lin, Z., and Song, S. 2014. "Media-aware quantitative trading based on public Web information," *Decision Support Systems* (61), pp. 93–105.

Li, X., and Chen, H. 2013. "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decision Support Systems* (54:2), pp. 880–890.

Lu, C.-J., Lee, T.-S., and Lian, C.-M. 2012. "Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks," *Decision Support Systems* (54:1), pp. 584–596.

Oh, C., and Sheng, O. 2011. "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement.," in *ICIS*, (available at http://www.misrc.csom.umn.edu/workshops/2011/fall/OliviaSheng_Paper.pdf).

Olson, D. L., and Chae, B. 2012. "Direct marketing decision support through predictive customer response modeling," *Decision Support Systems* (54:1), pp. 443–451.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. 2011. "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research* (12), pp. 2825–2830.

Vd Reijden, P., and Koppius, O. R. 2010. "The Value of Online Product Buzz in Sales Forecasting.," in *ICIS*, ICIS 2010 Proceedings, pp. 1–17 (available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.934&rep=rep1&type=pdf)71.

Sahoo, N., Singh, P. V., and Mukhopadhyay, T. 2012. "A Hidden Markov Model for Collaborative Filtering.," *MIS Quarterly* (36:4), pp. 1329–1356.

Schumaker, R. P. 2013. "Machine learning the harness track: Crowdsourcing and varying race history," *Decision Support Systems* (54:3), pp. 1370–1379.

Sermpinis, G., Dunis, C., Laws, J., and Stasinakis, C. 2012. "Forecasting and trading the EUR/USD exchange rate with stochastic Neural Network combination and time-varying leverage," *Decision Support Systems* (54:1), pp. 316–329.

Serrano-Cinca, C., and Gutiérrez-Nieto, B. 2013. "Partial Least Square Discriminant Analysis for bankruptcy prediction," *Decision Support Systems* (54:3), pp. 1245–1255.

Shin, H., Hou, T., Park, K., Park, C.-K., and Choi, S. 2013. "Prediction of movement direction in crude oil prices based on semi-supervised learning," *Decision Support Systems* (55:1), pp. 348–358.

Shmueli, G. 2010. "To Explain or to Predict?," *Statistical Science* (25:3), pp. 289–310.

Shmueli, G., and Koppius, O. R. 2011. "Predictive Analytics in Information Systems Research.," *MIS Quarterly* (35:3), pp. 553–572.

Su, P., Mao, W., Zeng, D., and Zhao, H. 2012. "Mining actionable behavioral rules," *Decision Support Systems* (54:1), pp. 142–152.

Watson, H. J., and Wixom, B. H. 2007. "The Current State of Business Intelligence," *Computer* (40:9), pp. 96–99.

Yang, C.-S., Wei, C.-P., Yuan, C.-C., and Schoung, J.-Y. 2010. "Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages," *Decision Support Systems* (50:1), pp. 325–335.

Yen, S. M.-F., and Hsu, Y.-L. 2010. "Profitability of technical analysis in financial and commodity futures markets — A reality check," *Decision Support Systems* (50:1), pp. 128–139.

Yolcu, U., Egrioglu, E., and Aladag, C. H. 2013. "A new linear & nonlinear artificial neural network model for time series forecasting," *Decision Support Systems* (54:3), pp. 1340–1347.

Zhao, H., Sinha, A. P., and Bansal, G. 2011. "An extended tuning method for cost-sensitive regression and forecasting," *Decision Support Systems* (51:3), pp. 372–383.

Appendix 1

Proposition 1: Proof of Non-Arbitrariness of the Model

Since variables can be freely rearranged, we simplify our notation without any loss of generality by letting i=n and j=n+1.

