September 2010

Performance Evaluation of Zero Net-Investment Strategies*

Abstract

Excess returns of zero net-investment strategies should be largely unpredictable under standard asset-pricing conditions with efficient markets. But obtaining systematically positive returns only requires a successful binary directional forecast of whether to go short or long, rather than accurate forecasts of returns themselves as judged by fit. With an emphasis on this important distinction, this paper introduces new statistical methods to directly assess the correct classification ability of alternative investment strategies based on binary evaluation tools commonly used in many fields, but rarely in economics and finance. We extend these techniques to evaluate strategies from a profit or gain-loss perspective and to provide methods for the three-way long/cash/short allocation problem. The formal tests that we derive allow us to determine a strategy's merits against a "coin-toss" null as well as comparing strategies against each other.

Keywords: correct classification frontier; receiver operating characteristic curve; area under the curve; gain-loss ratio.

JEL Codes: G14, G17, C14, C58

Òscar Jordà
Department of Economics
University of California
Davis CA 95616
e-mail (Jordà): ojorda@ucdavis.edu

Alan M. Taylor
Morgan Stanley
1585 Broadway, 38th Floor
New York NY 10036
and UC Davis and NBER
e-mail (Taylor): alan.taylor@morganstanley.com

*Taylor is has been supported by the Center for the Evolution of the Global Economy at UC Davis and Jordà by DGCYT Grant (SEJ2007-63098-econ); part of this work was completed whilst Taylor was a Houblon-Norman/George Fellow at the Bank of England and another part whilst he was a Senior Advisor at Morgan Stanley; all of this research support is gratefully acknowledged. We thank Colin Cameron and Burkhard Schipper for comments and suggestions. Travis Berge and Yanping Chong provided superb research assistance. All errors are ours.

One of the enduring problems in empirical finance is the quest to evaluate the performance of trading strategies. In many cases, such as the canonical test of the efficient markets hypothesis, this means judging performance relative to a random-walk or a coin-toss null hypothesis to decide whether a strategy has truly generated risk-adjusted predictable excess returns. At other times, when two or more potential trading strategies are being considered, the researcher may wish to evaluate whether one is better than the other in a statistically significant way.

Although they are amongst the most basic trading strategies available, simple directional or long-short strategies represent an important and widely-used class of investment rules. In these strategies the goal is not to craft a carefully weighted portfolio, but merely to make a binary choice as to whether to go long or short in one or more securities in each trading period. Pick the right direction and upside is guaranteed; pick the wrong direction and losses ensue.

These observations prompt the question whether useful directional forecasts are truly possible. That is, if some such models were proposed, how would one evaluate their claims to superiority against the coin-toss null, or even against each other? This paper provides new tools to answer these questions based on the receiver operating characteristic (ROC) curve: a collection of nonparametric statistical methods designed to evaluate classification ability in binary-outcome decision problems. ROC tools originated in the field of signal detection theory (Peterson and Birdsall 1953). They are currently broadly applied to evaluate diagnostic tests from different biomarkers in medicine as well as ranking radiological readings (see Pepe 2003 for an extensive monograph). They are also widely used in a variety of other fields of science, such as psychometrics (see Swets and Pickett 1982 for a classical treatment), machine learning (see Spackman 1989 for an early reference), and atmospheric sciences, where they have become part of the World Meteorological Organization's (WMO) Standard Verification System for assessing the quality of weather forecasts (see Stanski, Wilson, and Burrows 1989; and WMO 2000). With this paper we hope to popularize these procedures in economics and finance as well and, to that end, we introduce the *correct classification frontier* (CC), which summarizes the information in the ROC curve in a manner that is more natural to economists and for the type of problem we consider in this paper.

Despite their simplicity, directional trading strategies offer a well trodden playing field for statistical analysis. For a start, at a theoretical level, even if we cannot forecast returns well as judged by fit, an ability to make at least a systematic directional forecast might yield significant excess returns and would be sufficient to reject the classic, risk-neutral efficient

markets hypothesis. To take an example from forex trading strategies, which comprise one of the applications used this paper, we shall step outside the well known Meese and Rogoff (1983) puzzle concerning the unimpressive root-mean-square error (RMSE) of most exchange rate forecasts compared to the random-walk null, since this puzzle by itself does not rule out the possibility of good directional forecasts that generate statistically and quantitatively significant profit opportunities.[1] Directional tests in this tradition include Pesaran and Timmermann (1992) and Anatolyev and Gerko (2005), which we discuss in more detail below.

If we are to move beyond evaluation based on fit, as with the RMSE loss function, then what criteria can we use instead? When the predictive model is a correct representation of the data generating process, one obtains unbiased estimates of the parameters of the true model under any proper loss function. However, when the statistical model is only an approximation, different loss functions result in different models and parameter estimates, and therefore possibly different conclusions about the usefulness of a particular model (see Hand and Vinciotti 2003). The methods that we propose here recognize that the decisions on whether to go long or short on an investment vary with an investor's preferences and attitudes toward risk and for this reason, we aim to design methods where comparisons can be made, as much as possible, robustly with respect to the unknown loss function.

The tools that we introduce here are simple, often nonparametric (which is important because financial data tend to be distributed with heavy tails), and have well-understood large-sample properties that facilitate construction of classical inferential procedures. In fact, some of these procedures are closely related to the theory of rank tests (see, e.g. Hájek, Šidák, and Sen 1999) in that they effectively measure the distance between two distributions, in our case, the distance between the distributions of the forecast signal when returns are positive versus negative. In this paper we discuss some of these tools and provide appropriate results to evaluate binary prediction outcomes in the more realistic case of directional strategies with variable payoffs. We also investigate more complex multi-categorical investment strategies, such as long/short/cash investment positions, with an extension of the CC frontier to multiple dimensions.

Our CC frontier analysis is advantageous in that it is robust to the loss function (or preferences) of individual investors. Traditional statistics for the evaluation of binary deci-

[1] Cheung et al. (2005) have revisited the Meese-Rogoff puzzle and surveyed the entire gamut of exchange rate models; although RMSE performance was still lackluster, they did find that directional forecasts did rather better at outperforming the random walk.

sion problems (such as log and quadratic probability scores, Brier scores, misclassification probabilities and other commonly reported statistics) are only appropriate as long as the implied loss function coincides with the investor's loss function. On the other hand, while summary statistics associated with the CC frontier provide general statements on stochastic dominance, the CC frontier summarizes the space of all possible trade-offs implied by a particular set of preferences over an investment strategy. Thus, a strategy that is more successful in predicting correctly long-short positions may nevertheless be particularly vulnerable to extreme events, whereas a less successful strategy may still produce positive returns but with better protection against catastrophic losses. Analysis using CC frontiers allows one to visualize the regions in which these trade-offs take place.

We conclude the paper with two empirical applications of the methods that we propose. The first application is based on Welch and Goyal's (2008) state-of-the-art investigation of signals that helps forecast U.S. equity returns. Our aim is to examine the value of these signals in constructing profitable investment strategies where one borrows/lends at the risk-free rate to purchase/sell U.S. equities. We find out-of-sample evidence that several of these signals generate consistently profitable trades in contrast to Welch and Goyal's (2008) results, which are based on tests of fit using conventional RMSE metrics.

A second application focuses on currency carry trades in which a speculator borrows in one currency to invest in another, thus arbitraging the interest rate differential while bearing the risk of a possibly adverse exchange rate movement. Berge, Jordà and Taylor (2010) discuss four basic carry trade strategies where an investor's only choices are which currency to go short and which to go long. In practice though, transactions costs and other considerations may make some of the trades unprofitable when predicted returns before costs are small. Thus, we examine long/cash/short strategies using the same four carry trade investments described in Berge, Jordà and Taylor (2010) and find that these more sophisticated methods rank the preferred strategies somewhat differently compared to the case where only binary long/short strategies were considered.

1 Motivation: Accuracy Versus Arbitrage

The problem of evaluating the risk-adjusted excess returns of an investment can be cast as a zero net-investment strategy with respect to the risk-free rate. Fundamental models of consumption-based asset pricing in frictionless environments with rational agents would then suggest that, if m_{t+1} denotes the stochastic discount factor and x_{t+1} denotes ex-post

excess returns (see e.g. Cochrane 2001), then

$$E_t(m_{t+1}x_{t+1}) = 0. \tag{1}$$

An example of x_{t+1} is a currency carry trade position in which case $x_{t+1} = \Delta e_{t+1} + (i_t^* - i_t)$ where e_{t+1} denotes the log of the (home) exchange rate and $i_t^* - i_t$ the one period interest rate differential between two countries (foreign minus home). Another example of x_{t+1} is an investment where the trader goes short/long on a risky asset by lending/borrowing at the risk free rate. In both cases, the trader will be interested in determining whether, given information at time t, he should go long or short. For the moment, we abstract from many well-known frictions, such as short-selling constraints, transactions fees, and so on.

Under such conditions, one may presume that it would be difficult to predict x_{t+1} given information available up to time t and this seems to be generally the case (for surveys of the relevant literature on beating the random walk see, e.g, on currencies, Kilian and Taylor 2003; on equities, Welch and Goyal 2008). However, notice that the problem facing the trader in practice is somewhat simpler: he does not need to predict excess returns per se, but only whether to go long or short. That is, the actual performance of the trading strategy using one-period ahead forecasts $\widehat{x}_{t+1}$ is given by the realized returns

$$\widehat{\mu}_{t+1} = \text{sign}(\widehat{x}_{t+1})x_{t+1}. \tag{2}$$

Thus, the key insight here is that while $\widehat{x}_{t+1}$ may or may not be a very good forecast for x_{t+1} as judged by fit, it may be good enough to correctly pick the direction of trade often. Let us denote the *ex post* correct direction of trade as $d_{t+1} = sign(x_{t+1}) \in \{-1, +1\}$, with -1 denoting a loss (trader should go short), and $+1$ a gain (trader should go long).

In fact, because a trader's problem is now a classification problem for d_{t+1}, it is possible that a return forecast $\widehat{x}_{t+1}$ does not even provide the best way to generate a prediction for d_{t+1}, say $\widehat{d}_{t+1}$. For this reason, and to maintain full generality, we shall consider any $\widehat{\delta}_{t+1}$ which can serve as a generic scoring classifier for d_{t+1}, where the directional forecast takes the form $\widehat{d}_{t+1} = \text{sign}(\widehat{\delta}_{t+1} - c)$ and $c \in (-\infty, \infty)$ is a threshold parameter. For the moment, we set aside the discussion on the best method to obtain $\widehat{\delta}_{t+1}$. Here all that we require is that $\widehat{\delta}_{t+1}$ be a scalar that takes on any values in $(-\infty, \infty)$.

One of the most important distinctions to be noted at this stage is that the classification problem of determining a "good" $\widehat{d}_{t+1}$ is usually far simpler than the forecasting problem

of determining a "good" $\widehat{x}_{t+1}$. This fact is of fundamental importance for the application of the methods we develop in this paper: if directional classification is an easier problem than forecast fit, then directional tests should impose a much higher hurdle for the "coin-toss" null in tests of market efficiency.

Moreover, by focusing on the classification problem, we are not constrained by traditional loss functions (such as the ubiquitous *root mean squared error,* RMSE) associated with $\widehat{x}_{t+1}$. In fact, many of the methods that we discuss below are considerably less reliant on any specific loss function: $\widehat{\delta}_{t+1}$ could be a probability forecast from a binary regression model, a single-index model from a dimension-reduction procedure, an ordinal variable generated from a discrete-state model, or even $\widehat{x}_{t+1}$. For this reason, many of the methods we introduce are non-parametric.

Of course, if one can correctly specify the data generating process, the choice of loss function (under general conditions) is less relevant: one still obtains unbiased estimates of the true parameters. Elliott and Lieli (2007) offer an alternative and attractive view on the classification problem that consists in tailoring the estimator to the agent's specific utility function. However, when one thinks of the statistical model as simply an approximation, different loss functions result in different models and parameter estimates, and therefore different conclusions about what variables are preferable (see Hand and Vinciotti, 2003 for a discussion on this point).

