View metadata, citation and similar papers at core.ac.uk brought to you by CORE

provided by Research Papers in Economics

NBER WORKING PAPER SERIES

PERFORMANCE EVALUATION OF ZERO NET-INVESTMENT STRATEGIES

Òscar Jordà
Alan M. Taylor

Working Paper 17150
http://www.nber.org/papers/w17150

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2011

Taylor has been supported by the Center for the Evolution of the Global Economy at UC Davis and Jorda 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 later when he was a Senior Advisor at Morgan Stanley; all of this research support is gratefully acknowledged. We thank Colin Cameron, Miguel Delgado, Martin Evans, Pedro Gete, Gloria Gonz alez-Rivera, Guido Kuersteiner, Burkhard Schipper, Enrique Sentana and seminar participants at the Federal Reserve Bank of San Francisco, Georgetown University, Universidad Carlos III de Madrid, and University of California, Riverside for comments and suggestions. Travis Berge and Yanping Chong provided superb research assistance. All errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2011 by Òscar Jordà and Alan M. Taylor. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Performance Evaluation of Zero Net-Investment Strategies
Òscar Jordà and Alan M. Taylor
NBER Working Paper No. 17150
June 2011
JEL No. C14,C59,G14,G17

ABSTRACT

This paper introduces new nonparametric statistical methods to evaluate zero-cost investment strategies. We focus on directional trading strategies, risk-adjusted returns, and the investor's decisions under uncertainty as the core of our analysis. By relying on classification tools with a long tradition in the sciences and biostatistics, we can provide a tighter connection between model-based risk characteristics and the no-arbitrage conditions for market efficiency. Moreover, we extend the methods to multicategorical settings, such as when the investor can sometimes take a neutral position. A variety of inferential procedures are provided, many of which are illustrated with applications to excess equity returns and to currency carry trades.

Òscar Jordà
Department of Economics
University of California, Davis
One Shields Ave.
Davis, CA 95616
ojorda@ucdavis.edu

Alan M. Taylor
Morgan Stanley
1585 Broadway, 38th Floor
New York NY 10036
and NBER
alan.taylor@morganstanley.com

When is an investment opportunity attractive? In the standard set-up with frictionless, complete markets, asset pricing theory says that the representative investor's attitude toward risk, and the desire to insure against variation of consumption across different states of nature, determines the price of each asset. In this ideal world there is no opportunity to arbitrage returns between any two assets (in risk-adjusted terms) if markets are efficient.

Reality is more complex. Investor's preferences are unknown and possibly heterogeneous; markets are incomplete; trading frictions abound; and an asset's risk and returns can only be characterized with finite samples of data. This paper proposes nonparametric statistical methods to evaluate zero-net investment strategies which build on the canonical notion of absence of arbitrage and allow returns to be risk-adjusted. These methods turn out to have a natural connection to the concept of the gain-loss ratio introduced in Bernardo and Ledoit (2000) in the context of asset pricing in incomplete markets, among other features that we now briefly discuss.

Simple directional long-short designs are amongst the most basic trading strategies and represent an important and widely used class of investment rules, the currency carry trade being a classic example (see, e.g. Jordà and A. M. Taylor, 2009). At a theoretical level, even if we cannot forecast returns as judged by the prediction's mean squared error (MSE), the ability to systematically sort the long-short direction might yield significant excess returns and would be sufficient to reject the risk-neutral efficient markets hypothesis. Moreover, 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, 2001). The methods we propose here do not eliminate this dependence on the unknown loss function for the investor, but go a long way to minimize its influence to facilitate comparability.

1 Accuracy, Direction, and 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 endowed with continuously differentiable von Neumann-Morgensten utility indicate that risk-adjusted excess returns should be zero, that is

$$E_{t-1}(m_t x_t) = 0, \tag{1}$$

where m_t denotes the stochastic discount factor given by the investor's intertemporal marginal rate of substitution, and x_t denotes ex-post excess returns (Cochrane, 2001).

An example of x_t is a currency carry trade with $x_t = \Delta e_t + (i^*_{t-1} - i_{t-1})$, and where e_t denotes the logarithm of the (home) exchange rate and $i^*_{t-1} - i_{t-1}$ the one period interest rate differential, 'foreign' minus 'home.' Another example of x_t is an investment where the trader arbitrages the expected returns of a risky asset with the risk-free rate by going short/long in one and taking the opposite position in the other. 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 for an econometrician to predict x_t given information available up to time $t-1$ and this seems to be generally the case. For example, see Kilian and M. P. Taylor (2003) for a survey on currency trades, or the seminal work of Meese and Rogoff (1983); and see Goyal and Welch (2008) for a survey on equities. However, the trader faces a simpler problem, that of determining what to short and what to go long in. Thus, the realized returns of the trading strategy based on $\widehat{x}_t$ are:

$$\widehat{\mu}_t = \text{sign}(\widehat{x}_t)x_t. \tag{2}$$

Notice then that $\widehat{x}_t$ may simultaneously be a poor forecast, as judged by mean squared error, yet consistently determine the profitable direction of trade. This observation is one more example of the interplay between forecasting and decision theory discussed in Granger and Machina (2006) and references therein, that is, that the usefulness of a forecast depends on the rewards associated with the actions taken by the agent as a result of the forecast.

A simple example illustrates this observation further. Table 1 summarizes a hypothetical investment problem, a one-period currency return, say, which has four discrete outcomes: percentage returns of -2, -1, +1, and +2, and each outcome is equiprobable. Two trading signals are 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 just the opposite, it provides accurate predictions of the ±2 outcomes but has the same additive $N(0, 10)$ error as signal A for the ±1 outcomes. Which signal will the investor prefer?

The statistical mean squared error criterion is completely uninformative: both signals attain the same RMSE of 7.071. However, measuring the directional accuracy of each signal may also be misleading since signal A predicts the correct direction of trade 78.96%

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. RMSE denotes root mean squared error of the prediction.

of the time, whereas signal B predicts the correct direction of trade 76.99% of the time, although the difference is small. Yet a risk-neutral investor motivated by returns would clearly find signal B (with a yield of 1.04) much more profitable than signal A (with a yield of 0.66).

The purpose of this numerical example is simple. What matters to investors is not predictive accuracy as usually measured by statisticians by RMSE type measures of fit. What matters is the ability to get direction right, especially when big profits or losses are at stake. If that is the right performance measure, we need to develop tools better suited to the purpose of directional classification, with allowance for variable payoffs.

More generally, denote the *ex post* correct direction of trade as $d_t = sign(x_t) \in \{-1, 1\}$, with -1 denoting a long-side loss (trader should go short), and $+1$ a long-side gain (trader should go long). Let $\widehat{\delta}_t$ denote a *scoring classifier* for d_t such that $\widehat{d_t} = sign(\widehat{\delta}_t - c)$ and $c \in (-\infty, \infty)$ is a threshold parameter. $\widehat{\delta}_t$ can be set just to be $\widehat{x}_t$, or even $\widehat{m}_t \widehat{x}_t$ but more broadly, it can denote a probability forecast from binary regression, a single-index from a dimension reduction procedure, an ordinal variable generated from a discrete state-space model, and so on. All that is required is for $\widehat{\delta}_t$ to be a scalar that takes values in $(-\infty, \infty)$.

In what follows we focus on evaluating the classification ability of $\widehat{\delta}_t$,. We concentrate directly on the trader's classification problem first, and then expand these methods to introduce risk-weighted returns into the statistics that we present.[1] We first explain some advantages of our approach.

[1] We do not focus on the best method to generate $\widehat{\delta}_t$ itself, for which the literature offers numerous alternatives. For example, Elliott and Lieli (2007) provide a formal treatment of how the estimator can be tailored to the agent's specific utility function.

2 What's New

In economics and finance, works by Henriksson and Merton (1981) and Merton (1981), later formalized in Pesaran and Timmermann (1992), provide well known tests of directional accuracy. These tests were later extended to include raw returns by Anatolyev and Gerko (2005). However, the problem of assessing classification in binary-outcome decision problems has a much longer tradition in statistics, and it is within this tradition that we craft some new methods and seek to build a tighter nexus to the economics of the problem.

This older tradition has its origin in the field of signal detection theory (Peterson and Birdsall, 1953) with the introduction of the *receiver operating characteristic* (ROC) curve. ROC-based methods are used extensively to evaluate diagnostic tests from different biomarkers in medicine, as well as to rank radiological readings (see Pepe, 2003 for an extensive monograph; and Zhou, Obuchowski and McClish, 2011 for a monograph on diagnostic medicine). But these methods are also common in other fields of science, such as in psychometrics (see Swets and Pickett, 1982), machine learning (see Spackman, 1989), and atmospheric sciences, where ROC methods form part of the World Meteorological Organization's Standard Verification System for assessing the quality of weather forecasts (see Stanski, Wilson and Burrows, 1989; and World Meteorological Organization, 2000).

The procedures that we examine are closely related to the theory of rank tests (see, e.g. Hájek, Šidák, and Sen 1999) in that they essentially measure the distance between two distributions, that is, the empirical distribution of the forecast signal for the long-short direction versus the distribution for the short-long direction. There are many procedures that measure the distance between two distributions, perhaps the best-known being the Kolmogorov-Smirnov statistic.[2] However, even in the simplest case, in which there are no estimated parameters, no conditioning variables, and the data are i.i.d., the distribution of the statistic is characterized by a function of a Brownian Bridge. The analysis of more general cases requires an advanced toolkit from empirical process theory.[3]

[2] See Kolmogorov (1933) and Smirnov (1939); a more modern reference is Shorack and Wellner (1986).