Since every variance-covariance-matrix is positive semi-definite, every positive semi-definite matrix has an inverse, which is also positive semi-definite, and every positive semi-definite matrix has a root, we know that there exists a positive semi-definite matrix $\mathbf{V^{-1/2}_n}$ that is the square root of the variance-covariance matrix. Let $\mathbf{(V^{-1/2}_n)_{xy}}$ denote the element in the xth row and the yth column of the square root matrix. Finally, define matrix **A** as follows:

$$\mathbf{A} = \begin{bmatrix} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{11}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{12}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{21}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{22}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n2}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \end{bmatrix} \tag{A1}$$

Lemma 1.1:

$$\mathbf{AX_{n+1}} = \mathbf{V_n^{-\frac{1}{2}}X_n} \tag{Lemma 1.1}$$

Proof:

$$\mathbf{AX_{n+1}} = \begin{bmatrix} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{11}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{12}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{21}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{22}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n2}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ -\sum_{a=1}^{n} X_a \end{bmatrix} \tag{A2}$$

$$= \begin{bmatrix} \sum_{a=1}^{n-1} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}} X_a - \sum_{a=1}^{n-1} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} X_a + \sum_{a=1}^{n} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} X_a \\ \sum_{a=1}^{n-1} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}} X_a - \sum_{a=1}^{n-1} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} X_a + \sum_{a=1}^{n} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} X_a \\ \cdots \\ \sum_{a=1}^{n-1} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}} X_a - \sum_{a=1}^{n-1} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} X_a + \sum_{a=1}^{n} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} X_a \end{bmatrix} = \begin{bmatrix} \sum_{a=1}^{n} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}} X_a \\ \sum_{a=1}^{n} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}} X_a \\ \cdots \\ \sum_{a=1}^{n} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}} X_a \end{bmatrix} = \mathbf{V_n^{-\frac{1}{2}}X_n}$$

q.e.d

Lemma 1.2:

$$\mathbf{A}\sigma_{\mathbf{n+1}} = \mathbf{V_n^{-\frac{1}{2}}}\sigma_{\mathbf{n}} \tag{Lemma 1.2}$$

Proof:

$$\mathbf{A}\sigma_{\mathbf{n+1}} = \begin{bmatrix} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{11}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{12}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{21}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{22}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n2}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \end{bmatrix}$$

$$\begin{bmatrix} (\sigma_{\mathbf{n}})_{\mathbf{11}} & (\sigma_{\mathbf{n}})_{\mathbf{12}} & \cdots & (\sigma_{\mathbf{n}})_{\mathbf{1m}} \\ (\sigma_{\mathbf{n}})_{\mathbf{21}} & (\sigma_{\mathbf{n}})_{\mathbf{22}} & \cdots & (\sigma_{\mathbf{n}})_{\mathbf{2m}} \\ \cdots & \cdots & \cdots & \cdots \\ -\sum_{a=1}^{n-1}(\sigma_{\mathbf{n}})_{\mathbf{a1}} & -\sum_{a=1}^{n-1}(\sigma_{\mathbf{n}})_{\mathbf{a2}} & \cdots & -\sum_{a=1}^{n-1}(\sigma_{\mathbf{n}})_{\mathbf{am}} \end{bmatrix} \tag{A3}$$

$$= \begin{bmatrix} \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}}(\sigma_{\mathbf{n}})_{\mathbf{a1}} & \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}}(\sigma_{\mathbf{n}})_{\mathbf{a2}} & \cdots & \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}}(\sigma_{\mathbf{n}})_{\mathbf{am}} \\ \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}}(\sigma_{\mathbf{n}})_{\mathbf{a1}} & \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}}(\sigma_{\mathbf{n}})_{\mathbf{a2}} & \cdots & \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}}(\sigma_{\mathbf{n}})_{\mathbf{am}} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}}(\sigma_{\mathbf{n}})_{\mathbf{a1}} & \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}}(\sigma_{\mathbf{n}})_{\mathbf{a2}} & \cdots & \sum_{a=1}^{n}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}}(\sigma_{\mathbf{n}})_{\mathbf{am}} \end{bmatrix} = \mathbf{V_n^{-\frac{1}{2}}}\sigma_{\mathbf{n}}$$

q.e.d.