We find it useful to provide a simple example in Table 1 to illustrate these basic points. We consider a hypothetical investment problem: a one-period currency return, say, which has four discrete outcomes, percentage returns of $+2, +1, -1, -2$. Each outcome occurs $1/4$ of the time. If the investor goes long he receives the payoff outcome as above; if he goes short he receives minus one times the payoff.

There are two candidate trading signals available to the investor. Signal A is perfectly accurate in predicting the ± 1 outcomes but has an additive white-noise $N(0, 10)$ error on the ± 2 outcomes. Signal B is perfectly accurate in predicting the ± 2 outcomes but has the additive $N(0, 10)$ error on the ± 1 outcomes. Which signal is preferred? The answer depends on the criterion used.

According to the traditional RMSE criterion widely used in the forecast evaluation literature in economics and finance, there is nothing to choose between the two signals. They each have exactly the same error variance which appears on 50% of the observations. In large samples the RMSE of both signals is $\sqrt{10^2/2}$ or about 7.07. Signals A and B are equally good (or bad), and a decent test of predictive ability based on RMSE performance

Table 1: Mixed Signals: RMSE versus Direction versus Profitability

Signal type	Outcome y	Signal x	RMSE	Correct sign (%)	Profit
A	$y = \pm 1$	$x = y$	7.071	78.96	0.6585
	$y = \pm 2$	$x = y + \epsilon$			
B	$y = \pm 1$	$x = y + \epsilon$	7.071	76.99	1.0398
	$y = \pm 2$	$x = y$			

Note: $\epsilon \sim N(0, 10)$ is an i.i.d error.

should say so.

According to a directional criterion that measures the fraction of correct directional calls made, the answer is that Signal A is slightly better. Both signals get half the calls exactly right. What about the calls subject to noise? Here Signal A is only wrong a fraction $\Phi(-2/10)$ of the time; but Signal B is wrong a fraction $\Phi(-1/10)$ of the time, which is to say, slightly more often, where Φ denotes the cumulative Gaussian distribution. Overall, signal A makes correct calls 79% of the time, and signal B 77% of the time. A good directional accuracy test should prefer Signal A.[2]

Are either of the above the "right" answer in terms of investment performance or market efficiency? For a risk neutral investor, the answer is no. If we look at expected profits, then clearly Signal B is the better signal. It makes mistakes when the stakes are small, but it gets on the right side of trades when there are large potential profits or losses, whereas Signal A does just the opposite. The expected profit from Signal B is 1.04 on average per period, but from Signal A it is just 0.66.[3]

We can take away three important lessons from this discussion. First, the widely used RMSE statistic may be a good way to evaluate forecast accuracy or fit, but it is a misleading indicator of the presence of profitable arbitrage opportunities, which is the key metric for judging market inefficiency. Second, we can see that if signal errors were evenly spread over all outcomes, a good directional performance would clearly deliver excess returns. Finally, as we have just seen, if signal accuracy varies with returns, it will prove to be of

[2] The closed form expression for correct calls with Signal A is $\frac{1}{2} + \frac{1}{2}\Phi(2/10)$, and with Signal B it is $\frac{1}{2} + \frac{1}{2}\Phi(1/10)$.

[3] The closed form expression for profit with Signal A is $\frac{1}{2} + \frac{2}{2}(\Phi(2/10) - \Phi(-2/10))$, and with Signal B it is $\frac{2}{2} + \frac{1}{2}(\Phi(1/10) - \Phi(-1/10))$.

greater import for traders' profits (and, thus, for claims of market inefficiency) to make good directional forecasts when potential profits are large.

2 The Trader's Classification Problem

We find it useful to consider the following classification table associated with the binary decision problem the trader faces for a given investment strategy:

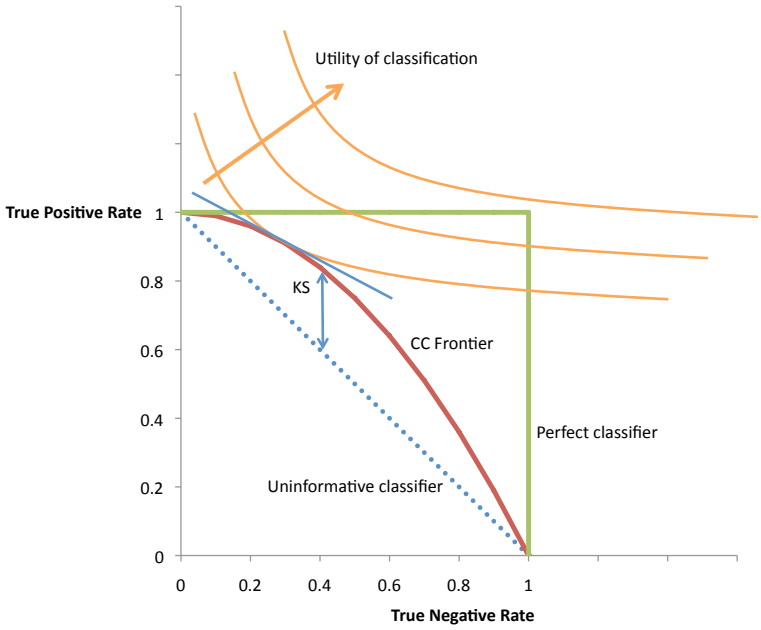
		Prediction	
		Negative/Short	Positive/Long
Outcome	Negative/Short	$TN(c) = P\left(\widehat{\delta}_t < c \mid d_t = -1\right)$	$FP(c) = P\left(\widehat{\delta}_t > c \mid d_t = -1\right)$
	Positive/Long	$FN(c) = P\left(\widehat{\delta}_t < c \mid d_t = +1\right)$	$TP(c) = P\left(\widehat{\delta}_t > c \mid d_t = +1\right)$

Here, $TN(c)$ and $TP(c)$ refer to the true classification rates of negatives and positives, respectively; and $FN(c)$ and $FP(c)$ refers to the false classification rates of negatives and positives, respectively. Clearly, $TN(c) + FP(c) = 1$ and $FN(c) + TP(c) = 1$. In statistics, $TP(c)$ is sometimes also called *sensitivity* and $TN(c)$, *specificity*. It may also be helpful to think of $\widehat{\delta}_t$ as the value of a test statistic and c as its critical value. Then $FP(c)$ would refer to the Type I error rate or size of the test, and $TP(c)$ its power.

The space of combinations of $TP(c)$ and $TN(c)$ for all possible values of $c \in (-\infty, \infty)$ summarizes a sort of "production possibilities frontier" (to use the traditional microeconomics nomenclature in a market for two goods) for the classifier $\widehat{\delta}_t$, that is, the maximum $TP(c)$ achievable for a given value of $TN(c)$. We will call the curve that summarizes all possible combinations $\{TN(c), TP(c)\}$ the *correct classification frontier* or *CC frontier*.

Of course, this is not the only way to summarize the performance of the classifier. Note that $FP(c) = 1 - TN(c)$, so another curve that is widely used in statistics and that summarizes all possible combinations $\{FP(c), TP(c)\}$ is called the *receiver operating characteristic* (ROC) curve, as we discussed in the introduction. And combinations $\{FN(c), TN(c)\}$ can be collected in a plot that is called the *ordinal dominance curve* (ODC) as discussed in Bamber (1975). Notice that the *CC* frontier and the ROC curve are the mirror image of one another (if one were to place the mirror at the vertical axis).

Figure 1: The Correct Classification Frontier (CCF)



A stylized plot of a CC frontier is presented in Figure 1. Notice that as $c \to -\infty$ then $TP(c) \to 1$ and $TN(c) \to 0$, and the limits are reversed as $c \to \infty$. For this reason, it is easy to see that the CC frontier lives in the unit square $[0,1] \times [0,1]$. A perfect classifier is one for which $TP(c) = 1$ for any $TN(c)$ and this corresponds to the north and east sides of the unit-square. An uninformative classifier on the other hand, is one where $TP(c) = FP(c) = 1 - TN(c)$ $\forall c$ and this corresponds to the north-west/south-east "coin-toss" diagonal. Using the language of the pioneering statistician Charles Sanders Peirce (1884), the classifiers corresponding to these two extreme cases would be referred to as the "infallible witness" and the "utterly ignorant person" (Baker and Kramer 2007, 343). Most CC frontiers in practice live in between these two extremes.

2.1 The Trader's Decision Problem

Using this type of decision theory, the investment decisions taken by the trader will depend on the CC frontier and on the utility the trader derives from each outcome—just as equi-

librium in the textbook two-goods market depends on the interaction of the production possibilities frontier and the consumer's utility. It might be reasonable (under conditions we shall examine in a moment) to assume that a risk-neutral investor's preferences in a frictionless world with symmetric returns will be tangent to the CC frontier where the marginal rate of substitution between true positives (profitable longs) and true negatives (profitable shorts) is -1. It turns out that the vertical distance between this point on the CC frontier and the coin-toss diagonal is given by the Kolmogorov-Smirnov (KS) statistic.

The KS statistic is based on the distance between the maximum value of the average correct classification rate for a given classifier, $\frac{1}{2}(TN(c)+TP(c))$, and the average correct classification rate for a coin-toss classifier. Since for the later $TP(c) = 1 - TN(c)\ \forall c$, this last average is easily seen to be $1/2$. Specifically the formula for the KS statistic is:

$$KS = \max_c \left| 2\left(\frac{TN(c)+TP(c)}{2} - \frac{1}{2} \right) \right|.$$

The KS statistic can be easily computed in practice. Let T_N (T_P) indicate the total number of observations in the sample $t = 1, ..., T$ for which $d_t = -1(+1)$, respectively, using the mnemonics N for negative and P for positive. Suppose that $T_P/T_N \to \lambda > 0$ as $T \to \infty$, where $T = T_N + T_P$, and empirically

$$\widehat{TN}(c) = \frac{\sum_{j=1}^{T_N} I\left(\widehat{\delta}_j < c\right)}{T_N}; \qquad \widehat{TP}(c) = \frac{\sum_{i=1}^{T_P} I\left(\widehat{\delta}_i > c\right)}{T_P} \tag{3}$$

where the indices $j(i)$ run over two sets of re-numbered observations, with each one mapping to a unique t such that $d_t = -1(+1)$, respectively and $I(.)$ is the indicator function that takes on the value of 1 when the argument is true and 0 otherwise. Then

$$\sqrt{\frac{T_N T_P}{T}}\widehat{KS} \to \sup_t |B(t)| \tag{4}$$

where $B(t)$ is a Brownian-bridge. That is, $B(t) = W(t) - tW(1)$ where $W(t)$ a Wiener process (see e.g. Conover 1999). Notice that $KS \in [0, 1]$ and is equivalent to maximum of the Youden (1950) J index, which is defined as

$$J(c) = TP(c) - FP(c). \tag{5}$$

Under the assumption that the investor's goal is to maximize the J index, we can identify the optimal operating point as the threshold c_{KS} where the KS statistic is maximized.

However, in practice returns may not be symmetric, and even a risk-neutral investor may face transaction costs when short-selling that he does not face when going long. Therefore, it is useful to explicitly cast the trader's utility of classification to account for all possible outcomes as

$$\begin{aligned} U(c) &= U_{pP}TP(c)\pi + U_{nP}(1-TP(c))\pi + \\ &\quad U_{pN}(1-TN(c))(1-\pi) + U_{nN}TN(c)(1-\pi). \end{aligned} \tag{6}$$

where $\pi = P(d = +1)$, that is, the unconditional probability of a positive; and U_{aA} for $a \in \{n, p\}$ and $A \in \{N, P\}$ is the utility associated with each of the possible four states defined by the (classifier,outcome) pair.

It may seem that applying the above methods to investment performance takes us down an unfamiliar track. But if we explore some simple utility weights we can show how J and U actually encompass some familiar and widely used investment performance criteria. If we weigh correct calls with a utility of 1, and incorrect calls with a utility of 0, then the utility measure U reduces to an accuracy rate

$$\text{Accuracy rate} = \frac{T_{pP}}{T} + \frac{T_{nN}}{T} = TP\pi + TN(1-\pi). \tag{7}$$

If we exchange these weights, then we obtain an error rate

$$\text{Error rate} = \frac{T_{nP}}{T} + \frac{T_{pN}}{T} = (1-TP)\pi + (1-TN)(1-\pi). \tag{8}$$

But the accuracy and error rates sum to one, so these are inversely related, and attain their respective maximum and minimum at the same choice of c.