[3] See Kosorok (2008) and Shorack and Wellner (1986). Some references are available in the econometrics literature. For example, Bai (1994) provides some weak convergence results for sequential empirical processes of residuals in ARMA models; Andrews (1997) introduces conditional Kolmogorov-Smirnov tests for parametric models with regressors; Bai (2003) uses the Khamladze transform (Khamaladze, 1981) to introduce nonparametric tests for parametric conditional distributions of dynamic models. And if one thinks of the problem in terms of stochastic dominance of the empirical distribution for the long-short direction over the distribution for the short-long direction, one could apply Linton, Maasoumi and Whang's (2005) extended Kolmogorov-Smirnov resampling procedures, for example (see also references therein).

Instead, the methods we propose build on the Mann-Whitney rank-sum statistic, whose large-sample distribution has been shown to be Gaussian (Hsieh and Turnbull 1996). Thus, if trading signals come from a conditional model with the usual $\sqrt{T}$ parametric rate of convergence to normality, the large-sample distribution remains Gaussian. Furthermore, Bickel and Freedman (1981) provide justification for the bootstrap, and some of the inferential procedures we discuss are based on the distribution-free permutation approach (see Mielke and Berry, 2007). As a result, inference in our framework is fairly straightforward.

Simplicity of implementation aside, we structure the statistical problem as an economic decision under uncertainty and for this purpose we introduce a variant of the ROC curve that we call the *correct classification* (CC) *frontier.* The virtue of the CC frontier is that it 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 correctly predicting 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. Traditional statistics for the evaluation of binary decision problems (such as log and quadratic probability scores, Brier scores, Kuipers scores, misclassification probabilities, and some of the tests we have discussed above) often lack sufficient texture to frame the problem and are only appropriate as long as the implied loss function coincides with that of the investor. Instead, our analysis of the CC frontier allows one to visualize the regions in which these trade-offs take place.

Forcing the investor to continuously arbitrage the returns between two assets is somewhat restrictive, especially in more realistic situations where the expected returns to the arbitrage are insufficient to cover transactions costs. For this reason, we also investigate more complex multicategorical investment strategies, the simplest of which allows the investor to remain in a neutral "cash" position in addition to taking a long/short position.

This and many of the procedures that we discuss are illustrated with two empirical applications. The first application is based on Goyal and Welch's (2008) state-of-the-art investigation of signals that help 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 Goyal and Welch (2008), whose results are based on the conventional performance metric of mean squared error loss (MSE), which may be relevant for statisticians, but not so much for portfolio managers.

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 A. M. Taylor (2011) 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 these four carry trade investments and find that the more sophisticated methods rank the preferred strategies somewhat differently compared to the case where only binary long/short strategies were considered.

3 The Trader's Classification Problem

The following contingency table summarizes the four conditional probabilities associated with the $\{$*prediction,outcome*$\}$ pair $\{\widehat{\delta}_t, d_t\}$:

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

Here $TN(c)$ and $TP(c)$ refer to the true classification rates of negatives ($d_t = -1$) and positives ($d_t = 1$); and $FN(c)$ and $FP(c)$ refer to the false classification rates of negatives and positives. 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, or $TN(c)$ as the Type II error. The probability measure $P(.)$ can be conceived more broadly as reflecting the risk-neutral measure or even as reflecting risk-adjusted probabilities, as we shall see.

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 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).

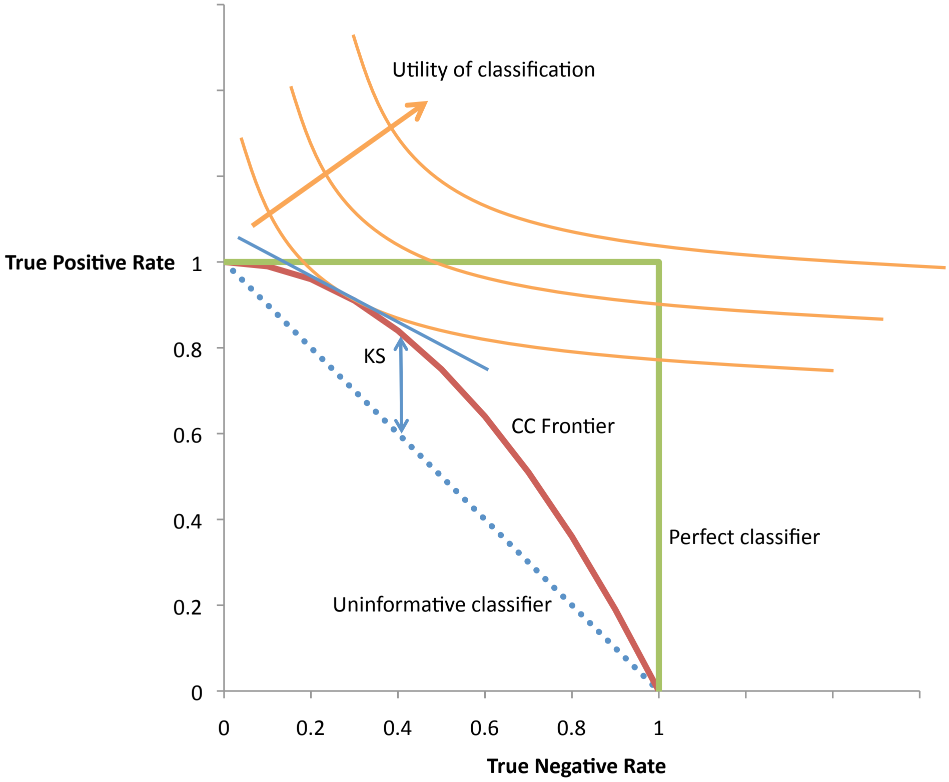
A stylized plot of a CC frontier is presented in Figure 1. Notice that as $c \to -\infty$ then $TP(c) \to 1$ and $TN(c) \to 0$, and the limits are reversed as $c \to \infty$. For this reason, it is easy to see that the CC frontier lives in the unit square $[0, 1] \times [0, 1]$. A perfect classifier 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 "chance" 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.

3.1 The Trader's Decision Problem

The interaction of the CC frontier with the investor's utility over the two choices short/long provides a convenient decision framework with a long tradition that perhaps dates as far back as Peirce's (1884) 'utility of the method.' In modern parlance, Peirce's (1884) concept is akin to expected utility maximization and can be formulated in terms of the optimal threshold parameter c as follows:

$$\begin{aligned} U(c) &= U_{pP}TP(c)\pi + U_{nP}(1 - TP(c))\pi + \\ &\quad U_{pN}(1 - TN(c))(1 - \pi) + U_{nN}TN(c)(1 - \pi). \end{aligned} \tag{3}$$

Figure 1: The Correct Classification Frontier



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. Notice that since $TP(c) + FN(c) = 1 = TN(c) + FP(c)$, everything can be expressed in terms of the true classification rate alone.

It is instructive to discuss a few special cases first to gain further intuition. For a risk-neutral investor, it would be natural to normalize his utility symmetrically such that $U_{pP} = U_{nN} = 1$ and $U_{nP} = U_{pN} = -1$. Further, if returns are themselves symmetric as well (i.e. $\pi = 1/2$), then the investor's utility is maximized where the marginal rate of substitution between true positives and true negatives is -1.

It turns out that the vertical distance between this maximum point on the CC frontier and the chance diagonal is the same distance that forms the basis of the well-known

Kolmogorov-Smirnov (KS) statistic:

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

that is, the maximum difference between the average correct classification rate and the average rate of an uninformative classifier for which $TP(c) = 1 - TN(c)\ \forall c$, scaled by 2 so that $KS \in [0, 1]$. We make a brief detour to discuss some of the salient features of this special case before we introduce the methods that we propose.

Intuitively, the KS statistic measures the uniform distance between the empirical distribution of $\widehat{\delta}$ when $d = -1$ and the empirical distribution of $\widehat{\delta}$ when $d = 1$. For this reason, the KS statistic is useful in determining the stochastic dominance features of two investment strategies (see, e.g., Linton, Maasoumi and Whang, 2005 and references therein).

In a sample with T observations, let T_N and T_P indicate the number of observations for which $d = -1, 1$ respectively, using the mnemonics N for negative and P for positive. For a given threshold c, the finite sample correct classification rates can be calculated nonparametrically as:

$$\widehat{TP}(c) = \frac{\sum_{i=1}^{T_P} I\left(\widehat{\delta}_i > c\right)}{T_P}; \qquad \widehat{TN}(c) = \frac{\sum_{j=1}^{T_N} I\left(\widehat{\delta}_j < c\right)}{T_N} \tag{4}$$

where the indices i, j relabel the original t subscript to keep the notation simple, and run over two sets of re-numbered observations, with each one mapping to a unique t such that $d_t = 1(-1)$, respectively and $I(.) \in \{0, 1\}$ is the indicator function that takes on the value of 1 when the argument is true and 0 otherwise. When $\widehat{\delta}$ is *i.i.d.* and is not generated by a fitted model with estimated parameters (such as the 'momentum' strategy in one of our carry trade examples below), then:

$$\sqrt{\frac{T_N T_P}{T}} \widehat{KS} \rightarrow \sup_t |B(t)| \tag{5}$$

where $B(t)$ is a Brownian-bridge, that is, $B(t) = W(t) - tW(1)$ where $W(t)$ a Wiener process (see e.g. Shorack and Wellner, 1986). Notice that $KS \in [0, 1]$ and is equivalent to the maximum of the Youden (1950) J index, which is defined as

$$J(c) = TP(c) - FP(c). \tag{6}$$

Under the symmetry assumption that $P(d = -1) = P(d = 1) = \pi = \frac{1}{2}$, a risk-neutral investor will want to maximize the J index out of which one can identify the optimal operating point as the threshold c_{KS} where $KS = J(c_{KS})$.