Lemma 1.3:

$$\mathbf{A'A} = \mathbf{V_{n+1}^{-1}} \tag{Lemma 1.3}$$

Proof:

$$\mathbf{V_{n+1}A'A} = \begin{bmatrix} (\mathbf{V_n})_{\mathbf{11}} & (\mathbf{V_n})_{\mathbf{12}} & \cdots & -\sum_{\mathbf{a=1}}^{\mathbf{n}}(\mathbf{V_n})_{\mathbf{1a}} \\ (\mathbf{V_n})_{\mathbf{21}} & (\mathbf{V_n})_{\mathbf{22}} & \cdots & -\sum_{\mathbf{a=1}}^{\mathbf{n}}(\mathbf{V_n})_{\mathbf{2a}} \\ \cdots & \cdots & \cdots & \cdots \\ -\sum_{\mathbf{a=1}}^{\mathbf{n}}(\mathbf{V_n})_{\mathbf{a1}} & -\sum_{\mathbf{a=1}}^{\mathbf{n}}(\mathbf{V_n})_{\mathbf{a2}} & \cdots & \sum_{\mathbf{a=1}}^{\mathbf{n}}\sum_{\mathbf{b=1}}^{\mathbf{n}}(\mathbf{V_n})_{\mathbf{ab}} \end{bmatrix}$$

$$\begin{bmatrix} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{11}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{21}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{12}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{22}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n2}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \\ \cdots & \cdots & \cdots & \cdots \\ -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \end{bmatrix}$$

$$\begin{bmatrix} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{11}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{12}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{21}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{22}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n2}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{a=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}} & \sum_{a=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}} & \cdots & \sum_{a=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}} \\ \sum_{a=1}^{n}(\mathbf{V_n})_{\mathbf{2a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}} & \sum_{a=1}^{n}(\mathbf{V_n})_{\mathbf{2a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}} & \cdots & \sum_{a=1}^{n}(\mathbf{V_n})_{\mathbf{2a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}} \\ \cdots & \cdots & \cdots & \cdots \\ -\sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1a}} & -\sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2a}} & \cdots & -\sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{na}} \end{bmatrix}$$

$$\begin{bmatrix} (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{11}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{12}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{1n}} \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{21}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{22}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{n1}} - (\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} & \cdots & -(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{nn}} \end{bmatrix} \tag{A4}$$

$$= \begin{bmatrix} \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{b1}} & \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{b2}} & \cdots & 0 \\ \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{2a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{b1}} & \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{2a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{b2}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ -\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ca}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{c1}} & -\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ca}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{c1}} & \cdots & 0 \end{bmatrix}$$

$$-\begin{bmatrix} \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{bn}} & \cdots & \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{bn}} \\ \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{bn}} & \cdots & \sum_{a=1}^{n}\sum_{b=1}^{n}(\mathbf{V_n})_{\mathbf{1a}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{bn}} \\ \cdots & \cdots & \cdots \\ -\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ca}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{cn}} & \cdots & -\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n}(\mathbf{V_n})_{\mathbf{ba}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{ca}}(\mathbf{V_n^{-\frac{1}{2}}})_{\mathbf{cn}} \end{bmatrix}$$

Since the square root of a symmetrical, positive semi-definite matrix is also symmetrical:

$$= \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ -1 & -1 & \ldots & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ -1 & -1 & \ldots & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = \mathbf{I} \tag{A5}$$

$$\Rightarrow \mathbf{A'A} = \mathbf{V_{n+1}^{-1}}$$

q.e.d.