We can now begin to see that the KS statistic is applicable only in a very special case. If positive and negative outcomes are equiprobable, that is $\pi = 1/2$, and if the utility from correct predictions (whether positive or negative) is normalized to be the same and equal to 1, and conversely, that the disutility from incorrect predictions (whether positive or negative) is the same and equal to -1, then expression (6) simplifies in such a way that

$U = J$:

$$\begin{aligned} U &= \frac{1}{2}TP(c) - \frac{1}{2}(1 - TP(c)) - \frac{1}{2}FP(c) + \frac{1}{2}(1 - FP(c)) \\ &= TP(c) - FP(c) = J(c). \end{aligned} \tag{9}$$

This is the same J index as in expression (5). In this case it is easy to show that the accuracy and error rates are $\frac{1+J}{2}$ and $\frac{1-J}{2}$. Thus, for this case only, all performance measures are monotonic in J, and on all performance criteria the same optimal c will be chosen.

In the seminal work of Peirce (1884), the expression for $J(c)$ was referred to as "the science of the method" and the general expression for $U(c)$ as "the utility of the method" (Baker and Kramer 2007). In Peirce's example, the applied problem was forecasting tornadoes, and his hypothetical utility weights corresponded to the net benefits of lives saved under true positives versus the costs of wasted resources or panic under false positives.

But in general, as Peirce understood, the choice of c that maximizes U need not be the one that maximizes J (or the accuracy rate, discussed below). In what follows we explore methods that allow π and U_{ij} to be generic. In finance problems, for realism, we want to allow the U_{ij} to be unrestricted since payoffs vary continuously, and we also want to allow for $\pi \neq \frac{1}{2}$ to admit the possibility of skewed payoff distributions.

Whether realized returns are systematically positive and significantly different from zero is determined primarily by a classifier's properties. In general settings, the utility derived from a given classifier will depend on the investor's attitude toward risk since each investment strategy is characterized by different combinations of returns, volatility and extreme events. Thus, the maximum of the Youden J index obtained from the KS statistic in expression (4) insufficiently characterizes an investor's choices—in other words, the simplifying assumptions used to derive expression (4) may not hold in practice.

To sidestep this problem, we can use the CC frontier to allow comparisons among classifiers without a need for specific statements about underlying preferences of the investor by considering all operating points simultaneously. Given the CC frontier's usefulness, it is helpful to develop further some intuition about the shape of the CC frontier and the properties of the the optimal operating point.

The CC frontier can be defined in terms of two distributions. Let u denote values of $\widehat{\delta}$ for which $d = 1$ and denote G its distribution and g its density so that $TP(c) = 1 - G(c)$. Similarly, let v denote values of $\widehat{\delta}$ for which $d = -1$ and denote F its distribution function and f its density so that $TN(c) = F(c)$.

We can now use the distributions F and G to define the CC frontier. Let us denote by $CC(r)$ the true positive rate corresponding to a true negative rate of r (since c uniquely determines both rates, this mapping is one-to-one). Hence $CC(r) = 1 - G(F^{-1}(r))$ with $r \in [0, 1]$. Notice then that the maximum utility from expression (6) is achieved when

$$\frac{dCC(r)}{dr} \equiv \frac{g(F^{-1}(r))}{f(F^{-1}(r))} = -\frac{1-\pi}{\pi}\frac{(U_{nN} - U_{pN})}{(U_{pP} - U_{nP})} \tag{10}$$

so that it is easy to see that the slope of the CC frontier is the likelihood ratio between the densities f and g. If this likelihood ratio is monotone, then the CC frontier is concave. In practice, one can make parametric assumptions about f and g and hence construct parametric models of the CC frontier. However, in the remainder of the paper we restrict our attention to non-parametric estimators because returns distributions are often poorly characterized by conventional Gaussian assumptions. The reader is referred to Pepe (2003) for an overview of parametric ROC models, which can be applied to CC frontier estimation.

The main point of the last equation is to show that, in general, the optimal operating point is at a slope that is skewed away from -1 in a way that depends on the relative probability of each outcome, and the utility weights. For example, in the last expression, suppose P is the event "cancer of type X" and upon that signal surgery will occur. All else equal, i.e., holding utility weights constant, if X gets very rare (π smaller, and the first fraction is larger), then a more conservative classifier should be used, with CC frontier steeper at the optimal point, typically nearer to (1,0) in Figure 1. One the other hand, holding the probability π constant, if, say, X is a more dangerous type of cancer then the costs of a false negative (U_{nP}) go up all else equal, then the second fraction gets smaller, and a more aggressive classifier should be used, with CC frontier flatter at the optimal point, typically nearer to (0,1) in Figure 1. These results are very intuitive indeed, although again we caution that the utility space is limited to four discreet outcomes, a restriction we shall seek to relax in a moment as we adapt these techniques for applications with variable payoffs in economics and finance.

3 A Standard Measure of Classification Ability: The Area under the CC Frontier

As a complement to the KS statistic we discussed, a more general measure of classification ability with conventional statistical properties is the area under the CC frontier or AUC. From Figure 1, it is clear that a perfect classifier will have $AUC = 1$ corresponding to the unit square. A coin-toss classifier has a CC frontier given by the diagonal that bisects this unit square and hence has $AUC = 0.5$. Formally

$$AUC = \int_0^1 CC(r)dr.$$

We should note that the AUC defined here has the same properties as the area under the ROC curve and the area under the ordinal dominance curve (see Hsieh and Turnbull 1996).

Of course, a perverse classifier can generate an $AUC < 0.5$, but, by reversing the classifier's predictions, one can obtain a classifier with $AUC > 0.5$, which we shall henceforth take to be the typical case. We remark that for two classifiers based on models A and B, $CC_A(r) > CC_B(r)\ \forall r$ means that classifier A stochastically dominates classifier B regardless of investor preferences. However, although in this case it is necessarily implied that $AUC_A > AUC_B$, it does not follow that $AUC_A > AUC_B$ is a sufficient condition for $CC_A(r) > CC_B(r)\ \forall r$. This is because the two CC frontier might cross, and, in fact, it could be the case that $CC_A(r) < CC_B(r)$ precisely for a value (or range of values) of r that is optimal for the trader. We will return to procedures that directly test the null $H_0 : CC_A(r) = CC_B(r)$ for any r momentarily, but before discussing the more difficult statistical procedure to test this hypothesis, it is more convenient to first present the results for tests that can be used for inference on AUC.

Green and Swets (1966) provide a nice interpretation of $AUC = P[v < u]$. Thus, like the Kolmogorov-Smirnov statistic, the AUC is a comparison of the distance between the distributions for v and u, except that it is F minus G averaged over all values of r rather than evaluated at the maximum:

$$\int_0^1 [F(c) - G(c)]dr = \int_0^1 [F(F^{-1}(r)) - 1 + CC(r)]dr = \int_0^1 CC(r)dr = AUC.$$

Not surprisingly, then, AUC is related to the Wilcoxon-Mann-Whitney U-statistic (see

Bamber 1975; Hanley and McNeil 1982), which is a rank-sum statistic. A simple empirical estimate of $P[v < u]$ and hence the AUC, can be obtained as

$$\widehat{AUC} = \frac{1}{T_N T_P} \sum_{j=1}^{T_N} \sum_{i=1}^{T_P} \left\{ I(v_j < u_i) + \frac{1}{2} I(u_i = v_j) \right\} \tag{11}$$

where the last term is used to break ties.

The empirical AUC turns out to have convenient statistical properties. If $T_P/T_N \to \lambda > 0$ as $T \to \infty$; F and G have continuous densities f and g respectively; and the slope of the CC frontier is bounded on any subinterval (a, b) of $(-1, 0)$, with $-1 < a < b < 0$; then Hsieh and Turnbull (1996) show that

$$\sqrt{T}\left(\widehat{AUC} - P[v < u]\right) \to N\left(0, \sigma^2\right). \tag{12}$$

In the special case that $F = G$

$$\sigma^2 = \frac{1}{12}\left(\frac{1}{T_N} + \frac{1}{T_P}\right).$$

Hanley and McNeil (1982) and Obuchowski (1994) provide a convenient approximation for the variance of AUC under general conditions given by

$$\sigma^2 = AUC(1 - AUC) + (T_P - 1)(Q_1 - AUC^2) + (T_N - 1)(Q_2 - AUC^2), \tag{13}$$

where

$$Q_1 = \frac{AUC}{2 - AUC}, \quad Q_2 = \frac{2AUC^2}{1 + AUC}.$$

Bootstrap procedures are also available (Obuchowski and Lieber 1998), although large sample approximations have been found to do well in relatively small samples (Pepe 2003).

4 Formal Inference for Directional Trading Strategies

Sections 2 and 3 have introduced tools to evaluate the classification performance of competing trading strategies. In this section we discuss four basic inferential procedures associated to this problem: (1) a formal test to determine whether an investment strategy is statistically superior to a coin-toss null hypothesis; (2) an overall performance test between

competing investment strategies; (3) a test that assesses whether two investment strategies are statistically equivalent at all operating points; and (4) confidence intervals for the CC frontier at an investor's optimal operating point.

The asymptotic results in the previous section provide the obvious test for the first of these four hypotheses. Since the AUC of a coin-tosser is 0.5, then the asymptotic normal result of the AUC in expression (12) provided by Hsieh and Turnbull (1996); and the approximate formula of the variance of the AUC in expression (13) from Hanley and McNeil (1982) are all that is required to construct a typical z-ratio with an approximate standard normal distribution in large samples. Obuchowski and Lieber (1998) investigate bootstrap percentiles, bootstrap-z and the bootstrap bias-corrected accelerated method and conclude, from a Monte Carlo study, that the asymptotic approximation can become unstable when the true $AUC \to 1$ and with samples of 200 observations or less. However, in financial data, sample sizes tend to be considerably larger and the AUC close to 0.5 (since AUCs close to 1 would indicate wildly profitable strategies!).

The relative performance of two investment strategies (say A and B) is of obvious interest. Here one could take advantage of the asymptotic Gaussian result in expression (12) to construct the z-ratio of the null hypothesis $H_0 : AUC_A = AUC_B$, that is

$$z = \frac{AUC_A - AUC_B}{\left(\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B\right)^{1/2}} \to N(0,1), \tag{14}$$

where σ_j^2 for $j \in \{A, B\}$ refers to the variance of the AUC given in expression (13) for each investment strategy, and ρ refers to the correlation between AUC_A and AUC_B. Hanley and McNeil (1983) propose estimating ρ as the average of the correlation $\rho(v_A, v_B)$ for the short positions and the correlation $\rho(u_A, u_B)$ for the long positions.

However, rejecting $H_0 : AUC_A = AUC_B$ (e.g., because $AUC_A > AUC_B$) could be because $CC_A(r) > CC_B(r)$ for $r \in R_1$ and $CC_A(r) < CC_B(r)$ for $r \in R_2$, where R_1 and R_2 denote two non-overlapping regions that span the real line. In such a case, some investors will prefer strategy B because it is superior precisely at the operating points where their utility is maximized. In such cases, the outer-envelope of CC_A and CC_B will be stochastically larger than any individual strategy. Of course, if an investor's utility function is known, then there is no difficulty in determining which classifier would be preferred.

Venkatraman and Begg (1996) provide a testing procedure that allows one to directly

test the null that the CC frontier for strategies A and B are statistically equivalent *at all operating points*. This test has the further attraction that it is distribution-free and can be implemented as a permutation test from which p-values are easily constructed by simulation.