Several practicalities deserve comment. Investment strategies are often the result of models with estimated parameters, in which case, the distribution of the KS statistic has to be recalculated as shown by Rubin (1973). See also Andrews (1997) for conditional KS asymptotic results which require the bootstrap. Linton, Maasoumi and Whang (2005) provide tests of stochastic dominance based on the KS statistic but these require resampling procedures. Bai (2003) probably contains the most relevant results for tests of parametric conditional distributions of dynamic models derived from empirical process theory using the Khamaladze (1981) martingale transformation. However, the test's critical values have to be obtained by simulation in each case.

If instead one uses utility weights of 1 for correct calls and 0 for incorrect calls, then expression (3) is just the *accuracy rate:*

$$A(c) = TP(c)\pi - TN(c)(1 - \pi)$$

whereas exchanging these weights, we obtain the *error rate:*

$$E(c) = FN(c)\pi + FP(c)(1 - \pi).$$

But the accuracy and error rates sum to one, so these are inversely related and attain their respective maximum and minimum for 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 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, expression (3) simplifies in such a way that $U(c) = J(c)$, and

$$A(c) = \frac{1 + J(c)}{2}; \qquad E(c) = \frac{1 - J(c)}{2}.$$

Thus, for this case only, all performance measures are monotonic in $J(c)$, 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 previously). 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 (5) insufficiently characterizes an investor's choices—in other words, the simplifying assumptions used to derive expression (5) may not hold in practice.

To sidestep this problem, the CC frontier allows us to make comparisons among classifiers without specific assumptions about underlying preferences by considering all operating points simultaneously. Given the CC frontier's usefulness, it is helpful to develop some further intuition about its shape and the properties of 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 (3) is achieved when

$$\frac{dCC(r)}{dr} \equiv \frac{g(F^{-1}(r))}{f(F^{-1}(r))} = -\frac{1-\pi}{\pi}\frac{(U_{nN} - U_{pN})}{(U_{pP} - U_{nP})} \tag{7}$$

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 nonparametric 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. On 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.

4 Building Blocks: The Area under the CC Frontier

In this section we introduce an alternative statistic to the KS that will form the basis for the procedures that we discuss below. Several advantages will become apparent throughout the discussion, among them, more convenient asymptotic properties and a tighter link to the investor's decision problem.

From Figure 1 it is clear that an arbitrage opportunity occurs for a perfect classifier. In that case the CC frontier in Figure 1 is given by the north-east edges of the unit-square and the tangent to the investor's utility takes place at the (1,1) corner so that the investor's specific preferences become irrelevant. The area under this CC frontier, which we denote AUC for *area under the curve*, is the area in the unit-square and therefore $AUC = 1$. Conversely, an uninformative classifier given by the chance diagonal that bisects the unit-square from (1,0) to (0,1) has $AUC = 0.5$.

In practice, the AUC for most investment strategies will be between these two values. Values below 0.5 are possible in small samples but reversing the predictions would generate an $AUC \geq 0.5$.

Formally, we have the definition

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

We 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).

Given two investment strategies, A and B such that $CC_A(r) > CC_B(r)$ $\forall r$ then A stochastically dominates B regardless of investor preferences and it will be the case that $KS_A > KS_B$ and $AUC_A > AUC_B$. However, the reverse is not a sufficient condition for stochastic dominance since even if $KS_A > KS_B$ and/or $AUC_A > AUC_B$, there could be a crossing between the two CC frontiers and hence regions where $CC_A(r) < CC_B(r)$. In such a case, whether A is preferred to B or vice versa depends on what region are investor preferences tangent to the CC frontier. We will discuss momentarily several inferential procedures, including tests of the null: $H_0 : CC_A(r) = CC_B(r)$.

Before then though, we discuss several properties of the AUC. Green and Swets (1966) provide a nice interpretation of the AUC as $AUC = P[v < u]$ where v and u where defined in the previous section as the $\widehat{\delta}$ for which $d = -1, +1$ respectively. Thus, like the KS statistic, the AUC measures the distance between the empirical distribution for v and u, except that instead of a uniform norm, we use:

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

Not surprisingly, the AUC is related to the Wilcoxon-Mann-Whitney U-statistic (see Bamber 1975; Hanley and McNeil 1982), which is a rank-sum statistic. The AUC can be estimated nonparametrically and in the simplest form can be obtained as

$$\widehat{AUC} = \frac{1}{T_N T_P} \sum_{j=1}^{T_N} \sum_{i=1}^{T_P} \left\{ I\left(v_j < u_i\right) + \frac{1}{2} I\left(u_i = v_j\right) \right\} \tag{8}$$

where the last term is used to break ties. However, we should point out that there are alternative estimators based on making parametric assumptions for f and g, and other kernel-weighted nonparametric estimators all of which are discussed, e.g., in Pepe (2003).

Here we prefer to retain the simpler estimator presented in (8) for reasons that will become clear momentarily.

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:

$$\sqrt{T}\left(\widehat{AUC} - P\left[v < u\right]\right) \to N\left(0, \sigma^2\right), \tag{9}$$

and in the special case that $F = G$ (which is the null case in which there is no classification ability), then $P[v < u] = 0.5$ and:[4]

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

We make one final comment on the effect of estimated parameters on the large sample results available for the AUC. The intuition is fairly straightforward to communicate: if the estimated parameters converge in distribution to a normal at the same $\sqrt{T}$ rate as the AUC, the asymptotic result in expression (8) remains valid except that the variance of the AUC would need to reflect parameter estimation uncertainty as well—an issue that can be easily resolved in practice by using the bootstrap.[5]

[4] The proof of this result can also be derived from the theory of empirical processes using lemma 12.3 and theorem 2.8 in Kosorok (2008), who also shows in theorem 12.1 that the delta method bootstrap is applicable and can be used under more general hypotheses on F and G. The result hinges on $CC(r)$ being a Hadamard differentiable map, a convenient assumption which cannot be verified in practice.

Hanley and McNeil (1982) and Obuchowski (1994) provide a convenient approximation for the variance of AUC under general conditions using the Mann-Whitney interpretation of the AUC. However, DeLong, DeLong and Clarke-Pearson (1988) provide a more general formula that is available in SAS and STATA and allows for correlated AUCs. Jackknife procedures are available in Hanley and Hajian-Tilaki (1997) and Obuchowski and Lieber (1998) provide standard bootstrap results under a variety of assumptions, although large sample approximations have been found to do well in relatively small samples (Pepe 2003).

[5] Early proofs can be found in Darling (1955) and Durbin (1973), although Bai (2003) contains a more modern treatment, and in particular, assumptions A1-A4 in theorem 1. Without repeating the assumptions here, one needs the densities f and g and the score functions to be continuously differentiable with respect to the parameter vector. The parameter estimates are assumed to converge in distribution at rate $\sqrt{T}$. Bai's fourth assumption ensures that, for dynamic models, past information becomes less relevant as time progresses. For another treatment of the same problem the reader is referred to theorem 19.23 in van der Vaart (1998).

5 Inferential Procedures with AUCs

This section discusses four inferential procedures: (1) tests of the null that an investment strategy picks profitable positions no better than random luck; (2) tests of he null that two strategies have the same AUC (an overall comparison); (3) tests of the null that two strategies have the same CC frontier at all operating points; and (4) confidence intervals for the CC frontier at specific operating points.

To test against the random null, the large-sample results in the previous section provide the sampling distribution of $\widehat{AUC}$, which can be used to construct Wald tests for the null that $AUC = \frac{1}{2}$. If the strategy is based on a model with estimated parameters, then the bootstrap percentiles, bootstrap-z or the bootstrap bias-corrected accelerated method discussed in Obuchowski and Lieber (1998) should be used.

For comparing the value of two AUCs, Hanley and McNeil (1983) suggest that a test of $H_0 : AUC_A = AUC_B$ can be constructed using the z-ratio

$$z = \frac{AUC_A - AUC_B}{\sqrt{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}} \rightarrow N(0,1), \tag{10}$$

where σ_j^2 for $j \in \{A, B\}$ refers to the variance of the AUC 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, as discussed in the previous section, which investment strategy an investor will prefer cannot be assessed with this type of null, nor can one test whether a strategy dominates another, unless $\widehat{CC}_A(r) > \widehat{CC}_B(r)\ \forall r$ and $H_0 : AUC_A = AUC_B$ can be rejected. A more appropriate hypothesis is $H_0 : CC_A(r) = CC_B(r)\ \forall r$. Venkatraman and Begg (1996) provide a test of this null using distribution-free permutation methods from which p-values can be easily constructed by simulation, as follows.

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

$$E = \sum_{k=1}^{T-1} \left| \sum_{i=1}^{T} e_{ik} \right| . \tag{11}$$

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 (11) 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.

Finally, one may be interested in providing error bands for the CC frontier at a given operating point (say the point that is tangent to the investor's preferences). In principle, such an interval can be constructed from the large sample results in Hsieh and Turnbull (1996) as follows. For a given operating point r and a confidence level $1 - \alpha$, then

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

with:[6]

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

where one can substitute for all theoretical quantities with their empirical counterparts. However, in more general situations than those contemplated in the assumptions in Hsieh and Turnbull (1996) one would have to rely on straightforward bootstrap methods.

[6] 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 $o(T^{-2/3})$.

6 Adjusting for Returns and Risk

As we have stressed in our discussion so far, correct classification alone is insufficient to assess the appeal of an investment strategy: correctly picking 99-in-100 penny trades while missing the 1-in-100 dollar trade results in a loss although the strategy has almost perfect directional accuracy.[7] And when there are transactions costs, it is reasonable to expect that a trader may choose to sit on the sidelines when the expected returns of a position are insufficient to cover transactions costs. In this section we show how such adjustments can be made to the basic framework introduced in previous sections.