Define matrix **V** as follows:

$$\mathbf{V} = \mathbf{V_Y} - \sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}}\sigma_{\mathbf{n+1}} \tag{A6}$$

Then, by block-wise inversion of a matrix:

$$\begin{bmatrix} \mathbf{X_{n+1}'} & \mathbf{Y'} \end{bmatrix} \begin{bmatrix} \mathbf{V_{n+1}} & \sigma_{\mathbf{n+1}} \\ \sigma_{\mathbf{n+1}}' & \mathbf{V_Y} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_{n+1}} \\ \mathbf{Y} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X_{n+1}'} & \mathbf{Y'} \end{bmatrix} \begin{bmatrix} \mathbf{V_{n+1}^{-1}} + \mathbf{V_{n+1}^{-1}}\sigma_{\mathbf{n+1}}\mathbf{V^{-1}}\sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}} & -\mathbf{V_{n+1}^{-1}}\sigma_{\mathbf{n+1}}\mathbf{V^{-1}} \\ -\mathbf{V^{-1}}\sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}} & \mathbf{V^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{X_{n+1}} \\ \mathbf{Y} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X_{n+1}'V_{n+1}^{-1}} + \mathbf{X_{n+1}'V_{n+1}^{-1}}\sigma_{\mathbf{n+1}}\mathbf{V^{-1}}\sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}} - \mathbf{Y'V^{-1}}\sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}} & -\mathbf{X_{n+1}'V_{n+1}^{-1}}\sigma_{\mathbf{n+1}}\mathbf{V^{-1}} + \mathbf{Y'V^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{X_{n+1}} \\ \mathbf{Y} \end{bmatrix}$$

$$\begin{aligned} &= \mathbf{X_{n+1}'V_{n+1}^{-1}X_{n+1}} + \mathbf{X_{n+1}'V_{n+1}^{-1}}\sigma_{\mathbf{n+1}}\mathbf{V^{-1}}\sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}X_{n+1}} \\ &\mathbf{-Y'V^{-1}}\sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}X_{n+1}} \quad \mathbf{-X_{n+1}'V_{n+1}^{-1}}\sigma_{\mathbf{n+1}}\mathbf{V^{-1}Y} + \mathbf{Y'V^{-1}Y} \\ &= \mathbf{X_{n+1}'V_{n+1}^{-1}X_{n+1}} + (\mathbf{Y} - \sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}X_{n+1}})'\mathbf{V^{-1}}(\mathbf{Y} - \sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}X_{n+1}}) \end{aligned} \tag{A7}$$

Recall that the the square root of a symmetrical, positive semi-definite matrix is also symmetrical. By Lemmas 1.1, 1.2 and 1.3:

$$\mathbf{V} = \mathbf{V_Y} - \sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}}\sigma_{\mathbf{n+1}} = \mathbf{V_Y} - \sigma_{\mathbf{n+1}}'\mathbf{A'A}\sigma_{\mathbf{n+1}} = \mathbf{V_Y} - \sigma_{\mathbf{n}}'\mathbf{V_n^{-1}}\sigma_{\mathbf{n}} \tag{A8}$$

By Lemmas 1.1, 1.2 and 1.3:

$$\begin{aligned} &\Rightarrow \mathbf{X_{n+1}'V_{n+1}^{-1}X_{n+1}} + (\mathbf{Y} - \sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}X_{n+1}})'\mathbf{V^{-1}}(\mathbf{Y} - \sigma_{\mathbf{n+1}}'\mathbf{V_{n+1}^{-1}X_{n+1}}) \\ &= \mathbf{X_{n+1}'A'AX_{n+1}} + (\mathbf{Y} - \sigma_{\mathbf{n+1}}'\mathbf{A'AX_{n+1}})'\mathbf{V^{-1}}(\mathbf{Y} - \sigma_{\mathbf{n+1}}'\mathbf{A'AX_{n+1}}) \\ &= \mathbf{X_n'V_n^{-1}X_n} + (\mathbf{Y} - \sigma_{\mathbf{n}}'\mathbf{V_n^{-1}X_n})'\mathbf{V^{-1}}(\mathbf{Y} - \sigma_{\mathbf{n}}'\mathbf{V_n^{-1}X_n}) = \begin{bmatrix} \mathbf{X_n'} & \mathbf{Y'} \end{bmatrix} \begin{bmatrix} \mathbf{V_n} & \sigma_{\mathbf{n}} \\ \sigma_{\mathbf{n}}' & \mathbf{V_Y} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_n} \\ \mathbf{Y} \end{bmatrix} \end{aligned} \tag{A9}$$

q.e.d.