Specifically, let $\{S_i^A\}_{i=1}^T$ and $\{S_i^B\}_{i=1}^T$ denote the ranks of $\{\widehat{\delta}_i^A\}_{i=1}^T$ and $\{\widehat{\delta}_i^B\}_{i=1}^T$ respectively (the signals from each investment strategy). Let the index $k = 1, ..., T-1$. Then define the empirical error matrix by

$$e_{ik} = \begin{cases} 1 & \text{if } (S_i^A \leq k, S_i^B > k, d_i = -1) \text{ or } (S_i^A > k, S_i^B \leq k, d_i = 1), \\ -1 & \text{if } (S_i^A > k, S_i^B \leq k, d_i = -1) \text{ or } (S_i^A \leq k, S_i^B > k, d_i = 1), \\ 0 & \text{otherwise,} \end{cases}$$

and the associated statistic

$$E = \sum_{k=1}^{T-1} \left| \sum_{i=1}^{T} e_{ik} \right|. \tag{15}$$

It is easy to see that this statistic focuses on the differences in predictions at each k^{th} operating point. To obtain the critical value for this statistic, one can obtain the percentile from a large number of draws of the statistic (15) from randomly exchanging the ranks between the two investment strategies and reranking them. To do this in practice, let $(q_1, ..., q_N)$ denote randomly drawn sequences of 0's and 1's and generate resamples $\{\widehat{S}_i^A, \widehat{S}_i^B\}$ using

$$\widehat{S}_i^A = q_i S_i^A + (1 - q_i) S_i^B \text{ and } \widehat{S}_i^B = q_i S_i^B + (1 - q_i) S_i^A$$

with a random coin-toss rule to break ties introduced by the permutation process.

The final inferential problem of interest concerns the error bands for the CC frontier for a given operating point. Such error bands can be constructed using asymptotic approximations based on results in Hsieh and Turnbull (1996). Specifically, an interval at the operating point r with approximate $1 - \alpha$ probability coverage is

$$S = \left\{ \widehat{CC}(r) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \widehat{\sigma}(r) \right\},$$

so that

$$P(CC(r) \in S) = 1 - \alpha,$$

and we can use Hsieh and Turnbull's (1996) formula for the variance:[4]

$$\sigma^2(r) = \frac{G\left\{F^{-1}(r)\right\}\left(1 - G\left\{F^{-1}(r)\right\}\right)}{T_P} + \left[\frac{g\left\{F^{-1}(r)\right\}}{f\left\{F^{-1}(r)\right\}}\right]^2 \frac{r(1-r)}{T_N},$$

substituting for all theoretical quantities with the empirical counterparts.

5 Beyond Basics: Returns-Weighted Classification Ability

This section improves on the framework presented thus far in a number of practical directions of interest to finance practitioners. Classification alone is insufficient to assess an investment strategy's success. A strategy that correctly picks 99-in-100 one penny trades and misses the direction on the 1-in-100 dollar trade has almost perfect classification skill but is a money-loser. Hence the first improvement we consider is to construct return-weighted variants of the KS and CC frontier and related statistics. This is done in the next section. Anatolyev and Gerko (2005) make this same point in regard to Pesaran and Timmermann's (1992) directional accuracy test.

In more realistic scenarios, transactions costs may make staying in cash positions appealing when expected returns are insufficient to cover costs. The second improvement we investigate extends the binary classification problem to include a third category and therefore account for long/cash/short positions. This gives rise to a modification of the AUC statistic to three dimensions called the *volume under the surface* or VUS. And just as we worry about weighing classification by returns in binary classification, we show how to construct return-weighted VUS statistics and explore their properties.

[4] Pepe (2003) suggests that such intervals may be imprecise when r is close to 0 or 1 and proposes instead a back-transformation of the interval generated by

$$\text{logit}\left(\widehat{CC}(r)\right) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\widehat{\sigma}(r)}{\widehat{CC}(r)\left(1 - \widehat{CC}(r)\right)}.$$

When there is reason to fear that the asymptotic approximation is inadequate, one can construct the usual t-percentile bootstraps although Hall, Hyndman and Fan (2004) caution that the theoretical properties of the bootstrap in this type of problem are quite complex because of the different smoothing choices to be made in calculating $\widehat{G}$, $\widehat{F}$, $\widehat{g}$, and $\widehat{f}$. The reader is referred to their paper to notice the oversmoothing bandwidths required to manage the coverage error rate to be $o(T^{-2/3})$.

5.1 The Return-Weighted CC Frontier: $CC^\star$

The maximally attainable profits of a "perfect" trader who always takes the right side of every trade can be broken down into two parts according to whether the return outcome is positive or negative, P or N (so the perfect trader would take the position that is long or, respectively, short). These maximal profits are given by

$$B = \sum_{d=+1} x_t, \quad C = \sum_{d=-1} |x_t| .$$

We can then construct some weights for each P and N outcome:

$$\begin{aligned} w_i &= \tfrac{x_t}{B} \quad \text{if} \quad \widehat{\delta}_i > c \text{ and } d_i = +1 \quad \text{for } i = 1, ..., T_P, \\ w_j &= \tfrac{x_t}{C} \quad \text{if} \quad \widehat{\delta}_j < c \text{ and } d_j = -1 \quad \text{for } j = 1, ..., T_N, \end{aligned}$$

where, as before, it is understood that the indices i and j each map P and N outcomes (respectively) to a unique observation t.

Using these weights we can modify the expressions in (3) to calculate return-weighted statistics $TN^\star(c)$ and $TP^\star(c)$ as

$$\widehat{TN^\star}(c) = \sum_{j=1}^{T_N} w_j I(\widehat{\delta}_j < c), \quad \widehat{TP^\star}(c) = \sum_{i=1}^{T_P} w_i I(\widehat{\delta}_i > c). \tag{16}$$

To develop some intuition, note that the former expressions represent actual profits as a fraction of the potential profits achieved by a "perfect" trader in the case where the outcome is N or P.

We can see that these fractions must lie between zero and one. Thus, by analogy with the CC frontier, we may define the $CC^\star$ frontier as the set of points $\{TN^\star(c), TP^\star(c)\}$ for $c \in (-\infty, \infty)$. By construction, $CC^\star$ still inhabits the unit square $[0, 1] \times [0, 1]$, and in a sense that we shall shortly make precise, return-weighted classification ability can be said to improve the further is the $CC^\star$ frontier from this diagonal. However, notice now that the slope of the $CC^\star$ frontier will be the likelihood ratio of the long/short densities weighed by the return/loss ratio. This can be easily seen from expression (10) by noticing that $w_j = U_{nN} - U_{pN}$ and $w_i = U_{pP} - U_{nP}$, i.e., when utility is just defined for a risk-neutral investor and normalized appropriately.

In the same way that the CC frontier leads to the AUC summary statistic, so the

$CC^\star$ leads to an analogous $AUC^\star$ statistic. Empirically, $AUC^\star$ can be estimated from expression (11) as

$$\widehat{AUC^\star} = \sum_{j=1}^{T_N} w_j \sum_{i=1}^{T_P} w_i \left\{ I(u_i > v_j) + \frac{1}{2} I(u_i = v_j) \right\}.$$

In terms of statistical properties, the Hsieh and Turnbull (1996) asymptotic normality results for the AUC in expression (12) provide, with minimal modification, the basis for large sample approximations for the $AUC^\star$. Notice that the weighting simply reranks the observations in each of the distributions associated with the long/short positions (but there are no observations that switch distributions as a result of the weighting). The asymptotic results only require regularity conditions on the resulting densities, call them $f^\star$ and $g^\star$. For this reason, it is necessary to introduce additional assumptions regarding the distribution of returns in long/short positions. Specifically, it is natural to require that in addition to the original assumptions in Section 3, $B/C \to \alpha > 0$ as $T \to \infty$ and that the densities of returns in long/short positions be continuous so that the resulting convolution with the original densities f and g, results in continuous densities $f^\star$ and $g^\star$.

An estimate of the variance of the $AUC^\star$ can be obtained by modifying expression (13) appropriately

$$\sigma_\star^2 = AUC^\star(1 - AUC^\star) + B(Q_1^\star - AUC^{\star^2}) + C(Q_2^\star - AUC^{\star^2}),$$

where $Q_1^\star$ and $Q_2^\star$ are defined as Q_1 and Q_2 in expression (13) by replacing AUC with $AUC^\star$. If instead, bootstrap procedures are preferred, it is natural to construct

$$u_i^\star = w_i u_i, \quad v_j^\star = w_j v_j,$$

and resample from $\left\{ u_i^\star, v_j^\star \right\}$. Other inferential procedures described previously can be similarly adapted as well.

5.2 Relationship to Other Performance Criteria

In our discussion of the CC frontier and its related J and AUC statistics, we were careful to spell out the intuitive meaning of the new concepts, and relate them to some existing directional-based investment performance measures. Similarly, we now take a moment to

bring out the links between our new return-weighted concepts of $CC^\star$, $J^\star$, and $AUC^\star$ and some existing profit-based investment performance measures.

We can obviously define a correspondingly adjusted Youden $J^\star$ statistic as

$$J^\star(c) = TP^\star(c) - FP^\star(c)$$

where $FP^\star(c)$ can be constructed in analogous manner to $TP^\star(c)$. This statistic measures the height of the $CC^\star$ curve above the diagonal, and it has a corresponding Kolmogorov-Smirnov statistic $KS^\star$ at its maximum, which may be used for inference.

As for utility-based measures, we can make a start by computing the total upside gains G and downside losses L achieved by the classifier as

$$G = B.TP^\star + C.TN^\star,$$

$$L = B.FN^\star + C.FP^\star.$$

The former sums up over all winning bets, and the later over all losing bets.

Following our previous discussion, we might be inclined to define utility here as net profit, that is gains minus losses,

$$\text{Net profit} = G - L = B.TP^\star + C.TN^\star - B.FN^\star - C.FP^\star.$$

However, this suffers from the problem that B and C are potentially unbounded as the sample size grows large. Therefore, we elect to define utility as net profit attained divided by the total potential profit attained by the "perfect" trader, where the latter is equal to $B + C$, which in turn also happens to be equal to $G + L$. With that scaling we may define utility as

$$\begin{aligned} U^\star = \frac{G - L}{G + L} &= \frac{B.TP^\star + C.TN^\star - B.FN^\star - C.FP^\star}{B + C} \\ &= \frac{B(2TP^\star - 1) + C(2TN^\star - 1)}{B + C}, \end{aligned}$$

where we use $TP^\star + FN^\star = 1$ and $TN^\star + FP^\star = 1$. A similar, but in that case superfluous, rescaling could have been applied to our definition of utility U above, without changing any results.

It is now easy to see that

$$\mathrm{U}^\star = \frac{(G/L) - 1}{(G/L) + 1}.$$

Hence, maximizing the utility of the trading strategy is the same as maximizing the gain-loss ratio G/L of the strategy, where the definition of this ratio matches precisely the well-known Bernardo and Ledoit (2000) gain-loss ratio for the risk-neutral case, a measure widely used by finance practitioners.

Taking differentials, maximizing utility is achieved when

$$B\Delta TP^\star + C\Delta TN^\star = 0$$

or when

$$\frac{\Delta TP^\star}{\Delta TN^\star} = -\frac{C}{B}$$

Thus, the optimal threshold is the point on the $CC^\star$ curve with slope $-C/B$. The importance of utility asymmetries comes to the fore again. Here, if the investment has asymmetric returns then the trading strategy will be optimally tilted in this direction. When $B > C$, the total returns from correct long bets are larger than those from correct short bets. Even if a classifier isn't perfectly informative, it is therefore optimal to move to a flatter point on the CC frontier, with slope less than 1, and be aggressive in taking more long bets since that's where more of the money is. Conversely, when $C < B$ it is better to tilt toward fewer long bets, make the slope bigger than 1 (in absolute value), and move to a steeper point and tilt toward going short. These arguments generalize to the case where there are $N > 2$ positions the investor can take as we shall see momentarily.

Finally, there is a very natural special case to consider. We could call this the *Long-Run Risk-Neutral Efficient Markets Hypothesis* (LRRNEMH). Equivalently we can invoke "long run fair pricing" in the terminology of Bernardo and Ledoit (2000). What this means is that naïve repeated long bets on $r > 0$ (or repeated short bets on $r < 0$) should return zero on average over many draws. That is, gains and losses should cancel out for such bets, with $B = C$ in large samples. Note that the condition $B = C$ is *not* saying that up and down moves are equally likely, but that upside and downside cumulative returns are equal. This may be a very natural assumption to make in some financial markets, even if there are substantial deviations from fair price or efficient markets in the short run. For example, there is ample evidence that long-run holding returns on different currencies are

identical, even if short-run carry trade strategies seem to make profits.[5]

In the special case $B = C$, the above expressions simplify, such that

$$U^\star = J^\star,$$

$$\frac{G}{L} = \frac{1 + J^\star}{1 - J^\star}.$$

Now the optimal operating point which maximizes net profit extraction given by utility $U^\star$, also maximizes the adjusted $J^\star$ statistic and maximizes the trading strategy's gain-loss ratio G/L. This point corresponds to the point on the CC* frontier with slope $= -1$.

