Backtesting market risk models in a standard normality framework

Kevin Dowd

Centre for Risk and Insurance Studies, Nottingham University Business School, Jubilee Campus, Nottingham NG8 1BB, UK.

This paper looks at various ways in which the problem of evaluating a market risk forecasting model can be reduced to the simple problem of testing for iid standard normality. The transformations involved allow us to extend the one-step ahead Berkowitz backtesting framework and its resulting prediction of iid standard normality to encompass complications associated with truncated distributions, multi-step ahead density forecasting models, multivariate density forecasting models, risk aggregation and parameter uncertainty.

1 Introduction

One of the most important aspects of good risk modeling is the need to evaluate (or "backtest") risk forecasting models. However, risk models are more difficult to evaluate than traditional forecasts of expected values. This is partly because a risk model forecasts a complete density function and not just its mean, and partly because density forecasts will typically change from one period to the next. An elegant way to deal with these difficulties was suggested by Crnkovic and Drachman (1996) and Diebold *et al* (1998). They suggested that realized values of the variable whose density is being forecast should be mapped to their probability integral transform (PIT) or forecast cumulative density values. In the case that they considered, these PIT values are predicted to be independently and identically distributed (iid) standard uniform under the null hypothesis of model adequacy, so we can test model adequacy by testing whether the PIT values are uniformly distributed. However, testing for uniformity is not always straightforward, and Berkowitz (2001) suggested that backtesting could be simplified further by putting the PITs themselves through a transformation to make them iid standard normal under the null. This second transformation reduces backtesting to the straightforward task of testing for iid standard normality.

This Berkowitz backtesting strategy – of transforming the data to make them iid standard normal under the null and then testing them for this distribution – reduces an initially complicated problem to a very tractable one. However, the

This work was carried out under an ESRC Research Fellowship on "Risk Measurement in Financial Institutions" (RES-000-27-0014), and the author thanks the ESRC for their financial support. The author also thanks Barry Schachter for helpful comments on an earlier draft, but the usual caveat applies.

"basic" approach as originally set out by Berkowitz is limited in that it presupposes a rather simple backtesting problem, and many real-world backtesting problems are more complicated and so more difficult. To give an illustration, the iid prediction only arises where density forecasts are one step ahead, as might be the case if we had a VAR model that forecasts a one-day VAR on a daily basis. Indeed, this was exactly the case considered by Berkowitz and the earlier writers mentioned above. Their approaches work well in this case, because they generate iid predictions that enable us to apply conventional textbook tests of either standard uniformity or standard normality. Nevertheless, we can also get multi-step ahead models – an example of which would be a 10-day VAR model forecast on a daily basis – and in these cases we do not get an iid prediction. The absence of such a prediction then causes problems for backtesting, because it undermines the reliability of conventional tests. The absence of an iid prediction also means that we cannot interpret the results of an iid test as a test of model adequacy. A model evaluator who failed to appreciate this point would then be in danger of making a major error: should the model fail such a test, he/she would be tempted to interpret this result as indicating that the model was inadequate, whereas the "correct" interpretation was that the model should fail such a test, because the "correct" prediction is that the data should be dependent.

Besides the complications arising from multi-step ahead models, real-world backtesting problems are often complicated in other ways that the Berkowitz analysis does not easily accommodate. Such complications often arise in problems involving multivariate density forecasts, multiple models at different business-unit levels and parameter uncertainty. As with multi-step models, there is the danger that failure to take account of these complications might undermine the reliability of backtesting procedures and lead to incorrect conclusions about model adequacy.

This paper investigates these issues further and shows how the Berkowitz backtesting strategy can be successfully applied to a variety of more difficult backtesting problems. The paper is organized as follows. Section 2 outlines the base case of a one-step ahead density forecasting model, which was the focus of the Berkowitz paper and of the papers by Crnkovic and Drachman and Diebold *et al* that preceded it, and discusses how the data can be transformed to become iid $N(0,1)$ under the null hypothesis of model adequacy. Sections 3 to 7 then address how more complicated problems can be simplified so that they can be handled in the same way, ie, they show how the null of model adequacy in more involved situations can also be reduced to a prediction of iid $N(0,1)$: Section 3 discusses the case of a truncated distribution, where we are only interested in part of the return or loss distribution; Section 4 addresses multi-step ahead density forecasting models; Section 5 examines multivariate density forecasting models; Section 6 examines the case of multiple density forecasting models at different business-unit levels; and Section 7 examines the impact of parameter uncertainty on backtesting. Section 8 concludes.