Proposition 2: Calculating the Elements of V

Let $E(x^i_a * y^j_c * x^k_b * y^l_d)$ denote the expected value of $x^i_a * y^j_c * x^k_b * y^l_d$ if i, j, k and l are random variables.

Lemma 2.1:

$$E(x_a^i y_c^j x_b^k y_d^l \mid k \neq i, l \neq j) = \frac{(\sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k - \sum_{i=1}^N x_a^i x_b^i)(\sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l - \sum_{j=1}^N y_c^j y_d^j)}{N^2(N-1)^2} \quad \text{(Lemma 2.1)}$$

Proof:

$$E(x_a^i y_c^j x_b^k y_d^l \mid k \neq i, l \neq j) = \frac{\sum_{i=1}^N x_a^i \sum_{k=1,k\neq i}^N x_b^k}{N(N-1)} \frac{\sum_{j=1}^N y_c^j \sum_{l=1,l\neq j}^N y_d^l}{N(N-1)}$$

$$= \frac{(\sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k - \sum_{i=1}^N x_a^i x_b^i)(\sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l - \sum_{j=1}^N y_c^j y_d^j)}{N^2(N-1)^2} \quad \text{(A10)}$$

q.e.d.

$$cov(X_c^a, X_d^b) = E(X_c^a X_d^b) - E(X_c^a)E(X_d^b)$$

$$= N * E(x_a^i y_c^j x_b^k y_d^l) + N(N-1) * E(x_a^i y_c^j x_b^k y_d^l \mid k \neq i, l \neq j) - E(X_c^a)E(X_d^b)$$

$$= N * \frac{\sum_{i=1}^N x_a^i x_b^i \sum_{j=1}^N y_c^j y_d^j}{N^2} + N(N-1) * \frac{(\sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k - \sum_{i=1}^N x_a^i x_b^i)(\sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l - \sum_{j=1}^N y_c^j y_d^j)}{N^2(N-1)^2}$$

$$-\frac{\sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k \sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l}{N^2} = \frac{N(N-1)\sum_{i=1}^N x_a^i x_b^i \sum_{j=1}^N y_c^j y_d^j}{N^2(N-1)}$$

$$+\frac{N(\sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k - \sum_{i=1}^N x_a^i x_b^i)(\sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l - \sum_{j=1}^N y_c^j y_d^j)}{N^2(N-1)}$$

$$-\frac{(N-1)\sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k \sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l}{N^2(N-1)} = \frac{N^2 \sum_{i=1}^N x_a^i x_b^i \sum_{j=1}^N y_c^j y_d^j - N \sum_{i=1}^N x_a^i x_b^i \sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l}{N^2(N-1)} \quad \text{(A11)}$$

$$+\frac{-N \sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k \sum_{j=1}^N y_c^j y_d^j + \sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k \sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l}{N^2(N-1)}$$

$$= \frac{(N\sum_{i=1}^N x_a^i x_b^i - \sum_{i=1}^N x_a^i \sum_{k=1}^N x_b^k)(N\sum_{j=1}^N y_c^j y_d^j - \sum_{j=1}^N y_c^j \sum_{l=1}^N y_d^l)}{N^2(N-1)}$$

q.e.d.