5.3 Volume Under the Surface: *VUS* and *VUS**

In practical situations, expected returns may sometimes be insufficient to cover transactions costs and it will be of interest to consider investment strategies with long/cash/short directional positions. Specifically, suppose that the desired direction is given by

$$\begin{aligned} d_{t+1} &= -1 \quad if \quad x_{t+1} < \gamma_1, \\ d_{t+1} &= 0 \quad\;\; if \quad \gamma_1 \leq x_{t+1} \leq \gamma_2, \\ d_{t+1} &= +1 \quad if \quad x_{t+1} > \gamma_2, \end{aligned} \tag{17}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ are thresholds pre-determined by the economic problem and therefore known with certainty, and where $d = 0$ corresponds to a cash position (no bet). Here we could interpret the x_t to refer to the returns in excess of the risk-free rate or simply assign the no bet position some risk-free return. In addition a further simplification would consist in choosing $\gamma = -\gamma_1 = \gamma_2$ to represent transactions costs so that the investor will choose to trade only if $|x_{t+1}| > \gamma$. Alternatively, γ_1 and γ_2 could reflect the returns of a risk-free position which would differ if long/short positions incur in different transactions costs. The set-up is general enough to accommodate all of these possibilities. In fact, the methods that we discuss here extend readily to even more categories but will not be explored here explicitly (see Waegeman, De Baets and Boullart 2008 for such an extension).

Like before, we assume there is a model that generates a continuous signal $\widehat{\delta}_{t+1}$ but instead of varying a single operating point c as in section 4, we consider predictions of the

[5] On the evidence for the long run, see Alexius (2001), Fujii and Chinn (2000), and Sinclair (2005).

ordered categorical positions given by

$$\begin{aligned}\widehat{d}_{t+1} &= -1 \quad if \quad \widehat{\delta}_{t+1} \leq c_1, \\ \widehat{d}_{t+1} &= 0 \quad\;\; if \quad c_1 < \widehat{\delta}_{t+1} \leq c_2, \\ \widehat{d}_{t+1} &= +1 \quad if \quad c_2 < \widehat{\delta}_{t+1},\end{aligned}$$

for $c_1 < c_2$; $c_1, c_2 \in (-\infty, \infty)$. We collect these thresholds into the vector $\mathbf{c} = (c_1, c_2)$ for later use. The relation between the $\boldsymbol{\gamma}$ and the $\mathbf{c}$ is very indirect as it depends on what the $\widehat{\delta}_{t+1}$ refer to although of course, maximizing the TP for each category depends on the $\boldsymbol{\gamma}$ that define each category. In three dimensional space, the axes of the CC surface are slightly different than in the traditional case since each describes the true positive rate for each category (rather than the plot of the true positive rate against the true negative rate). The category-specific true positive rate can be easily estimated as

$$\widehat{TP}_h(\mathbf{c}) = \frac{1}{T_h} \sum_{d_h = \widehat{d}_h} I(\widehat{\delta}_{t+1} \leq c_1) + I(c_1 < \widehat{\delta}_{t+1} \leq c_2) + I(c_2 < \widehat{\delta}_{t+1})$$

where T_h refers to the number of observations in the sample that belong to category h. We collect these true positive rates into the vector $\mathbf{TP} = (TP_{-1}, TP_0, TP_{+1})$ and plot the resulting CC surface as the set of points $\mathbf{TP}$ $\forall \mathbf{c}$ which lies inside the three-dimensional unit cube $[0,1] \times [0,1] \times [0,1]$.

Associated with the CC surface described above is the volume under the surface (VUS). In the traditional 2-dimensional CC frontier, the AUC provides a measure of classification ability, with a coin-tosser establishing a lower bound $AUC = 0.5$ against which the classification ability of a model can be tested. As an alternative thought experiment, consider a signal which calls outcomes $1/2$ randomly with fixed probabilities p_1/p_2 where $p_1 + p_2 = 1$; the CC surface for this uninformative classifier is just the 1-simplex, under which lies an area equal to $1/2$.

Now in a 3-dimensional VUS, notice that the same coin-tosser achieves a lower bound of $1/6$ because although the probability of randomly classifying correctly any two categories is still $1/2$, there are now three possible classification pairs (hence $1/6 = 1/3 \times 1/2$). As an alternative thought experiment, consider a signal which calls outcomes $1/2/3$ randomly with fixed probabilities $p_1/p_2/p_3$ where $p_1 + p_2 + p_3 = 1$; the CC surface for this uninformative classifier is just the 2-simplex, under which lies a volume equal to $1/6$. The approach easily generalizes to classification problems with $N > 3$ categories: the CC hypersurface

then has $N-1>2$ dimensions, and the (hyper)volume under the CC (hyper)surface of $1/N!$ for an uninformative classifier.

The VUS is a direct extension of the Wilcoxon-Mann-Whitney statistic (see Mossman 1999). Let v_j denote the observations of $\widehat{\delta}$ for which $\widehat{d}=-1$; z_k when $\widehat{d}=0$; and u_i when $\widehat{d}=+1$ then an empirical estimate of $P[v<z<u]$ is readily seen to be

$$\widehat{VUS}=\widehat{P}[v<z<u]=\frac{1}{\prod_{h=1}^{3}T_h}\sum_{j=1}^{T_1}\sum_{k=1}^{T_2}\sum_{i=1}^{T_3}\left\{I\left(v_j<z_k<u_i\right),\right\} \tag{18}$$

where we omit a rule to randomly break ties in the interest of keeping the notation concise. Dreiseitl, Ohno-Machado and Binder (2000) provide analytic expressions for the variance of VUS as well as the covariance between VUS from two competing models and these are reproduced in the appendix for convenience although the reader is referred to their paper for a detailed explanation. These results can be used in conjunction with the asymptotic normality results for the Wilcoxon-Mann-Whitney statistic to conduct inference along the lines described in section 4, that is, a test of an investment strategy's performance against the coin-tosser, which takes the form of a test of the null $H_0: VUS=1/6$. They can also be used for comparisons between two alternative strategies with a test of the null $H_0: VUS_A=VUS_B$. In small samples one can rely instead on the bootstrap.

A $VUS^{\star}$ statistic that incorporates the returns of each category can be constructed in analogous manner to the construction of $AUC^{\star}$. Specifically, let

$$B=\sum_{d=+1}x_t; \qquad C=\sum_{d=-1}\left|x_t\right|; \qquad D=\sum_{d=0}i_t$$

where i_t refers to the returns of the no-bet position. However, in the interest of transparency, we will choose the no-bet position to be a cash position, in which case, $i_t=0$. In general multicategory settings, one would simply calculate the returns that the "infallible witness" would obtain for each category. We now construct the following weights

$$\begin{array}{ll} w_i=\frac{x_t}{B} & \text{if } \ \widehat{\delta}>c_2 \text{ and } d=1 \text{ for } i=1,...,T_3 \\ w_k=\frac{1}{T_2} & \text{if } \ c_1<\widehat{\delta}\leq c_2 \text{ and } d=0 \text{ for } k=1,...,T_2 \\ w_j=\frac{|x_t|}{C} & \text{if } \ \widehat{\delta}\leq c_1 \text{ and } d=-1 \text{ for } j=1,...,T_1 \end{array}$$

where recall from expression (18) that T_2 is the total number of observations in the cash bin and c_1 and c_2 are the thresholds that determine which decision is optimal for the investor.

Notice that in our example, the returns of all transactions in the no-bet position are zero, in other words, they are all equally good (or bad) and therefore, there is no reason to weigh one transaction more than another. However, if the middle category were associated with an alternative investment with unknown returns, the weights w_k would be constructed simply as $w_k = i_t/D$, just like with the other categories.

Hence, the empirical counterpart of the $AUC^\star$ for 3 categories is:

$$VUS^\star = \sum_{i=1}^{T_P} w_i \sum_{k=1}^{T_c} w_k \sum_{j=1}^{T_N} w_j I(v_j < z_k < u_i).$$

Standard errors can be calculated via bootstrap, an illustration of which is provided below.

Finally, we remark on the connection between our $VUS^\star$ statistic and other performance criteria, specifically Bernardo and Ledoit's (2000) gain-loss ratio. Consider the more general case in which there are $N > 2$ different investment categories indexed by h and let the per-category total winnings in each bin be B_h (which we defined as B, C, and D above). Given a classifier, the gains and losses from ex-ante good and bad bets are:

$$G = \sum_h B_h TP_h^\star; \qquad L = \sum_h B_h (1 - TP_h^\star)$$

where

$$(1 - TP_h^\star) = \sum_{\substack{j=1 \\ j \neq h}}^{N} FN_{h,j}^\star$$

since now a false negative can occur when one selects one of the other h categories. The net profit is then:

$$\text{Net Profit} = \sum_{h=1}^{N} B_h \left[TP_h^\star - (1 - TP_h^\star)\right] = \sum_{h=1}^{N} B_h \left[2TP_h^\star - 1\right].$$

Using the same rescaling $G + L$ that we used previously, then utility can be expressed as:

$$U^\star = \frac{G - L}{G + L} = \frac{\sum_{h=1}^{N} B_h \left[2TP_h^\star - 1\right]}{\sum_{h=1}^{N} B_h}.$$

The first-order conditions associated with this problem are now given by:

$$\frac{\partial TP_j^\star}{\partial TP_k^\star} = -\frac{B_k}{B_j}$$

for all pairs $j \neq k$ in N, holding fixed $TP_h^\star$ for $h \neq j, k$. Thus we get the same sort of equilibrium condition that we were getting in the two category, long/short case.

6 Empirical Applications: Equity and Currency Strategies

We illustrate the methods discussed in previous sections with examples based on some widely used equity and currency investment strategies. The equity strategies utilize a gamut of supposedly plausible signals, including data on prices, dividends, earnings, interest rates and spreads, and so on. For these signals we work with the well established data series developed by Goyal and Welch (2003) and Welch and Goyal (2008) and we can compare our forecasting metrics with theirs.[6] The currency strategies include the three common trading signals, carry, momentum and value, as well as a new strategy, all of which are discussed in detail Berge, Jordà and Taylor (2010).

6.1 *CC* and *CC** Frontiers for Equity Strategies

For our first application, we turn to one of the holy grails of financial economics, the problem of forecasting equity returns. In this section we will scrutinize the performance of stock trading rules drawn from a veritable kitchen sink of signals, following the most recent and state-of-the-art treatment by Welch and Goyal (2008). As these authors have shown, at first sight many signals may appear to be useful based on in-sample performance (IS), only to fail when confronted with the "gold standard" of predictive tests—the ability to provide an informative out-of-sample forecast (OOS).

The strategy to be evaluated is based on the monthly excess return on U.S. equities for 1927:1 to 2008:12 and defined as the return on the S&P 500 including dividends, minus the "risk-free rate" defined as the 3-month treasury bill rate. The investor's long/short positions are then determined by the following 14 indicators:[7] (1) the dividend price ratio,

[6] We use the new dataset of Goyal-Welch extended through 2008, available on Goyal's website: `www.bus.emory.edu/AGoyal/Research.html`. Their published paper uses data through 2005.