6.1 The Risk-and-Return-Adjusted CC Frontier: $CC^\star$

Risk-and-return-adjusted probability measures incorporate information about investor preferences for consumption in different states. It would be natural to characterize the relative trade-offs embodied in the CC frontier expressed in terms of risk-and-return-adjusted probabilities. This can be done easily as follows. Define:

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

where m_t is the pricing kernel derived for a benchmark risk-averse investor. Under risk neutrality, $m_t = 1$ and B then becomes the in-sample gains of all the long positions and C the in-sample gains of all the short positions. Here we assume that the positions are normalized to \$1 for convenience and therefore, we do not take into account how the benchmark investor's endowment may or may not affect his aversion to risk. B and C can then be used to construct the following weights:

$$\begin{array}{llll} w_i = \frac{m_i x_i}{B} & \text{if} \quad \widehat{\delta}_i > c \text{ and } d_i = 1 & \text{for } i = 1, ..., T_P, \\ w_j = \frac{|m_j x_j|}{C} & \text{if} \quad \widehat{\delta}_j < c \text{ and } d_j = -1 & \text{for } j = 1, ..., T_N, \end{array}$$

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, the risk-and-return-adjusted equivalent to (4) can be calculated

[7] For example, see Anatolyev and Gerko (2005), who build on Pesaran and Timmermann (1992).

as:

$$\widehat{TN^{\star}}(c) = \sum_{j=1}^{T_N} w_j I(\widehat{\delta}_j < c), \quad \widehat{TP^{\star}}(c) = \sum_{i=1}^{T_P} w_i I(\widehat{\delta}_i > c), \tag{12}$$

so that the risk-and-return-adjusted CC* frontier, $CC^*(r)$, is a map of all the combinations $\{TN^*(c), TP^*(c)\}$ for $c \in (-\infty, \infty)$. By construction, $CC^*(r)$ still inhabits the unit-square, with a CC* frontier hugging the north-east corner of the unit-square now representing a perfect arbitrage opportunity and the 'chance' diagonal representing risk-and-return-adjusted complete absence of arbitrage. Furthermore, now the slope of $CC^*(r)$ represents the likelihood ratio of the risk-and-return-adjusted densities f^* and g^* (the risk-and-return-adjusted equivalents to f and g introduced earlier), weighted by returns.

Notice that the expected gains, $\Gamma(c)$, and losses, $\Lambda(c)$, under the risk-and-return-adjusted probabilities and for a given threshold parameter c are:

$$\begin{aligned} E^*\left(\Gamma(c)\right) &= B.TP^*(c) + C.TN^*(c) \\ E^*(\Lambda(c)) &= B.(1 - TP^*(c)) + C.(1 - TN^*(c)), \end{aligned} \tag{13}$$

where E^* denotes the expectation under the risk-and-return-measure. Hence, the investor's expected profit ratio (that is, where we normalize expected profits by the maximally attainable profits in the sample for cross comparability) can be expressed as:

$$E^*(\Pi(c)) = \frac{E^*(\Gamma(c) - \Lambda(c))}{B + C} = \frac{B\left(2TP^*(c) - 1\right) + C(2TN^*(c) - 1))}{B + C}.$$

In the special case where $B = C$, then it is easy to see that

$$E^*(\Pi(c)) = 2 \times J^*(c)$$

where $J^*(c)$ is the risk-and-return-adjusted Youden (1950) index. Hence, profit maximization is achieved for $c = c^*_{KS}$ since

$$\max_c 2 \times J^*(c) = 2 \times \max_c |TP^*(c) + TN^*(c) - 1|,$$

the latter being twice the KS^* statistic using risk-and-return-adjusted probabilities. In this special case ($B = C$), Bernardo and Ledoit's (2000) gain-loss ratio can be simply calculated as:

$$\alpha \equiv \frac{\max_c E^*(\Gamma(c))}{\max_c E^*(\Lambda(c))} = \frac{1 + KS^*}{1 - KS^*}.$$

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.[8]

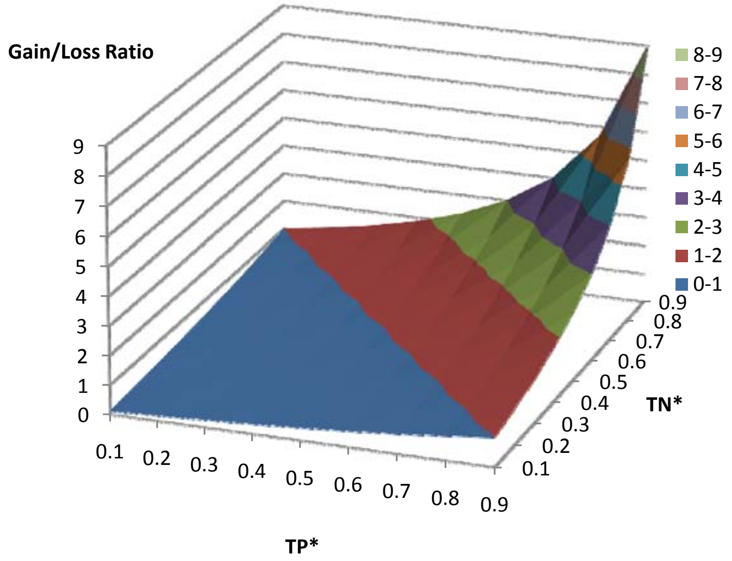
We pause our discussion momentarily to elaborate on this last equivalence to the Bernardo and Ledoit (2000) gain-loss ratio and its implications for asset pricing. First notice that, in the above equation, $\alpha \to \infty$ is an arbitrage opportunity: at least some gains can be made without the possibility of any loss. This will happen when $KS^* = 1$, the maximum value the KS^* can attain. Conversely, when $KS^* = 0$ (the minimum possible value), then expected gains and losses (under the risk-and-return-adjusted measure) are the same and $\alpha = 1$. Values in-between these two extremes are *approximate* arbitrage opportunities. Figure 2 shows the value of α in $\{TP^*(c), TN^*(c)\}$ to make the connection with CC^* space.

Moreover, theorem 1 in Bernardo and Ledoit (2000) provides an alternative characterization of all pricing kernels that correctly price all portfolio pay-offs and hence is an alternative to, for example, the duality result in Hansen and Jaganathan (1991), which show that a bound on the variance of the pricing kernel is equivalent to a bound on the maximum Sharpe ratio. Bernardo and Ledoit (2000) argue that, unlike the Sharpe ratio, α characterizes the set of arbitrage ($\alpha \to \infty$) and approximate arbitrage opportunities ($\alpha > 1$) and propose that for a given bound $1 < \overline{\alpha} < \infty$, the benchmark model will be reasonable if $\alpha \leq \overline{\alpha}$, but otherwise the model will be inconsistent with well functioning capital markets. Therefore $\overline{\alpha}$ controls the trade-off between the precision of a specific benchmark pricing model and the robustness of no-arbitrage bounds. As an example, Bawa and Lindeberg (1977) showed that if returns are Gaussian, then there is a one-to-one mapping with the Sharpe ratio under risk-neutrality so that, for example, a Sharpe ratio of 0.3 is equivalent to $\alpha = 2$. The reader is referred to Bernardo and Ledoit (2000) for more details.

In practice, several features of the data can disturb this nice connection between KS^* and α, such as $B \neq C$, asymmetry of returns, and so on. And in reality, the investor's

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

Figure 2: The Gain-Loss Ratio in $\{TP^{*}(c), TN^{*}(c)\}$ Space



benchmark preferences are unknown. For these reasons and along the lines of the arguments we used in previous sections, it is natural to compute an alternative to the KS^{*} in the spirit of the AUC, namely,

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

As long as the risk-and-return-adjusted probabilities conform with standard measurability conditions, nothing has changed with respect to the statistical properties previously discussed for $\widehat{AUC}$: the weighting simply re-ranks the observations within each empirical distribution, f^{*} or g^{*}, but no observations switch from one to the other as a result of the weighting. Thus, the asymptotic results only require that the regularity conditions now apply to f^{*} and g^{*}. As a result, we will now require that $B/C \to \beta > 0$ as $T \to \infty$ and that the densities of the returns in long/short positions be continuous so that the resulting convolution of f and g into f^{*} and g^{*} is also continuous. Under this scenario, the inferential and bootstrap procedures discussed in previous sections can be easily modified. It is useful to refer back to Figures 1 and 2 since the CC^{*} can now be seen as giving a map to α in Figure 2, but under more general conditions.

6.2 Multicategorical Positions

The excess returns of a continuously traded zero-cost strategy offer a natural benchmark with which to assess the attractiveness of an investment opportunity. In reality, however, investors face transactions costs or they may wish to place larger bets when risk-adjusted returns are expected to be specially high in a given period. In the first example, one would expect investors to take a neutral position if expected returns are insufficient to cover transactions costs, thus expanding the decision space to long/neutral/short positions. In the second example, an investor may consider doubling the bet when the expected returns of a position are specially high and hence the decision space would expand to four categories, say long(2)/long(1)/short(1)/short(2), or even five if one includes a neutral category. In what follows we show how to extend the analysis of the previous sections using the three category long/neutral/short example as our framework, although extensions to more categories should be obvious from our discussion (see also Waegeman, De Baets and Boullart, 2008).

Suppose that whether an investor goes short/neutral/long, that is, $d = -1, 0, 1$, is determined by

$$\begin{array}{lll} d_t = -1 & if & x_t < \psi_1, \\ d_t = 0 & if & \psi_1 \leq x_t \leq \psi_2, \\ d_t = +1 & if & x_t > \psi_2, \end{array} \tag{14}$$

where $\boldsymbol{\psi} = (\psi_1, \psi_2)$ are pre-determined thresholds known in advance. For example, if x_t represents excess returns, then ψ_1 could represent the costs of a short position and ψ_2 the costs of a long position so that if $\psi_1 < x_t < \psi_2$, the trader will choose to remain neutral.