2 Base case: evaluating a one-step ahead model

We begin with our base case involving a one-step ahead density forecasting model. An example is a model that forecasts the density of the financial return, X_t, over a horizon equal to a trading day, where forecasts are made and returns observed over a daily frequency. The X_t process will typically vary from one model to another (eg, it might be normal in one case, a t in another, etc) and the parameters for any given model will also typically change from one day to the next as VAR modelers recalibrate their models on a daily basis. The resulting changes in the density forecasts make it difficult to directly compare different observations of X_t over time, but we can make our observations comparable if we apply the following transformation:

$$p_t = F_t(X_t) \tag{1}$$

where $F_t(\cdot)$ maps the realized value X_t to the value of its forecast cumulative density function where the density forecast is made the previous day. This mapping is often known as the probability integral transform (PIT). A sample of PIT values is predicted to be distributed as standard uniform (ie, $U(0, 1)$) under the null that the forecasts are adequate. This prediction arises because forecast densities should match actual densities under the null, and this is so even if the forecast densities themselves change over time. p_t is also independent because consecutive values of p_t have no common factors. Thus, p_t is predicted to be iid $U(0, 1)$ under the null.[1]

We could test this prediction directly,[2] but it is more convenient to put the p_t through a second transformation to make them standard normal under the null. Following Berkowitz (2001), we work with the transformed variable:

$$z_t = \Phi^{-1}(p_t) \tag{2}$$

where $\Phi(\cdot)$ is the standard normal distribution function. This second (or Berkowitz) transformation is helpful because tests for standard normality are typically more straightforward (and also more numerous) than tests for standard uniformity,[3] and because dealing with a normal variable is more convenient when considering potential temporal dependence. Under the null, z_t will be distributed as iid $N(0, 1)$.[4]

[1] A formal proof of this prediction is provided by Diebold *et al* (1998, pp. 865, 867–9).

[2] We might test for standard uniformity in various ways: we might use Kolmogorov–Smirnov tests, Kuipier tests, Lilliefors tests, Anderson–Darling tests, Shapiro–Wilks tests, and so on. We might also bin observations and apply a chi-squared test to the hypothesis that actual frequencies match predicted ones. For more on these and similar tests, see Dowd (2005a, pp. 343–9).

[3] There are problems with some of the commonly used uniformity tests. For example, the most widely used test, the Kolmogorov–Smirnov test, is unreliable if model parameters are estimated rather than known, and is also known to be more sensitive to central-mass rather than tail departures from the null hypothesis. Other tests such as Anderson–Darling and Shapiro–Wilks can be difficult to implement, and chi-squared frequency tests have the disadvantage that they are dependent on arbitrary assumptions about the number and widths of bins.

[4] Note that z_t is predicted to be iid because the iid-ness of the p_t carries over to the z_t.

Testing model adequacy now boils down to testing whether z_t is distributed as iid $N(0,1)$,[5] and we would test this prediction using conventional tests.[6]

In short, the backtesting strategy is to transform the data so that they are iid $N(0,1)$ under the null, and then apply a battery of tests to determine whether the transformed data are consistent with the iid $N(0,1)$ prediction.

We now look at ways in which we can apply this same strategy to deal with more complicated backtesting problems. As we have nothing new to say about the tests themselves, we henceforth focus on the first part of this strategy, ie, how to transform the data to make them iid $N(0,1)$.

3 A truncated distribution

Our first case is where we are interested in only part of the return distribution: typically, we might be interested in the distribution of lower tail returns where we make losses in excess of VAR.[7] If we are only interested in the lower tail and are working to a confidence level of α, we can take our earlier p_t series and delete all non-tail observations from it. We then end up with a series that is iid uniformly distributed over the interval $[0,1-\alpha]$, and this implies that $p_t/(1-\alpha)$ is iid uniformly distributed over the interval $[0,1]$. Thus, all that is required is to delete the non-tail observations, divide p_t by $(1-\alpha)$ and apply the Berkowitz transformation to obtain the series $z_t = \Phi^{-1}\ (p_t/(1-\alpha))$, which is distributed as iid $N(0,1)$ under the null.

4 A multi-step ahead model

A more complicated situation arises where we have multi-step ahead density forecasting models. A typical case is a model that forecasts densities for $X_{k,t}$, the

[5] Before carrying out any formal tests, the model evaluator should first carry out some informal analysis to get a preliminary sense of the properties of the data. This analysis might involve tabulating the sample moments of z_t to see if they are "close" to their predicted values, plotting histograms and correlograms of z_t, checking QQ and similar plots to see if they are close to those predicted, etc. For more on these and similar methods, see, eg, Diebold *et al* (1998).

[6] We can think of these tests as falling into two categories. The first are tests for standard normality presupposing that the data are iid, and these would include textbook z-tests and t-tests of the prediction that the mean is zero, variance-ratio chi-squared tests of the prediction that the variance is 1, likelihood ratio (LR) tests of the joint prediction that the mean is 0 and the variance is 1, and tests such as the Jarque–Bera (JB) test of the higher moment (skewness and kurtosis) predictions of normality. The second category of tests are tests for iid itself, and these would include the Ljung–Box portmanteau test, runs tests (Wolfowitz, 1943; David, 1947), BDS tests (Brock *et al,* 1987), duration tests based on the amount of time that elapses between violations (Christoffersen and Pelletier, 2004), martingale tests based on the principle that violations should be unpredictable (see, eg, Engle and Manganelli, 2004) and spectral density tests (Durlauf, 1991).