[7] All data are taken from Goyal and Welch (2008) and the 2009 vintage updates on Goyal's website: `www.bus.emory.edu/AGoyal/Research.html`

dp, computed as the difference between the log of dividends and the log of prices; (2) the dividend yield ratio, *dy*, computed as the difference between the log of dividends and the log of lagged prices; (3) the earnings price ratio, *ep*, computed as the difference between the log of earnings and the log of prices; (4) the dividend payout ratio, *de*, computed as the difference between the log of dividends and the log of earnings; (5) the stock variance, *svar*, computed as the sum of squared daily returns on the S&P 500; (6) the cross-sectional beta premium, *csp*, which measures the relative valuations of high- and low-beta stocks; (7) the book to market ratio, *bm*; (8) the net equity expansion, *ntis*, which is one of two measures of corporate issuing activity; (9) the long term yield, *lty*, on government bonds; (10) the long term return, *ltr*, on government bonds; (11) the term spread, *tms*, computed as the difference between the yield on long-term government bonds and the T-bill rate; (12) the default yield spread, *dfy*, computed as the difference between BAA- and AAA-rated corporate bond yields; (13) the default return spread, *dfr*, computed as the difference between returns on long-term corporate bonds and returns on long-term government bonds; and (14) the inflation rate, *linfl*, based on the CPI and lagged one month to allow for publication lags.

These signals are used for IS prediction over the full period, and OOS prediction using a long window from 1970:1 to 2008:1.[8] The latter window is chosen to be roughly consistent with the OOS windows used by Welch and Goyal (2008), who find that the inclusion or exclusion of the 1970s oil shock period can dramatically affect the performance of prediction strategies. We now report our results and compare our findings to those of Goyal and Welch (2008). The key difference to remember is that we will be using directional and realized profit criteria to judge the presence of unexploited arbitrage opportunities, and not the prevailing RMSE fit-based criterion. Again, this turns out to be significant as some methods that have high accuracy, may have low profit (and vice versa), so we can draw attention to whether a particular strategy can "fit where it matters".

Briefly, at a monthly frequency Welch and Goyal (2008, section 5) found a handful of strategies whose IS predictive power surmounted conventional significance tests. Under their RMSE criterion eight strategies were judged successful. However, once these eight strategies were subjected to a further OOS prediction test, only one, the *eqis* signal, was found to have superior IS and OOS performance relative to the null of using the historical mean return. The term spread *tms* was found to have marginal IS significance, but OOS

[8] The data for *csp* are only available from May 1937 to December 2002, so the sample sizes for this variable are slightly smaller in what follows.

significance. A few more signals were found to be promising when various truncations were applied to the data.

In addition to IS and OOS statistical inference based on the RMSE criteria, Welch and Goyal (2008) also consider the profitability of the candidate strategies by constructing a certainty-equivalent gain after postulating a utility function (Brennan and Xia 2004; Campbell and Thompson 2008). They note that (p. 1488–89) "This allows a conditional model to contribute to an investment strategy not just by increasing the mean trading performance, but also by reducing the variance...." They found that "In order, among the IS reasonably significant models, those providing positive CEV gains were *tms* (14bp/month), *eqis* (14bp/month), *tbl* (10bp/month), *csp* (6bp/month), *cay3* (6bp/month), and *ntis* (2bp/month)." However, the authors also indicate the dearth of available tests geared towards evaluating such improvements, since "we know of no better procedure to judge the economic significance of forecasting models,..." One of the goals of this paper is to provide just such a procedure, and one that can not only measure the economic significance of any gains, but also their statistical significance.

With these preliminaries, we now turn to our results. Our IS predictions are shown in Table 2, panel (a) using the CC-based evaluation tools. The contrast between the AUC and $AUC^{\star}$ results is striking, and shows that ranking the profitability of strategies can lead to a very different impression of their merits as compared to tests based on fit. According to the directional AUC test, only three signals surmount the conventional 5% significance level: *csp*, *tms* and *linfl*. Of these only *csp* was found to be significant IS in the Welch-Goyal findings. However, when we turn to the $AUC^{\star}$ test based on profits, the picture is quite different. Five signals do well by this yardstick, but they are, on the whole, very different signals: *dp, dy, ep, csp, bm, ntis*. Only *csp* performs well as judged by both direction and profit.

However, as we have noted, in-sample performance alone will not convince an appropriately skeptical reader. We now repeat our exercise but this time we compute the OOS performance of the signals, and we report these results in Table 2, panel (b). Using the directional AUC test four signals are significant at the 5% level, namely: *de, csp, bm, tbl*. Note that only one of these was statistically significant in the IS tests using AUC, namely *csp*.

Turning to the profit based $AUC^{\star}$ test, three signals are significant at the 5% level, *de, svar, csp, ltr*. Yet again, these signals were for the most part not statistically significant in the IS tests when using the same $AUC^{\star}$ test. The clear exception is *csp* which is the only

signal to achieve statistical significance in all four of our tests; a close call would be *svar* which was significant at the 10% level IS, and 5% OOS, using the $AUC^{\star}$ test.

To sum up, our CC-based tests provide a different way of judging the performance of equity trading signals, as compared to the more usual reliance on RMSE based criteria. Comparing the results of our tests to the state-of-the-art methods in Welch and Goyal (2008), we find important differences in the relative merits of different signals, but at the end of the day we arrive at what is essentially common ground.

Among equity trading signals, even when we switch to a criterion like $AUC^{\star}$ specifically designed to make precise inferences on the relative profitability of different strategies, we tend to find no evidence of a robust and stable relationship across IS and OOS predictions for most of the mainstream proposed trading signals. The single exception to this generalization applies to our findings for the *csp* signal (the cross-section premium), which we found to be highly statistically significant in all of our CC-based tests, thus lending support to the findings of Polk et al. (2006). Figure 2 displays the CC and $CC^{\star}$ curves for the *csp* signals examined in Table 3.

However, this support is still subject to two caveats. The first is conceptual, for as Welch and Goyal (2008, p. 1494) note, "[w]hat we call OOS performance is not truly OOS, because it still relies on the same data that was used to establish the models. (This is especially applicable to *eqis* and *csp*, which were only recently proposed.)" The second is qualitative, and based on the potential profitability of a *csp*-based strategy. Suppose a hypothetical investor went long when the OOS forecast was positive, short otherwise, their excess return, assuming no transaction costs or margin costs, would have been 27 bps/month (s.d. = 460 bps); or, on annualized basis 3.3 percent per year with a Sharpe Ratio of 0.20. So whilst there may have been predictable returns that could be judged statistically significant, not everyone would judge them economically significant.

7 Currency Carry Trades with Long/Cash/Short Positions

Berge, Jordà and Taylor (2010) examine the returns from bilateral currency carry trade strategies in which a trader borrows in one currency and lends in another while bearing the risk of appreciation. Four benchmark trading signals are examined in that paper. The first three are based on simple strategies commonly found in a variety of exchange traded funds (ETFs) and investible indices, such as the Deutsche Bank currency ETFs and Goldman Sachs' FX Currents. The fourth signal is based on a vector error correction model (VECM).

Table 2: AUC and $AUC^{\star}$ for Equity Strategies

We compute the CC frontier and related AUC statistics for the monthly return to the S&P 500, obtaining the return and a range of candidate trading signals from the dataset in Goyal and Welch (2008). All signals are start of period, except inflation which is lagged one month to allow for lags in the CPI announcement. The sample data run from 1927:1 to 2008:12.

(a) In-sample prediction (1927:1–2008:12)

Signal	Description	N	AUC		$AUC^{\star}$	
dp	Dividend price ratio	983	0.5127		0.5385	**
			(0.0187)		(0.0185)	
dy	Dividend yield ratio	982	0.5132		0.5413	**
			0.0187		(0.0185)	
ep	Earnings price ratio	983	0.5201		0.5767	***
			(0.0186)		(0.0183)	
de	Dividend payout ratio	983	0.4973		0.5257	
			(0.0187)		(0.0186)	
svar	Stock variance	983	0.5064		0.5352	*
			(0.0187)		(0.0186)	
csp	Cross-sectional premium	788	0.5499	**	0.5730	***
			(0.0206)		(0.0204)	
bm	Book to market ratio	983	0.4918		0.5416	**
			0.0187		(0.0185)	
ntis	Net equity expansion	983	0.5243		0.5414	**
			(0.0186)		(0.0185)	
tbl	3 month T-bill rate	983	0.5348	*	0.5162	
			(0.0186)		(0.0186)	
lty	Long term yield	983	0.5363	*	0.5291	
			(0.0186)		(0.0186)	
ltr	Long term return	983	0.5063		0.5272	
			0.0187		(0.0186)	
tms	Term spread	983	0.537	**	0.531	*
			(0.0186)		(0.0186)	
dfy	Default yield spread	983	0.4737		0.4742	
			0.0187		(0.0187)	
dfr	Default return spread	983	0.5201		0.5319	*
			(0.0186)		(0.0186)	
linfl	Inflation, lagged one month	982	0.5526	***	0.5154	
			(0.0185)		(0.0187)	

Notes: The AUC and $AUC^{\star}$ are asymptotically distributed $N(0.5, \sigma)$. Standard errors in parentheses. * (**) (***) denote statistical significance at the 10% (5%) (1%) level for a test where the null is that the area under the curve is equal to 0.5.

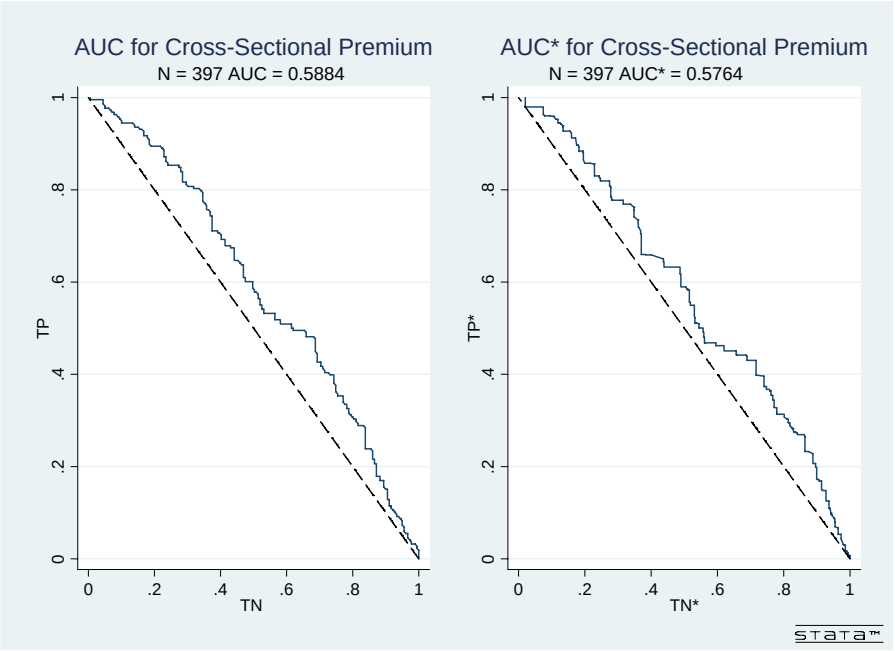
Table 2 (continued): AUC and $AUC^{\star}$ for Equity Strategies

(b) Out-of-sample prediction (1970:1–2008:12)

Signal	Description	N	AUC		$AUC^{\star}$	
dp	Dividend price ratio	468	0.4601		0.5073	
			(0.0269)		(0.0269)	
dy	Dividend yield ratio	468	0.463		0.5105	
			0.0269		(0.0269)	
ep	Earnings price ratio	468	0.4766		0.5335	
			(0.0269)		(0.0268)	
de	Dividend payout ratio	468	0.4277	***	0.4305	***
			(0.0267)		(0.0267)	
svar	Stock variance	468	0.503		0.4457	**
			(0.0269)		(0.0268)	
csp	Cross-sectional premium	397	0.5800	***	0.5753	***
			(0.0285)		(0.0286)	
bm	Book to market ratio	468	0.4358	**	0.4877	
			0.0268		(0.0269)	
ntis	Net equity expansion	468	0.5014		0.5036	
			(0.0269)		(0.0269)	
tbl	3 month T-bill rate	468	0.5541	**	0.5065	
			(0.0266)		(0.0269)	
lty	Long term yield	468	0.5499	*	0.4996	
			(0.0266)		(0.0269)	
ltr	Long term return	468	0.4933		0.5535	**
			0.0269		(0.0266)	
tms	Term spread	468	0.548	*	0.4973	
			(0.0266)		(0.0269)	
dfy	Default yield spread	468	0.4587		0.4952	
			0.0269		(0.0269)	
dfr	Default return spread	468	0.5065		0.5243	
			(0.0269)		(0.0268)	
linfl	Inflation, lagged one month	468	0.5459	*	0.5176	
			(0.0267)		(0.0268)	

Notes: The AUC and $AUC^{\star}$ are asymptotically distributed $N(0.5, \sigma)$. Standard errors in parentheses. * (**) (***) denote statistical significance at the 10% (5%) (1%) level for a test where the null is that the area under the curve is equal to 0.5.