Next, assume there is a model that generates a continuous signal $\widehat{\delta}_t$ that determines the investor's positions according to:

$$\begin{array}{lll} \widehat{d}_t = -1 & if & \widehat{\delta}_t < c_1, \\ \widehat{d}_t = 0 & if & c_1 \leq \widehat{\delta}_t \leq c_2, \\ \widehat{d}_t = +1 & if & c_2 < \widehat{\delta}_t, \end{array}$$

for $c_1 < c_2$; $c_1, c_2 \in (-\infty, \infty)$ with $\mathbf{c} \equiv (c_1, c_2)$.

The contingency matrix associated with each of the three states $d = -1, 0, +1$ means that the CC frontier is now a surface in the three-dimensional space of the per category true positive rates:

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

where T_h refers to the number of observations in the sample that belong to category $h \in \{N, 0, P\}$ with N standing for *negatives, 0* standing for *neutrals*, and P standing for positives. The CC surface belongs in the $[0,1] \times [0,1] \times [0,1]$ cube and as with the CC frontier, perfect classification implies a CC surface with unit volume, whereas the 'chance' simplex (rather than diagonal) bisects the cube at $(1,0,0)$, $(0,1,0)$, $(0,0,1)$.

The equivalent statistic to the AUC in three dimensions becomes the volume under the CC surface or VUS (for volume under the surface). A perfect classifier is one with $VUS = 1$, but now a chance classifier has $VUS = 1/6$ because the probability of randomly classifying any two categories, while still $1/2$, has to be considered for any three possible pairs. In general, classification problems with k categories result in a CC hypersurface in k dimensions and a VUS for the chance classifier of $1/k!$.

Although there is no comparable extension for the KS statistic, the VUS is a direct generalization of the Wilcoxon-Mann-Whitney statistic (see Mossman, 1999). Let v_j denote the observations of $\widehat{\delta}$ for which $d = 1$; z_k when $d = 0$; and u_i when $d = 1$, then the empirical estimate of $P[v < z < u]$ is readily seen to be

$$\widehat{VUS} = \widehat{P}[v < z < u] = \frac{1}{T_N T_0 T_P} \sum_{j=1}^{T_N} \sum_{k=1}^{T_0} \sum_{i=1}^{T_P} \{ I(v_j < z_k < u_i), \} \tag{15}$$

where we omit the rule to randomly break ties in the interest of keeping the notation concise.

Dreiseitl, Ohno-Machado and Binder (2000) provide Gaussian large sample approximations for $\widehat{VUS}$ and analytic formulas for the variance, as well as formulas for the covariance between the $\widehat{VUS}$ of two competing models. These can be used to craft inferential procedures along the lines described in the previous sections. More generally, concerns with the small sample properties of asymptotic approximations and/or models with $O(T^{1/2})$ estimated parameters, are good reasons to rely on bootstrap procedures instead.

Consider now VUS calculations using risk-and-return-adjusted probabilities. For this derivation, notice that in the neutral position, $x_t = 0$ and hence all observations within this category receive equal weight of $1/T_0$. Recall that we previously defined:

$$B = \sum_{d=1} m_t x_t; \text{ and } C = \sum_{d=-1} |m_t x_t|$$

and hence we can define the following weights:

$$\begin{array}{ll} w_i = \frac{m_i x_i}{B} & \text{if } \widehat{\delta} > c_2 \text{ and } d = 1 \text{ for } i = 1, ..., T_P \\ w_k = \frac{1}{T_0} & \text{if } c_1 \leq \widehat{\delta} \leq c_2 \text{ and } d = 0 \text{ for } k = 1, ..., T_0 \\ w_j = \frac{|m_j x_j|}{C} & \text{if } \widehat{\delta} < c_1 \text{ and } d = -1 \text{ for } j = 1, ..., T_N \end{array}$$

where the indices i, j, k are a shorthand device to denote the observations t belonging to each category.

The empirical counterpart to the AUC^* for three categories now becomes:

$$VUS^\star = \sum_{i=1}^{T_P} w_i \sum_{k=1}^{T_0} w_k \sum_{j=1}^{T_N} w_j I(v_j < z_k < u_i).$$

Notice now that expected gains, $\Gamma(\mathbf{c})$, and losses, $\Lambda(\mathbf{c})$, under the risk-and-return-adjusted probabilities and for a given vector of thresholds $\mathbf{c} = (c_1, c_2)$, $c_1 < c_2$ is similar to those calculated in expression (13), namely:

$$\begin{array}{rcl} E^*(\Gamma(\mathbf{c})) & = & B.TP_P^*(\mathbf{c}) + C.TP_N(\mathbf{c}) \\ E^*(\Lambda(\mathbf{c})) & = & B(1 - TP_P(\mathbf{c})) + C(1 - TP_N(\mathbf{c})) \end{array}$$

since in our example the neutral position offers no returns ($x_t = 0$). Similarly, the investor's normalized expected profits are:

$$E^*(\Pi(\mathbf{c})) = \frac{B\left(2TP_P^*(\mathbf{c}) - 1\right) + C\left(2TP_N(\mathbf{c}) - 1\right)}{B + C}; c_1 < c_2.$$

In the special case $B = C$ then $E^*(\Pi(\mathbf{c})) = 2 \times J^*(\mathbf{c})$ but where now $J^*(\mathbf{c})$ is a version of the Youden index constructed with $TP_P^*(\mathbf{c})$ and $TP_N(\mathbf{c})$.

However, because now there is a neutral position, choosing e.g., $c_1 \to -\infty$ so that $TP_N(\mathbf{c}) \to 0$ does not now imply that $TP_P(\mathbf{c}) \to 1$ since one could maintain c_2 fixed. In fact, if $c_1 \to -\infty$ and $c_2 \to \infty$, then $TP_N(\mathbf{c})$, $TN_P(\mathbf{c}) \to 0$ and $TP_0(\mathbf{c}) \to 1$.

Therefore the profit maximization problem is a little different and the maximum can be calculated from the simplified problem

$$\max_{c_1,c_2} 2\left(TP_N^*(c_1,c_2)+TP_P^*(c_1,c_2)-1\right)$$
$$s.t.\ c_1 < c_2.$$

In contrast to the binary case, there are three empirical distributions resulting from the values of $\widehat{\delta}$ assigned to $d=-1,0,1$, say f^*, h^*, and g^* respectively. For profit maximization the key is to determine the distance between f^* and g^* as before, but now there is the intervening distribution h^*. Although the connection between Bernardo and Ledoit's (2000) gain-loss ratio α and the KS^* statistic breaks down, the intuition for the result does not: the distance between f^* and g^* is still the critical element in calculating α, for which $\widehat{VUS}^*$ offers a useful measure.

7 Empirical Application I: Risk-and-Return Adjusted Excess Returns of Equities

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 Goyal and Welch (2008).[9] 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 a long-short trading strategy for U.S. equities for 1927:1 to 2008:12. The monthly excess return is 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 any one of the following 15 indicators used individually:[10] (1) the dividend price ratio, *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

[9] 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.

[10] 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`

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; (14) the inflation rate, *linfl*, based on the CPI and lagged one month to allow for publication lags; and (15) *tbl* the short-term return on the 6-to-3 month T-Bill rate.

These signals are used for IS prediction over the full period, and OOS prediction using a long window from 1970:1 to 2008:1.[11] 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 mean squared error (MSE) 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 Goyal and Welch (2008, section 5) found a handful of strategies whose IS predictive power surmounted conventional significance tests. Under their MSPE 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 significance. A few more signals were found to be promising when various truncations were applied to the data.

[11] 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.

In addition to IS and OOS statistical inference based on the MSPE criteria, Goyal and Welch (2008) also consider the profitability of the candidate strategies by constructing a certainty-equivalent gain after postulating a utility function (Brennan and Xia 2005; 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* (14 bps/month), *eqis* (14 bps/month), *tbl* (10 bps/month), *csp* (6 bps/month), *cay3* (6 bps/month), and *ntis* (2 bps/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 using the CC-based evaluation tools: the AUC and $AUC^\star$ using a risk-neutral returns adjustment only (in Table 4 below we examine stochastic discount factor adjustments with coefficients of risk aversion up to 80 for power utility). We also report the KS and $KS^\star$ as well as the Bernardo and Ledoit (2000) gain-loss ratio based on $KS^\star$ and under the assumption of symmetry. According to the directional AUC test, only two signals surmount the conventional 5% significance level: *csp* and *linfl,* although *tbl, lty,* and *tms* come very close (the lower bound of the 95% confidence interval calculated with 1,000 bootstrap replications is just at the null value of 0.50). Of these only *csp* was found to be significant IS in the Goyal-Welch findings. However, when we turn to the $AUC^\star$ test based on profits, the picture is a bit different. First, the $AUC^\star$ is noisier and this is reflected in 95% confidence intervals that are slightly wider than with the raw AUC. Now only two signals do well by this yardstick: *ep* and *csp* of which the first was not significant without the adjustment. Only *csp* performs well in both cases.

Yet, as we have noted, in-sample performance alone will not convince an appropriately skeptical reader. We therefore repeat our exercise and compute the OOS performance of the signals, as shown in Table 3. Using the directional AUC test four signals are significant at the 5% level, namely: *tbl, lty, tms* and *csp.* Note that only one of these was statistically significant in the IS tests using AUC, namely *csp.* Turning to the returns-based $AUC^\star$ test, only one signal is significant at the 5% level: *csp*, which is the only signal to achieve statistical significance in all four of our tests.