[7] The case of a truncated distribution was also considered by Berkowitz (2001, p. 479), who suggested dealing with it by means of a truncated log-likelihood. This is a reasonable suggestion, but the approach suggested in the text is easier.

average return over a horizon of k days:

$$X_{k,t} = \frac{1}{k}\sum_{i=1}^{k} X_{1,t-i+1} \tag{3}$$

where $X_{1,t}$ are observed daily returns and forecasts are made on a daily basis. An example is a model that every day forecasts a VAR over a horizon of 10 trading days. As before, we standardize our observations by applying the PIT transformation:

$$p_{k,t} = F_{k,t}(X_{k,t}) \tag{4}$$

where $F_{k,t}(\cdot)$ maps the realized k-day average return, $X_{k,t}$, to its forecast CDF value with the forecast made k days earlier. As with the one-step ahead forecasts, we can now deduce that $p_{k,t}$ is stationary and distributed as standard uniform under the hypothesis that the density forecasting model is adequate. However, $p_{k,t}$ is not independent (except in the special case of $k=1$, which is the earlier one-step ahead case) because consecutive values of $X_{k,t}$ share $k-1$ common factors.[8] Thus, the $p_{k,t}$ are predicted to be $U(0,1)$ but for $k>1$ the $p_{k,t}$ are not predicted to be iid $U(0,1)$. We now apply the Berkowitz transformation:

$$z_{k,t} = \Phi^{-1}(p_{k,t}) \tag{5}$$

and it follows that $z_{k,t}$ is $N(0,1)$ but not iid because the lack of iid-ness of $p_{k,t}$ carries over to lack of iid-ness of $z_{k,t}$. We now have a $z_{k,t}$ that is standard normal but has a dependence structure.

This implies that standard tests that assume iid might be unreliable. So what do we do?

4.1 Converting to a one-step ahead problem

One possible response is to convert the k-step ahead problem into a one-step ahead one. Suppose we have samples of n sets of k-day-horizon density forecasts and matching realized values of $X_{k,t}$. If we treat this problem as a k-step ahead forecasting problem, we have $p_{k,t}$ and $z_{k,t}$ samples of size n that are not independent. However, we can obtain independent samples if we reduce the observation frequency from one observation a day to one observation every k days: in effect, we throw away information about all density forecasts except the first, $k+1$st, $2k+1$st, … forecasts, and we roll our original set of n realized daily returns into a set of n/k non-overlapping realized k-day returns. This converts the multi-step ahead problem into a one-step ahead one, where the phrase "one-step ahead" now refers to a forecast horizon of k days. Because the problem is now one-step ahead, the analysis of Section 2 applies and we get an iid prediction that makes testing straightforward.

[8] For more on the dependence structure of the $p_{k,t}$, see also Diebold *et al* (1998, p. 880).

Nevertheless, this “solution” comes at a high price: it throws away a large amount of potentially useful data, which is statistically inefficient; and it reduces our effective sample size from n to n/k, which reduces the power of any tests we apply. Both problems also get worse as k gets bigger, and there comes a point, as k continues to increase, where so much information has been lost and our sample sizes have become so small that this “solution” becomes unfeasible.

4.2 Bonferroni approach

A second option is to work with Bonferroni sub-samples. The underlying idea is that the maximum length of the dependence structure is known, so we can create iid sub-samples from an original sample by selecting observations that are suitably spaced apart in time. This leads to the original sample of length n being broken down into k sub-samples of length n/k, each of which exhibits independence. However, the main disadvantage of this method is that it is not feasible if n is relatively small or k relatively high, because in such circumstances the sub-samples are too small to have much power.[9]

4.3 Iid resample bootstrap

A third approach is to make use of a modified bootstrap based on a resampling algorithm that makes the resample iid by construction. This algorithm works by selecting each odd element of the resample as any of the first $n-k$ elements of the original sample chosen at random. For each such element chosen, the corresponding even element of the resample is then taken as the sample observation occurring k periods later. The values of the odd and even elements of the resample are then independent because the temporal gap between them – k periods – exceeds the maximum length of the dependence structure, which is only $k-1$. One can easily show that all elements of the resample are independent, and their independence allows us to apply conventional tests to the resample that presuppose independence.[10] A (minor) drawback of this approach is that we have to live with a certain amount of resampling error (as with any bootstrap), but simulation evidence suggests that tests based on this approach have good power properties for “reasonable” sample sizes.[11] Indeed, one can show that for “reasonable” sample

[9] A partial way round this problem is to splice Bonferroni sub-samples together to produce a “resample” of the same length as the original sample. This new “resample” will be (nearly) iid but is very unlikely to be fully iid because of the potential for dependence in the “joins” between the different sub-samples. This “join problem” is similar to the join problems that arise with block bootstraps, where observations within each block are independent, but the adjacent observations across the joins between different blocks might be related. However, the statistical properties of tests based on such “resamples” are not well understood.

[10] However, in doing so, we also have to recalibrate the size or type I error rate of the test: this is because the resample is not drawn from a normal distribution under the null, but is instead drawn from the distribution of a sample that is itself drawn from the normal. More details on these issues are to be found in Dowd (2005b).

[11] To give an example, the simulation results presented in Dowd (2005b) suggest that for a

sizes tests applied to iid resamples have power rates approaching those of the same tests applied to iid samples.