Figure 2: Out-of-Sample CC and $CC^{\star}$ Frontiers for the *csp* Signal



Notes: CC and $CC^{\star}$ curves correspond to out-of-sample results in Table 3, panel (b), for *csp*. Standard errors for the AUC and $AUC^{\star}$ statistics are shown in parentheses.

We provide a brief description below but encourage the interested reader to refer to the original source for more details.

The *Carry Signal* c is computed as the interest differential between the local currency (LC) and the U.S. dollar (US). Under this strategy, the presumption is that high yield currencies will deliver profits despite the risk of depreciation. In this case uncovered interest parity either fails, or holds ex-ante but suffers ex-post from systematic and exploitable expectational errors. Thus $c_t = i_t^{LC} - i_t^{US}$, and the trader using this signal uses the model $\widehat{x}_{t+1} = c_t$ for each currency.

The *Momentum Signal* m is computed as rate of appreciation of the local currency exchange rate against the U.S. dollar $E^{LC/US}$ in the previous month. Under this strategy, the presumption is that appreciating currencies will have a tendency to keep appreciating on average. Thus $m_t = \Delta log E_t^{US/LC}$, and the trader using this signal uses the model $\widehat{x}_{t+1} = m_t$ for each currency.

The *Value Signal* v is computed as the undervaluation of the country's log CPI-

index-based real exchange rate level against the U.S. (IFS data) in the prior period $q = ln[E_{LC/US}P_{US}/P_{LC}]$, using deviation from average lagged levels $\bar{q}$ computed using a trailing window (to avert look-ahead bias). Under this strategy, the presumption is that currencies will have a tendency to return to their historic PPP value in the long run. Thus $v = q - \bar{q}$, and the trader using this signal uses the model $\widehat{x}_{t+1} = v_t$ for each currency.

Finally, the *vecm* signal is based on a panel VECM forecasting model for the holding return for each currency, where the dynamic interactions between nominal exchange rates, inflation and nominal interest rate deviations are its constituent elements.

The data include the nine currencies EUR, GBP, JPY, CHF, AUS, CAD, NZD, NOK, and SEK, with the USD as the base home currency (i.e., the "G-10" currencies), in a sample from 1986 to 2008 observed at monthly frequency. Table 3 presents the out-of-sample (OOS) $AUC/AUC^{\star}$ statistics associated to each of these four strategies for the 540 pooled currency-month observations in our chosen OOS sample window from 2004:1 to 2008:12

The results reported in Table 3 highlight once more the difference between good classification ability and profitability. For example, the *value* strategy does not classify direction significantly better than a coin-toss, but when trades are adjusted for return, clearly the *value* strategy outperforms a coin-tosser. Measured by this metric, the VECM strategy has the highest $AUC^{\star}$ at 0.6018, well above the 0.5 null and highly statistically significant. For a further perspective, the CC and $CC\star$ frontiers for the four strategies are shown in Figure 3.

Indeed, the trading profits delivered by the VECM strategy are not trivial. If a trader faced no transaction costs and could go long or short each currency at will, then a portfolio based on the signs of the signals from the OOS VECM model would have generated average returns of 34 bps *[checking with Travis for correct number]*per month on each position, with each trade having a standard deviation of 316 bps*[need new number]*. Thanks to diversification, the returns on the portfolio of 9 currencies had a standard deviation of 105 bps*[need new number]*. Annualized, the strategy would have delivered 1.04 percent per year compounded with a Annualized Sharpe Ratio of 1.13*[need new number]*.

Often times the signal generated by a strategy may be weak and the investor may prefer staying in a cash position, especially if there are transactions costs associated with each trade. In order to showcase how $VUS/VUS^{\star}$ statistics can be used in such situations, we redo the previous analysis but now allowing for long/cash/short positions. When the decision space is binary, there is no ambiguity in determining the ex-post profitable long/short

Table 3: Out-of-Sample AUC and $AUC^{\star}$ for Currency Strategies

We compute the AUC statistics based on the returns from currency trading using one of three signals, or a combination thereof. The nine currencies are EUR, GBP, JPY, CHF, AUS, CAD, NZD, NOK, and SEK, with the USD as the base home currency. The trading periods are months from 2004:1 to 2008:12. The data are from Jordà and Taylor (2009). The carry, momentum, and value signals are as described in the text. These signals are not based on any econometric model, only on lagged observables at the start of each trading period. In the case of the VECM, a country-fixed effect panel vector error-correction model is used to build an out-of-sample forecast, using an expanding lagged window which starts in 1986:1. The variables in the VECM model are the lagged change in the exchange rate, lagged interest differential, lagged inflation differential, and lagged real exchange rate deviation. The VECM signal is the one-step ahead forecast. In each period a long-short ±$1 bet is placed on each currency, based on the signal.

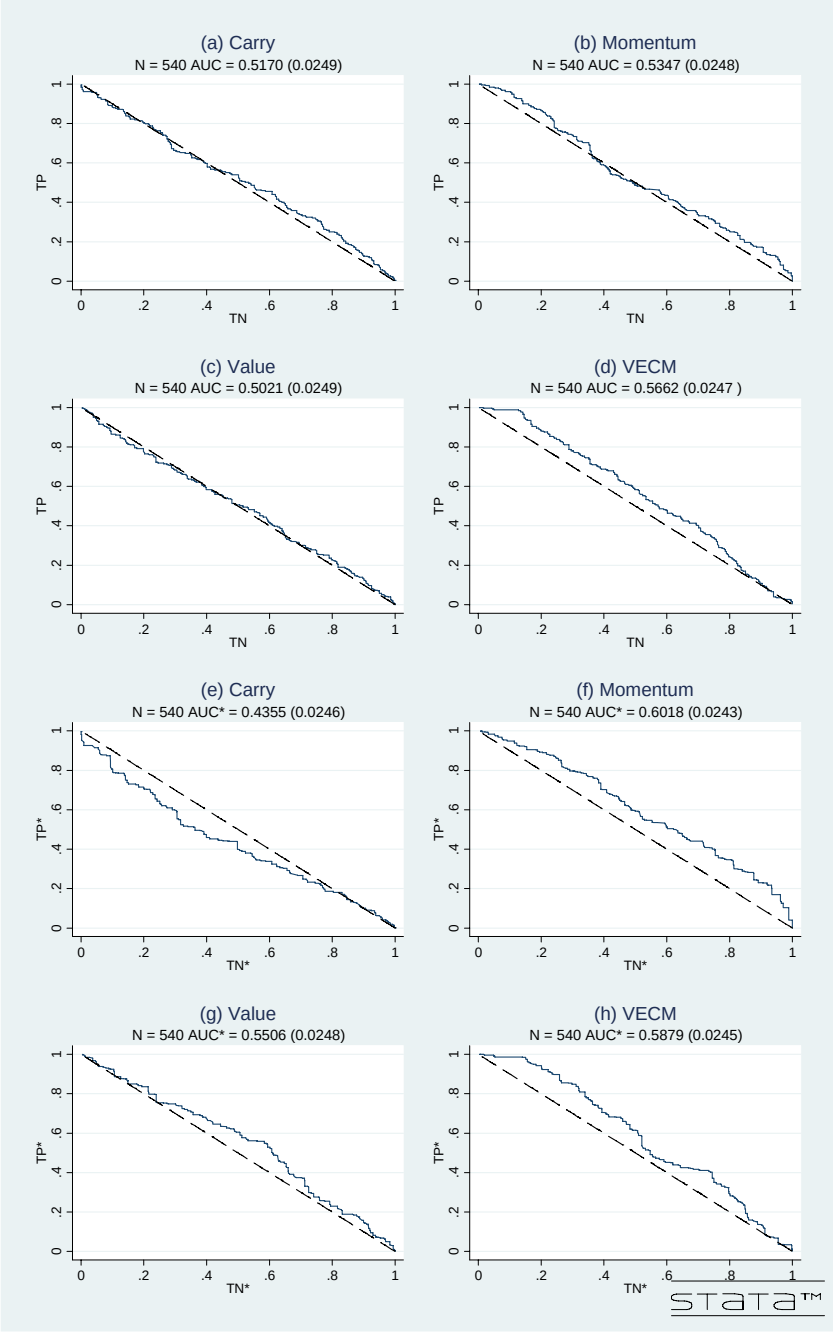
Signal	Description	N	AUC		$AUC^{\star}$	
c	Carry	540	0.5021		0.4355	***
			(0.0247)		(0.0243)	
m	Momentum	540	0.5170		0.5506	**
			(0.0248)		(0.0245)	
v	Value	540	0.5347		0.5879	***
			(0.0249)		(0.0246)	
$\hat{r}^{VECM}$	Panel VECM forecast	540	0.5662	***	0.6018	***
			(0.0249)		(0.0248)	

Notes: The AUC and $AUC^{\star}$ are asymptotically distributed $N(0.5, \sigma)$. Standard errors in parentheses. * (**) (***) denote statistical significance at the 10% (5%) (1%) level for a test where the null is that the area under the curve is equal to 0.5.

direction. However, by now adding a cash position, we need some criterion to determine the ex-post choice of long/cash/short positions.

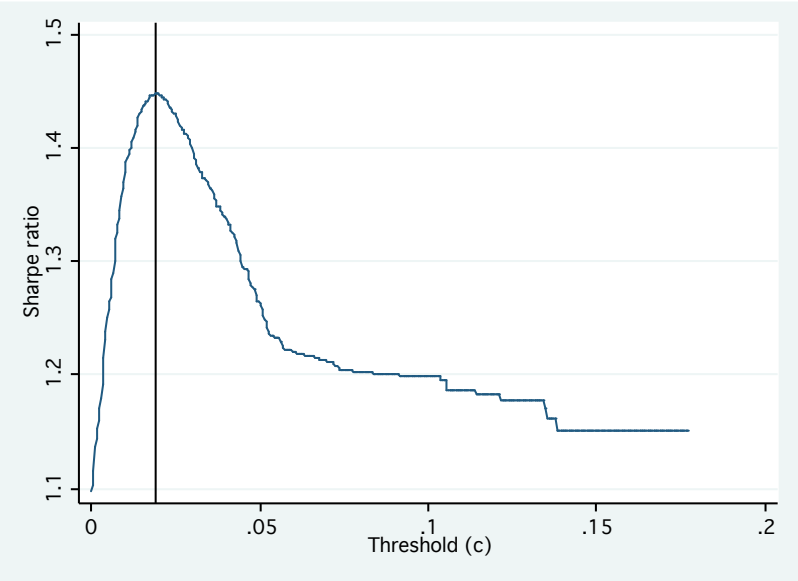
Absent good data on transactions costs, we decided to calculate a minimum symmetric return threshold γ beyond which a long/short perfect-foresight trade would be triggered, but otherwise the trader would remain in the cash position. In order to find such a threshold, we used a grid-search of values of γ that would maximize the ex-post Sharpe ratio for a $1 investment. This is reported in Figure 4 and shows that the ex-post Sharpe ratio is maximized for $\gamma = 1.91\%$. This results in a mean monthly return of 3.7% and an annualized Sharpe ratio of 1.45. These numbers may appear wildly optimistic but we remind the reader that they refer to the *perfect foresight* returns. With this choice of threshold, the investor would stay in the cash position about 50% of the time, and the other 50% of the time he would go long/short in equal proportion. Given this ex-post classification of the data, we

Figure 3: CC and $CC^{\star}$ for Four Currency Strategies



Notes: The CC and $CC^{\star}$ curves correspond to the results in Table 3.
Standard errors for the AUC and $AUC^{\star}$ statistics are shown in parentheses.

Figure 4: Sharpe Ratio as a Function of Threshold



can now ask how would the four benchmark carry trade strategies reported in Table 3 fare if one allowed for a cash position and for this we calculated each strategy's $VUS/VUS^{\star}$ statistics, the results of which are reported in Table 4.