Table 2: AUC, KS, and Gain-Loss Ratio, and Risk-Neutral $AUC^{\star}$, $KS^{\star}$, and Gain-Loss Ratio for Equity Strategies. In-sample Predictions: 1927:1–2008:12

Signal	**Statistic**	**Value**	**95% Conf. Interval**		**Statistic**	**Value**	**95% Conf. Interval**	
dp	AUC	0.51	0.48	0.55	AUC*	0.54	0.48	0.60
	KS	0.05	0.01	0.10	KS*	0.08	0.02	0.15
	g/l	1.12	1.02	1.22	g/l*	1.19	1.03	1.36
dy	AUC	0.51	0.48	0.55	AUC*	0.54	0.48	0.60
	KS	0.06	0.02	0.10	KS*	0.09	0.02	0.17
	g/l	1.12	1.03	1.23	g/l*	1.20	1.04	1.40
ep	AUC	0.52	0.48	0.56	AUC*	0.58	0.53	0.63
	KS	0.05	0.01	0.09	KS*	0.14	0.07	0.22
	g/l	1.11	1.03	1.19	g/l*	1.34	1.14	1.57
de	AUC	0.50	0.46	0.53	AUC*	0.53	0.47	0.58
	KS	0.07	0.02	0.11	KS*	0.08	0.00	0.16
	g/l	1.14	1.05	1.24	g/l*	1.18	1.00	1.39
svar	AUC	0.51	0.47	0.54	AUC*	0.54	0.48	0.59
	KS	0.03	-0.01	0.07	KS*	0.08	0.00	0.15
	g/l	1.05	0.97	1.14	g/l*	1.17	1.00	1.37
bm	AUC	0.49	0.46	0.53	AUC*	0.54	0.49	0.60
	KS	0.03	0.00	0.06	KS*	0.09	0.02	0.17
	g/l	1.06	0.99	1.13	g/l*	1.20	1.03	1.40
ntis	AUC	0.52	0.49	0.56	AUC*	0.54	0.48	0.60
	KS	0.09	0.03	0.15	KS*	0.11	0.03	0.18
	g/l	1.20	1.07	1.35	g/l*	1.24	1.07	1.45
tbl	AUC	0.53	0.50	0.57	AUC*	0.52	0.46	0.57
	KS	0.08	0.03	0.13	KS*	0.07	0.00	0.14
	g/l	1.17	1.07	1.29	g/l*	1.15	1.00	1.32
lty	AUC	0.54	0.50	0.57	AUC*	0.53	0.48	0.58
	KS	0.08	0.03	0.13	KS*	0.09	0.02	0.16
	g/l	1.18	1.06	1.31	g/l*	1.20	1.04	1.39
ltr	AUC	0.51	0.47	0.54	AUC*	0.53	0.48	0.58
	KS	0.06	0.00	0.11	KS*	0.06	0.00	0.12
	g/l	1.12	1.00	1.25	g/l*	1.12	1.00	1.26
tms	AUC	0.54	0.50	0.57	AUC*	0.53	0.48	0.58
	KS	0.08	0.03	0.13	KS*	0.10	0.03	0.18
	g/l	1.17	1.05	1.30	g/l*	1.23	1.06	1.43
dfy	AUC	0.47	0.44	0.51	AUC*	0.47	0.42	0.53
	KS	0.01	-0.02	0.04	KS*	0.05	-0.01	0.11
	g/l	1.03	0.97	1.09	g/l*	1.10	0.98	1.24
dfr	AUC	0.52	0.48	0.56	AUC*	0.53	0.48	0.59
	KS	0.07	0.02	0.13	KS*	0.09	0.01	0.17
	g/l	1.16	1.04	1.29	g/l*	1.20	1.03	1.40
linfl	AUC	0.55	0.52	0.59	AUC*	0.52	0.46	0.58
	KS	0.10	0.04	0.15	KS*	0.06	-0.01	0.14
	g/l	1.22	1.09	1.36	g/l*	1.14	0.98	1.32
csp	**AUC**	**0.55**	**0.51**	**0.59**	**AUC***	**0.57**	**0.52**	**0.63**
	KS	**0.08**	**0.03**	**0.14**	**KS***	**0.12**	**0.06**	**0.18**
	g/l	**1.18**	**1.06**	**1.31**	**g/l***	**1.27**	**1.12**	**1.44**

Notes: 95% confidence interval calculated with 1,000 bootstrap replications.

Table 3: AUC, KS, and Gain-Loss Ratio, and Risk-Neutral $AUC^{\star}$, $KS^{\star}$, and Gain-Loss Ratio for Equity Strategies. Out-of-sample Predictions: 1970:1–2008:12

Signal	Statistic	Value	95% Conf. Interval		Statistic	Value	95% Conf. Interval	
dp	AUC	0.51	0.48	0.55	AUC*	0.54	0.48	0.60
	KS	0.05	0.01	0.10	KS*	0.08	0.02	0.15
	g/l	1.12	1.02	1.22	g/l*	1.19	1.03	1.36
dy	AUC	0.51	0.48	0.55	AUC*	0.54	0.48	0.60
	KS	0.06	0.02	0.10	KS*	0.09	0.02	0.17
	g/l	1.12	1.03	1.23	g/l*	1.20	1.04	1.40
ep	AUC	0.52	0.48	0.56	AUC*	0.58	0.53	0.63
	KS	0.05	0.01	0.09	KS*	0.14	0.07	0.22
	g/l	1.11	1.03	1.19	g/l*	1.34	1.14	1.57
de	AUC	0.50	0.46	0.53	AUC*	0.53	0.47	0.58
	KS	0.07	0.02	0.11	KS*	0.08	0.00	0.16
	g/l	1.14	1.05	1.24	g/l*	1.18	1.00	1.39
svar	AUC	0.51	0.47	0.54	AUC*	0.54	0.48	0.59
	KS	0.03	-0.01	0.07	KS*	0.08	0.00	0.15
	g/l	1.05	0.97	1.14	g/l*	1.17	1.00	1.37
bm	AUC	0.49	0.46	0.53	AUC*	0.54	0.49	0.60
	KS	0.03	0.00	0.06	KS*	0.09	0.02	0.17
	g/l	1.06	0.99	1.13	g/l*	1.20	1.03	1.40
ntis	AUC	0.52	0.49	0.56	AUC*	0.54	0.48	0.60
	KS	0.09	0.03	0.15	KS*	0.11	0.03	0.18
	g/l	1.20	1.07	1.35	g/l*	1.24	1.07	1.45
tbl	AUC	0.53	0.50	0.57	AUC*	0.52	0.46	0.57
	KS	0.08	0.03	0.13	KS*	0.07	0.00	0.14
	g/l	1.17	1.07	1.29	g/l*	1.15	1.00	1.32
lty	AUC	0.54	0.50	0.57	AUC*	0.53	0.48	0.58
	KS	0.08	0.03	0.13	KS*	0.09	0.02	0.16
	g/l	1.18	1.06	1.31	g/l*	1.20	1.04	1.39
ltr	AUC	0.51	0.47	0.54	AUC*	0.53	0.48	0.58
	KS	0.06	0.00	0.11	KS*	0.06	0.00	0.12
	g/l	1.12	1.00	1.25	g/l*	1.12	1.00	1.26
tms	AUC	0.54	0.50	0.57	AUC*	0.53	0.48	0.58
	KS	0.08	0.03	0.13	KS*	0.10	0.03	0.18
	g/l	1.17	1.05	1.30	g/l*	1.23	1.06	1.43
dfy	AUC	0.47	0.44	0.51	AUC*	0.47	0.42	0.53
	KS	0.01	-0.02	0.04	KS*	0.05	-0.01	0.11
	g/l	1.03	0.97	1.09	g/l*	1.10	0.98	1.24
dfr	AUC	0.52	0.48	0.56	AUC*	0.53	0.48	0.59
	KS	0.07	0.02	0.13	KS*	0.09	0.01	0.17
	g/l	1.16	1.04	1.29	g/l*	1.20	1.03	1.40
linfl	AUC	0.55	0.52	0.59	AUC*	0.52	0.46	0.58
	KS	0.10	0.04	0.15	KS*	0.06	-0.01	0.14
	g/l	1.22	1.09	1.36	g/l*	1.14	0.98	1.32
csp	**AUC**	**0.55**	**0.51**	**0.59**	**AUC***	**0.57**	**0.52**	**0.63**
	KS	**0.08**	**0.03**	**0.14**	**KS***	**0.12**	**0.06**	**0.18**
	g/l	**1.18**	**1.06**	**1.31**	**g/l***	**1.27**	**1.12**	**1.44**

Notes: 468 out-of-sample observations except for *csp*, which has 397. 95% confidence interval calculated with 1,000 bootstrap replications.

Table 4 expands on these results by computing the risk-and-return-adjusted $AUC^\star$ using a power utility function and benchmark U.S. consumption time-series data. For the purposes of illustration, we explore risk-adjustment using a standard, simple constant relative risk aversion (CRRA) utility function $u(c) = c^{1-\gamma}/(1-\gamma)$. We then weight each period's returns (x) using the empirical next-period relative marginal utility weight (to get the stochastic discount factor m). We take the relevant consumption stream to be U.S. real per capita consumption, with growth rate g_t.[12] Each period, we treat real consumption "this period" as being normalized to 1 and then the random real consumption draw "next period" is given by $c = (1 + g_t)$, and for the case of CRRA utility the appropriate risk-weight to be applied to each return observation is then $m = u'(c) = (1 + g_t)^{-\gamma}$. That is, investment returns that deliver high payoffs in low consumption states are weighted more highly, as expected.

We utilize a wide range of coefficients of risk aversion, γ, from 1 (log utility) all the way to 80. Strikingly, the main lesson from Table 3 endures: only *csp* appears to provide consistently significant risk-adjusted returns (confidence intervals in Table 4 are omitted for brevity). The gain-loss ratio is around 1.35 (as low as 1.27 and as high as 1.39 depending on whether we look in-sample or out-of-sample with a coefficient of risk aversion, $\gamma = 80$), which roughly means that risk-adjusted expected gains are 35% larger than risk-adjusted expected losses. Moreover, a cursory look at Table 4 reveals another interesting feature worth highlighting: although we present a wide variation in risk aversion (including risk-neutral behavior), the $AUC^\star$ does not change very much at all. The main adjustment appears to come from returns relative to the pure directional signal in AUC, but once adjusted for returns, further adjustment for risk appear to make little difference. We will leave for a different paper a more detailed investigation of this attractive feature, but the bottom line here is that no matter what the risk aversion parameter γ, the stochastic discount factor is orthogonal to returns and consumption risk (with standard preferences) does not appear to have any ability to price equity returns.