4.4 Working with an estimated dependence structure

The three approaches discussed in Sections 4.1 to 4.3 have in common that they deal with a $z_{k,t}$ dependence structure treated as unknown. However, instead of treating the dependence structure as unknown, we could also work with an estimate of it.[12]

One way to do so is to estimate the $z_{k,t}$ dependence structure using ARMA methods: we would use Box–Jenkins analysis to identify the best-fitting ARMA process, taking care to ensure that we select a process that is as parsimonious as possible. For example, we might find that the "best-fitting" process is an AR(1):

$$z_{k,t} = \phi z_{k,t-1} + \varepsilon_t \tag{6}$$

where $|\phi| < 1$ and ε_t is an iid normal noise process. The variances of $z_{k,t}$ and ε_t are related in the following way:

$$\begin{aligned} &\operatorname{var}(z_{k,t}) = \phi^2 \operatorname{var}(z_{k,t-1}) + \operatorname{var}(\varepsilon_t) \\ &\Rightarrow \operatorname{var}(z_{k,t}) = \frac{1}{1-\phi^2}\operatorname{var}(\varepsilon_t) \end{aligned} \tag{7}$$

We also know that under the null, the variance of $z_{k,t}$ should be 1, so the variance of ε_t should be:

$$\operatorname{var}(\varepsilon_t) = 1 - \phi^2 \tag{8}$$

Under the null ε_t is therefore distributed as iid $N(0, 1-\phi^2)$. We then rearrange (6) to obtain

$$z_{k,t} - \phi z_{k,t-1} = \varepsilon_t \sim \text{iid } N(0, 1-\phi^2) \tag{9}$$

$$\Rightarrow \frac{z_{k,t} - \phi z_{k,t-1}}{\sqrt{1-\phi^2}} \sim \text{iid } N(0,1) \tag{10}$$

which gives us our prediction of iid $N(0, 1)$.

sample size of 250, the LR joint test of the standard normal mean and variance predictions applied to iid resamples has an average power rate equal to about 50% of the power rate of the same test applied to truly iid samples, where the departures from the null are reflected in a non-standard normal distribution. Where departures from the null involve non-normal skewness and kurtosis, comparable results suggest that the Jarque–Bera test has a power rate equal to nearly 80% of the power rate of the same tests applied to truly iid samples. Thus, standard tests seem to have respectable power properties applied to iid-resamples, relative to the powers of the same tests under "ideal" iid circumstances.

[12] From a statistical point of view, an estimate gives us some information about the dependence structure, so one can argue that it is statistically more efficient to work with the information provided by the estimate even though that information may be imperfect, than to ignore that information entirely.

A similar approach can also be applied if $z_{k,t}$ has a more general ARMA structure than a simple AR(1). To illustrate further, we might find that the best-fitting process is an ARMA(1,1):

$$z_{k,t} = \phi z_{k,t-1} + \varepsilon_t - \theta\varepsilon_{t-1} \tag{11}$$

The variance of $z_{k,t}$ is then given by[13]

$$\text{var}(z_{k,t}) = \frac{(1-2\phi\theta+\theta^2)}{1-\phi^2}\text{var}(\varepsilon_t) \tag{12}$$

Setting $\text{var}(z_{k,t}) = 1$ and rearranging then gives us

$$\text{var}(\varepsilon_t) = \frac{1-\phi^2}{1-2\phi\theta+\theta^2} \tag{13}$$

We now rearrange (11) to get

$$\begin{aligned} & z_{k,t} - \phi z_{k,t-1} + \theta\varepsilon_{t-1} = \varepsilon_t \sim \text{iid } N\left(0, \frac{1-\phi^2}{1-2\phi\theta+\theta^2}\right) \\ & \Rightarrow \sqrt{\frac{1-2\phi\theta+\theta^2}{1-\phi^2}} \times (z_{k,t} - \phi z_{k,t-1} + \theta\varepsilon_{t-1}) \sim \text{iid } N(0,1) \end{aligned} \tag{14}$$

and we again get our prediction of iid $N(0,1)$.

Nevertheless, we should keep in mind that this ARMA-based approach has two significant limitations. The first is that the approach works with estimates of the ARMA parameters and is therefore exposed to the risk of estimation error. This issue is discussed further below in Section 7. The other limitation is that the ARMA framework presupposes that the ARMA residuals, the ε_t, are normally distributed. This implies that any departures of $z_{k,t}$ from standard normality are assumed to manifest themselves in failures of the mean and variance predictions of $z_{k,t}$, ie, and not in terms of higher moment (or skewness and kurtosis) predictions. This means, in turn, that we cannot use this approach to test for possible failures of the skewness and kurtosis predictions of $z_{k,t}$.

5 A multivariate density model

Another interesting case is a multivariate density forecasting model, a model that predicts the density for more than one random variable. Examples in risk management would be models that predict densities for losses on two different portfolios, or models that predict densities for both credit exposures and losses on a credit portfolio. In macroeconomics, there has also been some discussion of multivariate "fan chart" models that predict the joint density of inflation and real economic growth (eg, Blix and Sellin, 2000)).

[13] See Tsay (2005, p. 57).