Proposition 3: Evaluating Statistical Significance Given Other Predictive Methods

Define matrix **V** as follows:

$$\mathbf{V} = \mathbf{V_2} - \sigma_{\mathbf{12}}' \mathbf{V_1^{-1}} \sigma_{\mathbf{12}} \quad \text{(A12)}$$

In analogy to (A7) we write:

$$\begin{bmatrix} \mathbf{X_1'} & \mathbf{X_2'} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \sigma_{\mathbf{12}} \\ \sigma_{\mathbf{12}}' & \mathbf{V_2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} = \mathbf{X_1'} \mathbf{V_1^{-1}} \mathbf{X_1} + (\mathbf{X_2} - \sigma_{\mathbf{12}}' \mathbf{V_1^{-1}} \mathbf{X_1})' \mathbf{V^{-1}} (\mathbf{X_2} - \sigma_{\mathbf{12}}' \mathbf{V_1^{-1}} \mathbf{X_1}) \quad \text{(A13)}$$

Define **Y** as follows:

$$\mathbf{Y} = \mathbf{X_2} - \sigma_{\mathbf{12}}' \mathbf{V_1^{-1}} \mathbf{X_1} \quad \text{(A14)}$$

It is easy to see that E(**Y**) is **0**. Consider its variance-covariance-matrix:

$$E(\mathbf{YY}') = E((\mathbf{X_2} - \sigma'_{\mathbf{12}}\mathbf{V_1^{-1}X_1})(\mathbf{X_2} - \sigma'_{\mathbf{12}}\mathbf{V_1^{-1}X_1})')$$

$$= E(\mathbf{X_2X_2'}) - E(\sigma'_{\mathbf{12}}\mathbf{V_1^{-1}X_1X_2'}) - E(\mathbf{X_2X_1}'\mathbf{V_1^{-1}}\sigma_{\mathbf{12}}) + E(\sigma'_{\mathbf{12}}\mathbf{V_1^{-1}X_1X_1}'\mathbf{V_1^{-1}}\sigma_{\mathbf{12}}) \quad \text{(A15)}$$

$$= \mathbf{V_2} - 2\sigma'_{\mathbf{12}}\mathbf{V_1^{-1}}\sigma_{\mathbf{12}} + \sigma'_{\mathbf{12}}\mathbf{V_1^{-1}V_1V_1^{-1}}\sigma_{\mathbf{12}} = \mathbf{V_2} - \sigma'_{\mathbf{12}}\mathbf{V_1^{-1}}\sigma_{\mathbf{12}} = \mathbf{V}$$

Therefore, $\mathbf{Y'V^{-1}Y}$ is a squared Mahalanobis distance.

$$\begin{bmatrix} \mathbf{X_1'} & \mathbf{X_2'} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \sigma_{\mathbf{12}} \\ \sigma'_{\mathbf{12}} & \mathbf{V_2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} = \mathbf{X_1'V_1^{-1}X_1} + \mathbf{Y'V^{-1}Y} \quad \text{(A16)}$$

$$\Leftrightarrow \mathbf{Y'V^{-1}Y} = \begin{bmatrix} \mathbf{X_1'} & \mathbf{X_2'} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \sigma_{\mathbf{12}} \\ \sigma'_{\mathbf{12}} & \mathbf{V_2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} - \mathbf{X_1'V_1^{-1}X_1}$$

If $\mathbf{X_1}$ and $\mathbf{X_2}$ are multivariate normal, then $\mathbf{Y}$ is a linear combination of multivariate normally distributed random variables and therefore in itself multivariate normal. Then:

$$\mathbf{X_1'V_1^{-1}X_1} = \begin{bmatrix} \mathbf{X_1'} & \mathbf{X_2'} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \sigma_{\mathbf{12}} \\ \sigma'_{\mathbf{12}} & \mathbf{V_2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} - \mathbf{X_1'V_1^{-1}X_1} \sim \chi^2(n_2) \quad \text{(A17)}$$

q.e.d.