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 MSPE based criteria. Comparing the results of our tests to the state-of-the-art methods in Goyal and Welch

[12] We follow Burnside et al. (2011) and construct U.S. real consumption growth as the expenditure-weighted sum of the real growth of nondurable goods and services consumption (SAAR series from BEA), and we subtract the monthly population growth (series from BLS) to obtain a monthly per capita growth rate g.

Table 4: Risk-Adjusted $AUC^{\star}$, $KS^{\star}$, and Gain-Loss Ratio for Equity Strategies: Various Coefficients of Risk Aversion and Power Utility. Out-of-sample Predictions: 1970:1–2008:12

		Coefficient of Risk Aversion								
Signal		$\gamma = 0$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 4$	$\gamma = 8$	$\gamma = 20$	$\gamma = 40$	$\gamma = 80$
dp	AUC*	0.498	0.498	0.498	0.498	0.498	0.497	0.495	0.492	0.487
	KS*	0.092	0.092	0.092	0.092	0.092	0.093	0.094	0.095	0.099
	G/L	1.202	1.202	1.203	1.203	1.203	1.204	1.207	1.210	1.219
dy	AUC*	0.505	0.505	0.505	0.505	0.505	0.504	0.502	0.499	0.493
	KS*	0.090	0.090	0.090	0.090	0.090	0.090	0.091	0.093	0.097
	G/L	1.197	1.197	1.197	1.197	1.198	1.198	1.201	1.205	1.214
ep	AUC*	0.516	0.516	0.516	0.516	0.515	0.515	0.513	0.509	0.501
	KS*	0.103	0.103	0.103	0.103	0.102	0.102	0.100	0.097	0.088
	G/L	1.229	1.229	1.229	1.229	1.228	1.226	1.222	1.214	1.193
de	AUC*	0.433	0.433	0.433	0.434	0.434	0.434	0.435	0.437	0.439
	KS*	0.017	0.017	0.017	0.017	0.017	0.017	0.018	0.020	0.024
	G/L	1.034	1.034	1.034	1.034	1.034	1.035	1.037	1.040	1.049
svar	AUC*	0.501	0.502	0.502	0.502	0.502	0.502	0.504	0.506	0.509
	KS*	0.057	0.057	0.057	0.057	0.057	0.058	0.059	0.060	0.062
	G/L	1.121	1.121	1.121	1.121	1.122	1.122	1.124	1.128	1.132
bm	AUC*	0.491	0.491	0.491	0.490	0.490	0.489	0.485	0.480	0.470
	KS*	0.073	0.073	0.073	0.073	0.072	0.071	0.069	0.065	0.059
	G/L	1.157	1.157	1.157	1.156	1.155	1.153	1.148	1.139	1.125
ntis	AUC*	0.512	0.512	0.512	0.512	0.512	0.512	0.511	0.511	0.512
	KS*	0.074	0.074	0.074	0.075	0.075	0.075	0.076	0.078	0.083
	G/L	1.161	1.161	1.161	1.161	1.161	1.162	1.165	1.170	1.182
tbl	AUC*	0.523	0.523	0.523	0.523	0.523	0.524	0.525	0.526	0.530
	KS*	0.109	0.109	0.109	0.109	0.109	0.109	0.110	0.112	0.117
	G/L	1.244	1.244	1.244	1.244	1.245	1.245	1.247	1.251	1.266
lty	AUC*	0.528	0.528	0.528	0.528	0.528	0.528	0.529	0.531	0.534
	KS*	0.079	0.079	0.079	0.079	0.079	0.078	0.077	0.080	0.084
	G/L	1.171	1.171	1.171	1.171	1.170	1.169	1.167	1.173	1.184
ltr	AUC*	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.563	0.564
	KS*	0.137	0.137	0.137	0.137	0.137	0.137	0.136	0.136	0.140
	G/L	1.317	1.317	1.317	1.316	1.316	1.316	1.315	1.314	1.325
tms	AUC*	0.528	0.528	0.528	0.528	0.529	0.529	0.530	0.531	0.533
	KS*	0.086	0.086	0.086	0.086	0.087	0.087	0.089	0.091	0.096
	G/L	1.188	1.188	1.189	1.189	1.189	1.191	1.194	1.201	1.213
dfy	AUC*	0.485	0.485	0.485	0.485	0.485	0.485	0.486	0.488	0.493
	KS*	0.087	0.087	0.087	0.087	0.087	0.087	0.086	0.086	0.085
	G/L	1.191	1.190	1.190	1.190	1.190	1.190	1.188	1.187	1.187
dfr	AUC*	0.477	0.477	0.477	0.477	0.477	0.477	0.476	0.474	0.469
	KS*	0.082	0.082	0.082	0.083	0.083	0.083	0.083	0.083	0.082
	G/L	1.180	1.180	1.180	1.180	1.180	1.180	1.181	1.182	1.178
linfl	AUC*	0.530	0.530	0.530	0.530	0.529	0.529	0.527	0.524	0.517
	KS*	0.105	0.105	0.105	0.105	0.104	0.103	0.100	0.094	0.081
	G/L	1.235	1.235	1.235	1.234	1.233	1.230	1.222	1.208	1.176
csp	**AUC***	**0.576**	**0.577**	**0.577**	**0.577**	**0.577**	**0.577**	**0.577**	**0.578**	**0.579**
	KS*	**0.147**	**0.147**	**0.147**	**0.147**	**0.148**	**0.149**	**0.151**	**0.156**	**0.164**
	G/L*	**1.344**	**1.344**	**1.345**	**1.345**	**1.347**	**1.349**	**1.357**	**1.369**	**0.393**

Notes: 468 observations except for *csp* strategy, with 397. *csp* is the only strategy with statistically significant AUC and KS (using the bootstrap), although standard errors not reported here for brevity. We use boldface to highlight this finding.

(2008), we find important differences in the relative merits of different signals, which mostly appear when we adjust for returns and to a lesser extent, risk.

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).

However, this support is still subject to two caveats. The first is conceptual, for as Goyal and Welch (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.

8 Empirical Application II: Currency Carry Trades with Long/Cash/Short Positions

Berge, Jordà and A. M. Taylor (2011) 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). 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 5 presents the out-of-sample (OOS) $AUC/AUC^\star$, $KS/KS^\star$ statistics and gain-loss ratios where we adjust for returns and then risk-adjust as in the equity example by considering values of $\gamma = 2$, 4, 8, 20, 40 and 80, associated to each of these four strategies for the 648 currency-month observations in our chosen OOS sample window from 2003:1 to 2008:12.

The results reported in Table 5 (with country-clustered bootstrap standard errors) highlight once more the difference between good classification ability, profitability and risk-adjusted 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 0.622, well above the 0.5 null and highly statistically significant, although *momentum* and *value* are both close with $AUC^\star = 0.604$ and 0.600 respectively. And as was the case for the equity strategies, adjusting for risk (even with risk-aversion jacked up to 80) does not alter this ranking significantly. The evidence here shows that

Table 5: Out-of-sample AUC, KS, and Gain-Loss: Raw, Risk-Neutral, and Risk-Adjusted for Several Values of the Risk Aversion Coefficient. Currency Strategies: 2003:1–2008:12

Signal	**Stat.**	**Raw**	**RN**	$\gamma = 2$	$\gamma = 4$	$\gamma = 8$	$\gamma = 20$	$\gamma = 40$	$\gamma = 80$
Carry	AUC	0.544	0.478	0.478	0.478	0.477	0.475	0.471	0.464
		(0.033)	(0.050)						
	KS	0.094**	0.040	0.040	0.040	0.039	0.039	0.037	0.035
		(0.044)	(0.049)						
	G/L	1.208	1.083	1.083	1.082	1.082	1.080	1.078	1.073
Moment.	AUC	0.552	0.604	0.604	0.604	0.604	0.604	0.605	0.605
		(0.038)	(0.063)						
	KS	0.105**	0.161**	0.161	0.161	0.161	0.161	0.161	0.159
		(0.042)	(0.073)						
	G/L	1.234	1.384	1.384	1.384	1.384	1.384	1.383	1.379
Value	AUC	0.536	0.600**	0.601	0.601	0.601	0.603	0.605	0.609
		(0.034)	(0.048)						
	KS	0.076**	0.076	0.193	0.194	0.194	0.197	0.201	0.211
		(0.043)	(0.043)						
	G/L	1.164	1.164	1.479	1.480	1.483	1.491	1.504	1.534
VECM	AUC	0.600**	0.622**	0.622	0.622	0.622	0.622	0.621	0.619
		(0.033)	(0.045)						
	KS	0.183**	0.241**	0.241	0.241	0.241	0.241	0.241	0.242
		(0.040)	(0.058)						
	G/L	1.448	1.633	1.634	1.634	1.634	1.634	1.636	1.640