For the sake of argument, suppose we wish to evaluate a model that predicts the multivariate density for two random variables X_1 and X_2. Let $F(x_1, x_2)$ (ie, the probability that X_1 is less than or equal to some constant x_1 and that X_2 is less than or equal to some other constant x_2) be their joint distribution function with continuous marginals $F_1(x_1) = u_1$ and $F_2(x_2) = u_2$ (ie, so $F_1(x_1) = u_1$ is the probability that $X_1 \leq x_1$, and this probability is set to u_1, etc.). To understand the issues involved, it helps to couch the problem in the language of copulas, and the famous Sklar theorem (Sklar, 1959) tells us $F_1(x_1, x_2)$ that can be written in terms of a unique function $C(u_1, u_2)$:

$$F(x_1, x_2) = C(u_1, u_2) \tag{15}$$

where $C(u_1, u_2)$ is known as the copula of $F(x_1, x_2)$. The copula function gives an alternative representation of $F(x_1, x_2)$ that describes how the multivariate function $F(x_1, x_2)$ is coupled with the marginal distribution functions $F_1(x_1)$ and $F_2(x_2)$. It can also can be interpreted as giving the dependence structure of $F(x_1, x_2)$.

From this perspective, the evaluation of a multivariate density forecasting model can be broken down into two steps.[14] The first is straightforward and involves the evaluation of the marginal functions, which provide the inputs to the copula. Under the null, the marginals should be distributed as $U(0, 1)$ and will often be (or can be transformed to be[15]) distributed as iid $U(0, 1)$. All we then need to do is to run each set of marginals through the Berkowitz transformation.

5.1 Evaluating the copula

The second step is to evaluate the copula itself. The evaluation of copulas is the subject of a rapidly growing literature in its own right,[16] but can also be handled in the spirit of our iid $N(0, 1)$ strategy. To see how this might work, first consider the value of the copula function as itself a random variable driven off random $U(0, 1)$ inputs. This means that the value of the copula has a distribution of its own. Now define two alternative copulas, an estimated (or H0) copula, which is the copula we wish to test, and an H1 copula, which is a hypothetically true alternative copula. Let c^{H0} be a set of H0 copula values with distribution function $F^{H0}(\cdot)$, and let c^{H1} and $F^{H1}(\cdot)$ be their H1 equivalents. Now note that if the two copulas match, we would expect $p = F^{H1}(c^{H0})$ to be uniformly distributed, and if the two copulas do not match, we would expect the p to be non-uniformly distributed.

These expectations therefore provide the basis for our evaluation, and we go through the following steps:

[14] One can also evaluate the complete multivariate density function as a whole. The evaluation of multivariate density forecasts is addressed in a number of places, eg, Diebold *et al* (1999) and Patton (2005). See also note 17.

[15] For example, it may be that returns follow GARCH-type processes (and are therefore not iid), and we might fit the copula to their residuals, which are iid.

[16] For more on the subject of copula evaluation, see, eg, Fermanian (2005), Savu and Trede (2004) and Dobrić and Schmid (2005). Dowd (2006) has a survey of these and other studies, and also presents some new tests of the adequacy of fitted copulas.

(1) We take the estimated (H0) copula, input random $U(0,1)$ variables to it, and obtain a corresponding set of copula values c^{H0}.[17]

(2) We write a function that enables us to obtain estimates of $F^{H1}(\cdot)$.

(3) After this, we map the copula values c^{H0} (obtained in step (1)) to their $F^{H1}(\cdot)$ values (using the function obtained in step (2)), ie, we obtain the series $p = F^{H1}(c^{H0})$.[18] If the estimated copula matches the assumed copula, then p should be $U(0,1)$, and if there is no temporal dependence in the marginals, p should also be iid.[19]

(4) Finally, we take the Berkowitz transformation and map the p series to their standard normal equivalents. Assuming that the p are iid, the resulting z series will also be iid $N(0,1)$.

5.2 Copula evaluation: a worked example

It might help to expand on this method using a worked example. Let us therefore suppose that we have a bivariate Gaussian copula as specified in (15) below:

$$C^{Ga}(u_1, u_2) = \Phi_\rho\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right) \tag{16}$$

where $\Phi_\rho(\cdot)$ is the joint bivariate Gaussian distribution function predicated on the Pearson correlation ρ.[20] Let us also suppose that we have a H0 (ie, estimated) copula predicated on $\rho = 0$ and an alternative (true) copula predicated on $\rho = 0.5$.

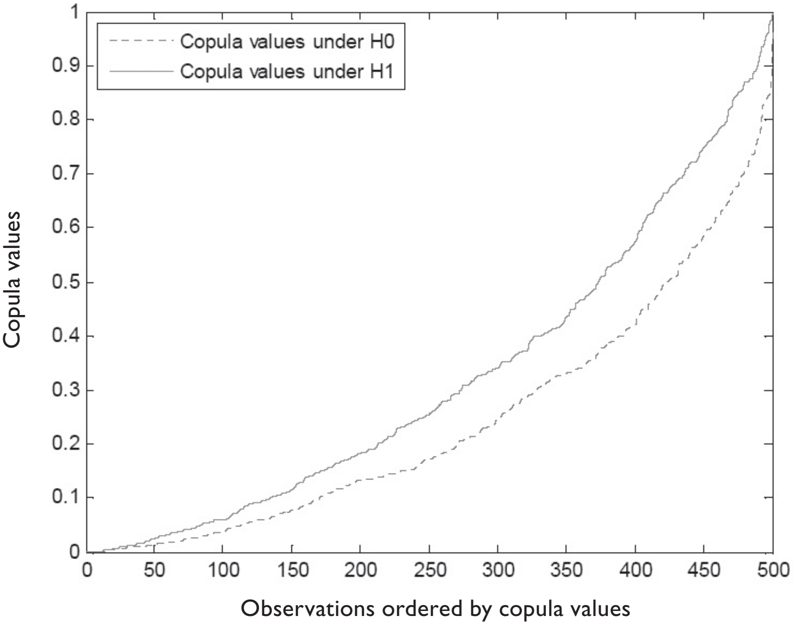
[17] We use random $U(0,1)$ inputs rather that the empirical marginals because we are specifically interested here in evaluating the copula on the assumption that it has correct inputs. Were we to use the empirical marginals as inputs, we would be jointly testing the hypothesis that the empirical marginals are correct as well as testing the hypothesis that the fitted copula is correct. This would be appropriate if we wanted a single test of the multivariate density forecasts as a whole.