Recall that the null of no classification ability (the equivalent of the coin-toss null in CC-space) is now $VUS = 1/6 \simeq 0.167$. By this metric, the *momentum, value* and *vecm* signals handily beat this simple null (standard errors are calculated with the bootstrap). The *carry* signal does not however, and in fact attains $VUS/VUS^{\star}$ values *below* 1/6, suggesting that the trader would be well advised to reverse the interpretation of the signal. However, it is interesting to see that when weighing by returns in $VUS^{\star}$, the *momentum* signal now appears to do better than the VECM signal (our previous favorite) and by a statistically significant amount (using bootstrapped confidence 95% confidence intervals). One explanation for this result is that, while VECM may be more consistent at picking the correct direction of a carry trade, it may be missing some of the high-profit trades that *momentum* is picking up. And in our VUS setup, the high profit trades take on even greater importance: remember that given our imposed thresholds, ex-post we remain in the cash position about 50% of the time and only trade when we can beat a 2% monthly return, which is rather conservative.

Table 4: Out-of-Sample VUS and $VUS^{\star}$ for Currency Strategies

We compute the CC frontier and related VUS statistics for the monthly return to the four trading strategies. The sample data run from 2004:1 to 2008:12.

Signal	Description	N	VUS		$VUS^{\star}$	
C	Carry	540	0.1648		0.1269	
			(0.017)		(0.014)	
M	Momentum	540	0.2067	***	0.2554	***
			(0.020)		(0.020)	
V	Value	540	0.1861	***	0.2025	***
			(0.024)		(0.020)	
vecm	VECM forecast	540	0.2061	***	0.2113	***
			(0.020)		(0.020)	

Notes: Bootstrap standard errors in parentheses. * (**) (***) denote statistical significance at the 10% (5%) (1%) level for a test where the null is that the area under the curve is equal to 1/6.

8 Conclusions

The presence of excess returns in a zero net-investment strategy does not per se violate the efficient markets hypothesis. But Bernardo and Ledoit (2000) construct bounds to these arbitrage opportunities, using the gain-loss ratio, that have implications for asset pricing in incomplete markets that are robust yet with sufficient texture to be economically compelling. Our paper is a compendium of statistical methods designed to investigate this sort of problem from a variety of angles interesting to academic researchers and investors alike.

We design techniques that allow one to compare alternative predictive models on the basis of profitability in a manner that is robust to variation in investor preferences. But our methods go beyond providing simple summary statistics, they also provide a complete description of an investor's choices. Formal inferential procedures are designed to test the null of absence of arbitrage; to test the relative overall profitability of competing investment strategies; to test whether a strategy is stochastically dominated by another; and to provide confidence bounds on optimal operating points.

In practice, specially (but not exclusively) when there are transaction costs, it is important to allow the investor to adopt a neutral position during those times when the expected return from the risky position is low. Allowing for such an extension can greatly enhance the overall profitability of a zero net-investment strategy and change the perceived

opportunities to arbitrage. Hence we develop extensions for such a case and along the way generalize our framework for more complex strategies involving multiple categories. We also show how these more sophisticated strategies can be related to Bernardo and Ledoit's (2000) gain-loss ratio.

We illustrate our methods with applications to the stock market and the carry trade. On the former, we show how Welch and Goyal's (2008) results based on RMSE metrics fare under our framework and show that, while there is perhaps one strategy with statistically significant returns, its risk-return characteristics are probably not in violation of conventional parameters of investor preferences. Our application to the carry trade is based on Berge, Jordà and Taylor (2010) and identifies a strategy that generates a statistically significant departure from no arbitrage (but not necessarily a violation of efficient markets), that is later shown to be dominated by another strategy if one allows the investor to adopt a neutral position.

The framework that we propose is non-parametric but simple to implements and makes explicit the connection between the statistical properties of the returns of investment positions, and the investor's preferences over such positions. Moreover, we show how this framework connects with a well established benchmark of asset pricing in incomplete markets, the gain-loss ratio. For these reasons we think our methods represent a viable standard approach to analyze an important class of problems in finance.

9 Appendix: The Variance of VUS

The variance for $\widehat{VUS}$ in expression (18) available in Dreiseitl, Ohno-Machado and Binder (2000) can be calculated as follows. Let $\widehat{\theta} = \widehat{VUS}$, then:

$$var(\widehat{\theta}) = \frac{1}{T_1 T_2 T_3}\left[\begin{array}{c}\theta(1-\theta) + (T_3-1)(q_{12}-\theta^2) + (T_2-1)(q_{13}-\theta^2) + (T_1-1)(q_{23}-\theta^2) + \\ (T_2-1)(T_3-1)(q_1-\theta^2) + (T_1-1)(T_3-1)(q_2-\theta^2) + (T_1-1)(T_2-1)(q_3-\theta^2)\end{array}\right]$$

where:

$$\begin{aligned}
q_1 &= P\left[I\left(v_j < z_k < u_i\right) = I\left(v_j < z_K < u_I\right)\right] \text{ for } K \neq k, I \neq i \\
q_2 &= P\left[I\left(v_j < z_k < u_i\right) = I\left(v_J < z_k < u_I\right)\right] \text{ for } J \neq j, I \neq i \\
q_3 &= P\left[I\left(v_j < z_k < u_i\right) = I\left(v_J < z_K < u_i\right)\right] \text{ for } J \neq j, K \neq k \\
q_{13} &= P\left[I\left(v_j < z_k < u_i\right) = I\left(v_j < z_K < u_i\right)\right] \text{ for } K \neq k \\
q_{23} &= P\left[I\left(v_j < z_k < u_i\right) = I\left(v_J < z_k < u_i\right)\right] \text{ for } J \neq j
\end{aligned}$$

and all population quantities can be substituted by their sample estimate equivalents.

References

Alexius, Annika. 2001. Uncovered Interest Parity Revisited. Review of International Economics, 9, 505–517.

Anatolyev, Stanislav and Alexander Gerko. 2005. A Trading Approach to Testing Predictability. *Journal of Business and Economic Statistics,* 23(4): 455–461.

Baker, Stuart G. and Barnett S. Kramer. 2007. Pierce, Youden, and Receiver Operating Characteristics Curves. *The American Statistician*, 61(4): 343–346.

Bamber, Donald. 1975. The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph. *Journal of Mathematical Psychology*, 12: 387–415.

Berge, Travis J., Òscar Jordà and Alan M. Taylor. 2011. Currency Carry Trades. In *International Seminar on Macroeconomics, 2010*, Richard H. Clarida, Jeffrey A. Frankel and Francesco Giavazzi (eds.), NBER. Forthcoming.

Bernardo, Antonio E. and Ledoit, Olivier. 2000. Gain, Loss and Asset Pricing. *Journal of Political Economy* 108(1): 144–172.

Brennan, Michael J.and Yihong Xia. 2004. Persistence, Predictability, and Portfolio Planning. Working paper, UCLA and Wharton.

Campbell, John Y. and Samuel B. Thompson. 2008. Predicting the Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Review of Financial Studies*, 21(4): 1509–1531.

Cheung, Yin-Wong, Menzie D. Chinn, and Antonio Garcia Pascual. 2005. Empirical Exchange Rate Models of the Nineties: Are Any Fit to Survive? *Journal of International Money and Finance* 24(7): 1150–75.

Cochrane, John H. 2001 *Asset Pricing* New Jersey: Princeton University Press.

Conover, W. J. 1999. *Practical Nonparametric Statistics.* 3rd edition. New York: John Wiley and Sons.

Dreiseitl, Stephan, Lucila Ohno-Machado and Michael Binder. 2000. Comparing Three-class Diagnostic Tests by Three-way ROC Analysis. *Medical Decision Making*, 20(3): 323–331.

Elliott, Graham and Robert P. Lieli. 2009. Predicting Binary Outcomes. Department of Economics, University of California, San Diego. Mimeograph.

Fujii, Eiji, and Menzie D. Chinn. 2000. Fin de Siècle Real Interest Parity. NBER Working Papers 7880.

Goyal, Amit and Ivo Welch. 2003. Predicting the Equity Premium with Dividend Ratios. *Management Science*, 49(5):639–654.

Green, David M. and John A. Swets. 1966. *Signal Detection Theory and Psychophysics.* Los Altos, CA: Peninsula Publishing.

Hájek, Jaroslav , Zbyněk. Šidák and P. K. Sen. 1999. *Theory of Rank Tests.* San Diego, CA: Academic Press.

Hall, Peter, Rob J. Hyndman and Yanan Fan. 2004. Nonparametric Confidence Intervals for Receiver Operating Characteristic Curves. *Biometrika*, 91(3): 743–750.

Hand, David J. and Veronica Vinciotti. 2003. Local versus Global Models for Classification Problems: Fitting Models Where It Matters. *The American Statistician*, 57(2): 124–131.

Hanley, James A. and Barbara J. McNeil. 1982. The Meaning and use of the Area Under the Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143: 29–36.

Hsieh, Fushing and Bruce W. Turnbull. 1996. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristics Curve. *Annals of Statistics*, 24: 25–40.

Jordà, Òscar, and Alan M. Taylor. 2009. The Carry Trade and Fundamentals: Nothing to Fear but FEER itself. NBER Working Papers no. 15518.

Kilian, Lutz, and Mark P. Taylor. 2003. Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics* 60(1): 85–107.

Meese, Richard A., and Kenneth Rogoff. 1983. Empirical Exchange Rate Models of the Seventies. *Journal of International Economics* 14(1–2): 3–24.

Mossman, Douglas. 1999. Three-way ROCs. *Medical Decision Making*, 19(1): 78–89.

Obuchowski, Nancy A. 1994. Computing Sample Size for Receiver Operating Characteristic Curve Studies. *Investigative Radiology*, 29(2): 238–243.

Obuchowski, Nancy A. and Michael L. Lieber. 1998. Confidence Intervals for the Receiver Operating Characteristic Area in Studies with Small Samples. *Academic Radiology*, 5(8): 561–571.

Peirce, Charles S. 1884. The Numerical Measure of the Success of Predictions. *Science* 4: 453–454.

Pepe, Margaret s. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press.

Peterson, W. Wesley and Theodore G. Birdsall. 1953. The Theory of Signal Detectability: Part I. The General Theory. Electronic Defense Group, Technical Report 13, June 1953. Available from EECS Systems Office, University of Michigan.

Pesaran, M.. Hashem and Allan Timmermann. 1992. A Simple Nonparametric Test of Predictive Performance. *Journal of Business Economics and Statistics*, 10(4): 461–65.

Polk, Christopher, Samuel Thompson and Tuomo Vuolteenaho. 2006. Cross-sectional Forecasts of the Equity Premium. *Journal of Financial Economics*, 81(1):101–41.

Sinclair, Peter J. N. 2005. How Policy Rates Affect Output, Prices and Labour, Open Economy Issues, and Inflation and Disinflation. In *How Monetary Policy Works*, edited by Lavan Mahadeva and Peter Sinclair. London: Routledge.

Spackman, Kent A. 1989. Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning. In *Proceedings of the Sixth International Workshop on Machine Learning.* Morgan Kaufman, San Mateo, Calif., 160–63.

Stanski, Henry R., Laurence J. Wilson and William R. Burrows. 1989. Survey of Common Verification Methods in Meteorology. Research Report No. 89-5, Atmospheric Environment Service, Forest Research Division, 4905 Dufferin Street, Downsview, Ontatio, Canada.

Swets, John A. and R. M. Pickett. 1982. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* New York: Academic Press.

Venkatraman, E. S. and Colin B. Begg. 1996. A Distribution-Free Procedure for Comparing Receiver Operating Characteristic Curves from a Paired Experiment. *Biometrika*, 83(4): 835–848.

Waegeman, Willem, Bernard de Baets and Luc Boullart. 2008. ROC Analysis in Ordinal Regression Learning. *Pattern Recognition Letters*, 29(1): 1–9.

Welch, Ivo and Amit Goyal. 2008. A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *Review of Financial Studies*, 21(4): 1455-1508. World Meteorological Organization. 2000. *Standard Verification System for Long-Range Forecasts.* Geneva, Switzerland: World Meteorological Organization.

Youden, W. J. 1950. Index for Rating Diagnostic Tests. *Cancer* 3, 32–35.