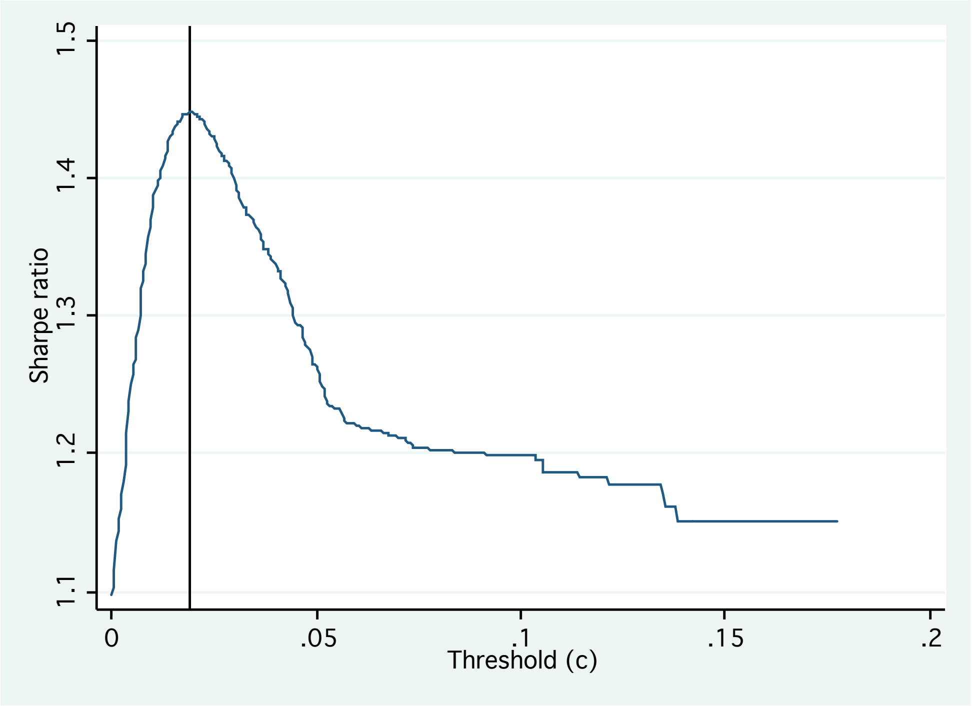
Notes: 648 out-of-sample observations. Raw refers to the un-weighted versions of the statistics; RN refers to the risk-neutral weighted version (i.e., weighted by returns); γ refers to the values of the coefficient of risk-aversion in a power utility function. Standard errors in parenthesis calculated with 1,000 clustered bootstrap replications to allow for country-level correlation. Notice the distribution of the KS statistic is not normal (see text). G/L refers to the gain-loss ratio calculated from the KS statistic under the assumption of symmetry.

allowance for standard consumption risk is unable to help price the returns from carry trades or other currency strategies.

Indeed, the trading profits delivered by the *vecm* strategy (returns-adjusted) 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 24 bps per month on each position, with each trade having a standard deviation of 315 bps. Thanks to diversification, the returns on the portfolio of 9 currencies had a standard deviation of 161 bps. Annualized, the strategy would have delivered 2.9 percent per year compounded with a Annualized Sharpe Ratio of 0.46.

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 deci-

Figure 3: Sharpe Ratio as a Function of Threshold



sion space is binary, there is no ambiguity in determining the ex-post profitable long/short direction. However, by now adding a cash position, we need some criterion to determine the ex-post correct choice of long/cash/short positions.

Absent good data on transactions costs, we decided to calculate a minimum symmetric return threshold ϕ 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 ϕ that would maximize the ex-post Sharpe ratio for a \$1 investment. This is reported in Figure 3 and shows that the ex-post Sharpe ratio is maximized for $\phi = 1.91\%$. 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 can now ask how would the four benchmark carry trade strategies reported in Table 5 fare if one allowed for a cash position and for this we calculated each strategy's

Table 6: Out-of-sample VUS, and Risk-Adjusted $VUS^{\star}$ for Several Values of the Risk Aversion Coefficient. Currency Strategies: 2003:1–2008:12

Signal	**VUS**	**VUS* - RN**	**VUS*:** $\gamma = 2$	**VUS*:** $\gamma = 4$	**VUS*:** $\gamma = 8$
Carry	0.183	0.147	0.147	0.148	0.148
	[0.128, 0.237]	[0.101, 0.194]			
Momentum	0.209	0.249	0.249	0.249	0.250
	[0.136, 0.283]	[0.088, 0.410]			
Value	0.209	0.232	0.233	0.233	0.234
	[0.146, 0.271]	[0.117, 0.348]			
VECM	0.229	0.227	0.227	0.227	0.228
	[0.162, 0.297]	[0.127, 0.326]			

Notes: 648 out-of-sample observations. VUS refers to the un-weighted volume under the CC surface statistic; $VUS^{\star}$ - RN refers to the risk-neutral weighted version (i.e., weighted by returns); γ refers to the values of the coefficient of risk-aversion in a power utility function. 95% confidence interval in brackets calculated with 1,000 clustered bootstrap replications to allow for country-level correlation. The chance value for the VUS statistic is 1/6 = 0.16667.

$VUS/VUS^{\star}$ statistics, the results of which are reported in Table 6. (Recall that this is for the risk-neutral case, so unlike our equity analysis, no consumption data are used for this analysis.)

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, all signals fail to beat this simple null (95% confidence intervals are calculated with the country-clusterered bootstrap) although we remark that these are estimated somewhat imprecisely (the confidence intervals are a bit larger than their analytic large-sample counterparts). 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 (the $AUCs$ reported in Table 5 are indeed close) although not 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. Just as in the results for the AUCs reported in Table 5, adjusting for risk did not change these conclusions substantively.

9 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 the MSE yardstick fare under our framework and show that there is perhaps one strategy with statistically significant returns. Our application to the carry trade is based on the data in Berge, Jordà and A. M. Taylor (2011) and identifies a strategy that generates a statistically significant departure from no arbitrage, 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 nonparametric but simple to implement 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 set of standards to analyze an important class of problems in empirical finance.

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.

Andrews, Donald W. K. 1997. A Conditional Kolmogorov Test. *Econometrica*, 65(5): 1097–1128.

Bai, Jushan. 1994. Weak Convergence of the Sequential Empirical Processes of Residuals in ARMA Models. *The Annals of Statistics*, 22(4): 2051–2061.

Bai, Jushan. 2003. Testing Parametric Conditional Distributions of Dynamic Models. *The Review of Economics and Statistics*, 85(3): 531–549.

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 Olivier Ledoit. 2000. Gain, Loss and Asset Pricing. *Journal of Political Economy* 108(1): 144–72.

Bickel, Peter J., and David A. Freedman. 1981. Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6): 1196–1217.

Brennan, Michael J., and Yihong Xia. 2005. Persistence, Predictability, and Portfolio Planning. Rodney L. White Center for Financial Research, Working Paper no. 25-05.

Burnside, Craig, Martin Eichenbaum, Isaac Kleshchelski, and Sergio Rebelo. 2011. Do Peso Problems Explain the Returns to the Carry Trade? *Review of Financial Studies* 24(3): 853–89.

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.

Darling, Donald A. 1955. The Cramér-Smirnov Test in the Parametric Case. *The Annals of Mathematical Statistics*, 26: 1–20.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44:837–845.

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. Photocopy.

Fujii, Eiji, and Menzie D. Chinn. 2001. Fin de Siècle Real Interest Parity. *Journal of International Financial Markets, Institutions and Money* 11(3–4): 289–308.

Goyal, Amit, and Ivo Welch. 2003. Predicting the Equity Premium with Dividend Ratios. *Management Science*, 49(5):639–54.

Goyal, Amit, and Ivo Welch. 2008. A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *Review of Financial Studies*, 21(4): 1455–1508.

Granger, Clive W. J., and Mark J. Machina. 2006. Forecasting and Decision Theory. In *Handbook of Economic Forecasting*, Vol. I, Graham Elliott, Clive W. J. Granger, and Allan Timmermann (eds.). Amsterdam: Elsevier.

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

Hájek, Jaroslav, Zbyněk Šidák, and P. K. Sen. 1999. *Theory of Rank Tests.* San Diego, Calif.: 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 Karimollah Hajian-Tilaki. 1997. Sampling variability of Nonparametric estimates of the Areas Under receiver Operating Characteristic Curves: An Update. *Academic Radiology*, 4: 49–58.

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.

Hansen, Lars P. and Ravi Jaganathan. 1991. Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy*, 99: 225–262.

Henriksson, Roy D. and Robert C. Merton. 1981. On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills. *Journal of Business* 54: 513–533.

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.

Khamaladze, Estate V. 1981. Martingale Approach in the Theory of Goodness-of-Fit Tests. *Theory of Probability and its Applications*, 26, 240–257.

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.

Kosorok, Michael R. 2008. *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer-Verlag.

Linton, Oliver, Esfandiar Maasoumi, and Yoon-Jae Whang. 2005. Consistent Testing for Stochastic Dominance under General Sampling Schemes. *The Review of Economic Studies*, 72:735–765.

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

Merton, Robert C. 1981. On Market Timing and Investment Performance. I. An Equilibrium theory of Value for Market Forecasts. *Journal of Business*, 54: 363–406.

Mielke, Jr., Paul W., and Kenneth J. Berry. 2007. *Permutation Methods: A Distance Function Approach.* New York: Springer-Verlag.

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.

Pesaran, M. Hashem, and Allan Timmermann. 1992. A Simple Nonparametric Test of Predictive Performance. *Journal of Business and Economic Statistics*, 10(4): 461–465.

Pesaran, M. Hashem, and Allan Timmermann. 2009. Testing Dependence Among Serially Correlated Multicategory Variables. *Journal of the American Statistical Association*, 104(485): 325–337.

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.

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

Rubin, James. 1973. Weak Convergence of the Sample Distribution Function When Parameters are Estimated. *The Annals of Statistics*, 1(2): 279–290.

Shorack, Galen R., and Jon A. Wellner. *Empirical Processes with Applications to Statistics.* Hoboken, N.J.: John Wiley & Sons.

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.

van de Vaart, Aad W. 1998. *Asymptotic Statistics.* Cambridge: Cambridge University 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: 835–48.

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

World Meteorological Organization. 2000. *Standard Verification System for Long-Range Forecasts.* Geneva: World Meteorological Organization.

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

Zhou, Xia-Hua, Nancy A. Obuchowski, and Donna K. McClish. 2011. *Statistical Methods in Diagnostic Medicine.* Hoboken, N.J.: John Wiley & Sons.