[18] One way to carry out the calculations is as follows. We first simulate sets of H0 and H1 copula values. We then create a do loop for each i from 1 to n, where n is the sample size. Inside this loop (and using obvious MATLAB notation) we then create the variable p(i)=length(H1_copula_values(H1_copula_values<=H0_copula_values(i))))/n. To avoid difficulties with extreme endpoints in the range $(0,1)$, it is prudent to give any such values arbitrary values that are close (but not too close) to 0 or 1 respectively: this ensures that the subsequent Berkowitz transformation can be implemented without the danger of it breaking down due to the computer trying to take the standard normal inverse of 0 or 1.

[19] We assume throughout that there is a single unchanging copula appropriate to the whole of our sample period, ie, so the dependence between the two random variables does not change. This assumption can be relaxed, but doing so would make the discussion in the text more complicated.

[20] This copula is chosen to illustrate the method and also fits in with the general theme of this paper, ie, how to transform our series so that we can apply tests for iid $N(0,1)$. However, in practice, it will sometimes be possible to apply easier methods. For example, if we are prepared to accept that the type of copula is correct, then in many cases we would be dealing with a single-parameter copula and the problem of copula evaluation would boil down to testing whether its parameter is correctly estimated. This is equivalent to testing a correlation estimate, and in such cases we might prefer to resort to an off-the-shelf correlation test instead. In the case of the example in the text, if r is the sample Pearson correlation

FIGURE 1 Plots of copula values under H0 and H1

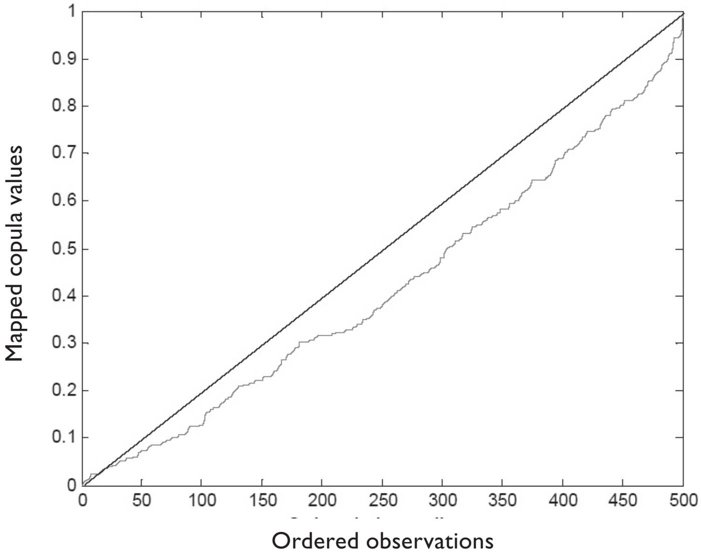


Based on 500 randomly generated values for a bivariate Gaussian copula with iid $U(0,1)$ marginals, where the (false) null hypothesis postulates $\rho = 0$ and the (true) alternative is $\rho = 0.5$.

The values of these two copulas are plotted in Figure 1: this shows that the values of the H0 copula lie well below those of the H1 copula, which indicates that the two copulas differ considerably. (This finding is as we would expect given our knowledge that the true value of ρ (0.5) is greater than the estimated one (0).) Figure 2 gives a plot of the H0 copula values mapped to the CDF values of the H1 copula. Were the fit a good one, this plot should be close to the 45° line. However, Figure 2 shows that the plot is well below this line, which again indicates that the estimated copula does not provide a good fit. Figure 3 then gives a histogram of the Berkowitz-transformed series, and Table 1 presents the sample moments of this series as well as the results of three illustrative tests (ie, a t-test of the mean prediction, a variance ratio test of the variance prediction, and a Jarque–Bera test of the skewness and kurtosis predictions). Figure 3 and the sample moments in Table 1 clearly show that the z are not distributed as standard normal. This conclusion is confirmed by the results of the mean and variance tests, which show that the predictions $\mu = 0$ and $\sigma = 1$ are both decisively rejected, though the Jarque–Bera

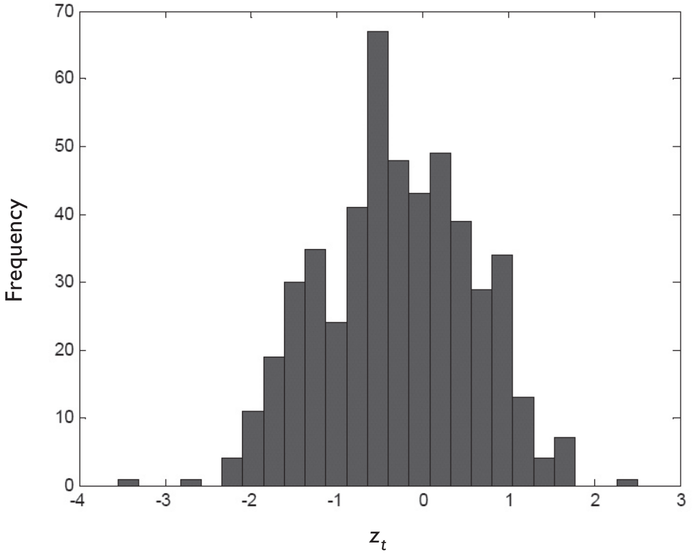
coefficient and this is predicted to be 0 under the null, then the test statistic $r\sqrt{n-2}\big/\sqrt{1-r^2}$ is distributed under the null as a t with $n-2$ degrees of freedom and we can apply a simple t-test. The method outlined in the text involves more work, but unlike the case with a simple correlation test, can also be applied to any copula we might encounter.

FIGURE 2 Plot of copula values mapped to their CDF values



Based on 500 randomly generated values for a bivariate Gaussian copula with iid $U(0,1)$ marginals, where the (false) null hypothesis postulates $\rho = 0$ and the (true) alternative is $\rho = 0.5$. The H0 copula values are mapped to their CDF values in terms of the H1 copula.

FIGURE 3 Histogram of copula values mapped to their CDF values.



Based on 500 randomly generated values for a bivariate Gaussian copula with iid $U(0,1)$ marginals, where the (false) null hypothesis postulates $\rho = 0$ and the (true) alternative is $\rho = 0.5$. The H0 copula values are mapped to their CDF values in terms of the H1 copula and then put through a standard inverse normal (Berkowitz) transformation.

TABLE 1 Sample moments of Berkowitz-transformed copula values.

	Sample moments			
	Mean	Standard deviation	Skewness	Kurtosis
$n = 500$	0.2908	0.8784	0.1103	2.7998
Test results				
t-test of zero mean:	Test stat = –7.4026	df = 499	Prob-value $\approx$ 0	
Variance ratio test of unit std:	Test stat = 385.0517	df = 499	Prob-value = 0.0001	
Jarque-Bera test of normality:	Test stat = 1.8495	df = 2	Prob-value = 0.3966	

The first row gives sample moments for based on 500 randomly generated values for a bivariate Gaussian copula with iid $U(0, 1)$ marginals, where the (false) null hypothesis postulates $\rho = 0$ and the (true) alternative is $\rho = 0.5$.

test result indicates that we cannot reject the hypothesis that z is normal. Our analysis therefore leads to the (correct) conclusion that the estimated (H0) copula is not consistent with the hypothesized (H1, true) copula.

6 Multiple models at different business-unit levels

A fourth application is where we have information on multiple models at different business-unit levels: for example, we might have information on the density forecasts and returns realized on different portfolios within an institution, where some of these portfolios make up the constituent components of others, and we might wish to evaluate these risk models both individually and collectively.

Evaluating these models individually is very straightforward – we evaluate the different models on a stand-alone basis using one of the approaches already set out – so our real focus of interest is now on how on to evaluate them collectively, and in so doing evaluate the adequacy of the process by which risk forecasts at the lower business unit level are aggregated to give us risk forecasts at the higher level. In this context, it is helpful to think of the lower-level models as providing univariate density forecasts, and to think of the higher-level model as providing multivariate ones. Recall that the previous section explained that a multivariate cumulative density function can be represented as a set of marginal univariate density forecasts combined with a copula function.

Suppose for the sake of argument that we have sufficient information to test the risk models at the lower business-unit level and that these models "pass" their evaluation tests. We are then interested in evaluating the firm's risk model at the higher business-unit level. Such an evaluation would tell us something not only about the adequacy of the higher level risk model, but also something about the adequacy of the way in which the firm is aggregating its risks. Now consider two possible scenarios:

(a) The firm might give copula forecasts that describe the inter-dependence between the risk factors at the lower business-unit levels. Given that the risk models at lower business units have passed their evaluation tests, evaluating the copula gives us an implicit test of the adequacy of the risk model at the higher business-unit level. In this case, we have an explicit evaluation of the copula and an implicit evaluation of the higher-level risk model.

(b) The firm might provide stand-alone information (ie, realized returns and density forecasts) for the risk forecasting model at the higher level. In this case, the evaluator can explicitly evaluate the higher-level model on a stand-alone basis using the same approaches used to evaluate the lower level models. If the higher-level model passes, we can infer that the risk-aggregation model (or copula) by which we move from the lower to the higher business levels must be a sound one, ie, we have an explicit evaluation of the higher-level risk model and an implicit evaluation of the copula.

In practice, it is possible that both, one or neither of these conditions might apply. If both (a) and (b) apply, then we can carry out both of these evaluation tests, and each of these gives a cross-check against the other: if the copula passes, then the higher-level risk model should pass too, and vice versa. If either (a) or (b) applies, then we can evaluate either the copula or the risk model, but not both; and if neither (a) nor (b) applies, then we can evaluate neither the copula nor the aggregate risk model. In this latter case, we must usually make do with the evaluation of the stand-alone models at the lower business-unit level.[21]

7 Dealing with parameter uncertainty

A final complication involves parameter uncertainty. We have assumed hitherto that the parameters of the "true" return process (or processes) are known to the modeler at the time the density forecasts were made. However, in practice, the modeler is unlikely to know the "true" return parameters and in such cases he/she will have to work with estimates of them instead.[22]

[21] Nevertheless, one possible response is to rely on ad hoc assumptions. For instance, we might assume (as is often done with regulatory VAR models) that correlations take the values plus or minus 1. Such assumptions make it very straightforward to infer the aggregate risk model from the lower level models because risks become additive. However, such assumptions have to be made with care, because they are extreme, and when interpreting results we have to keep in mind that we are testing these maintained assumptions as well as any information in the models themselves: a "fail" result might merely indicate that it is the additional assumptions that we have made that are empirically untenable.

[22] Whether parameter uncertainty is a significant problem in this context is a controversial issue. On the one hand, Diebold and Mariano (1995), Diebold *et al* (1998) and Berkowitz (2001) among others make persuasive arguments that the effects of parameter uncertainty are often likely to be of second-order importance. On the other, West (1996) and West and McCracken (1998) are less sanguine and suggest that parameter uncertainty can be a major problem. Perhaps the safest claim one can make is that the impact of parameter uncertainty depends on the problem at hand. As the text below suggests, with "well-behaved" problems and large sample sizes then parameter uncertainty will often be little more than a minor nuisance, but where conditions are less favorable we cannot rule out the possibility that parameter uncertainty might have an impact on the reliability of backtesting procedures.

7.1 Impact of parameter uncertainty with "best" parameter estimators

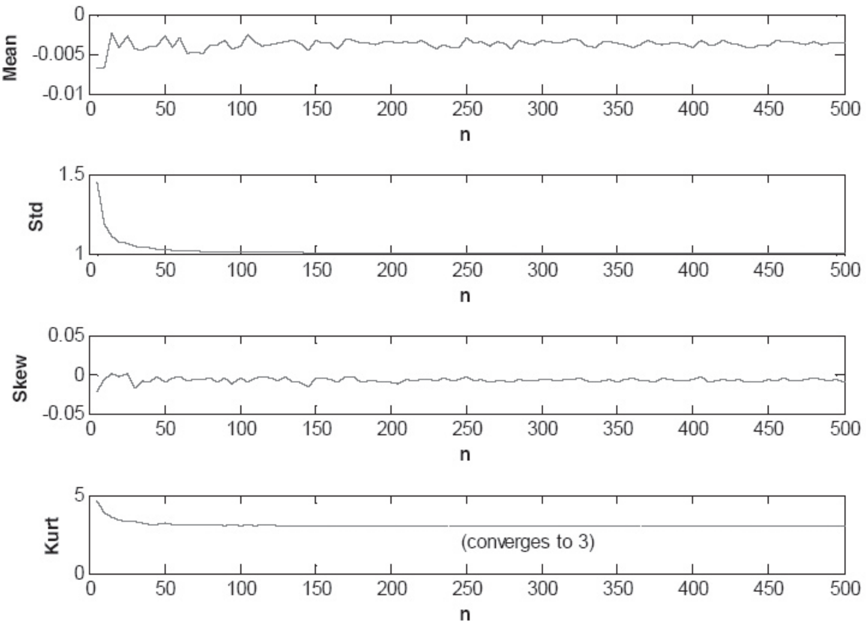
Suppose then that we relax the assumption that the parameters are known and replace it with the weaker assumption that the modeler has "best" estimators of them instead. This assumption implies that we are no longer working with true values of μ and σ, but are instead working with estimates m and s that satisfy (or nearly satisfy) the best possible sampling properties. We take "best" sampling properties to be those of standard normal samples, viz:

❑ If s is the sample estimate of the standard deviation σ from an underlying normal distribution, then the variance ratio $(n-1)s^2/\sigma^2$ is distributed under the null as a χ^2_{n-1}; in our case $\sigma = 1$ so the sample variance s^2 is distributed as $\chi^2_{n-1}/(n-1)$.

❑ Similarly, $m\sqrt{n}/s$ is distributed as a t with $n-1$ degrees of freedom, so m is distributed as $s/\sqrt{n}$ times such a t distribution.

In principle, the fact that the parameters are uncertain will affect the distribution of z_t. Some idea of this impact is illustrated by Figure 4. This shows plots of the moments of z_t against sample size, where each sample moment is simulated

FIGURE 4 Predicted moments of z_t in the presence of parameter uncertainty.



Based on 50,000 Monte Carlo simulation trials. z_t is drawn from a N(m,s), where m is drawn from a Student t with $n-2$ degrees of freedom, and $(n-1)s^2$ is drawn from a χ^2_{n-1}. n is the sample size, and "mean", "std", "skew" and "kurt" refer to the mean, standard deviation, skewness and kurtosis of z_t.

under parameter uncertainty as just described.[23] These plots show that the impact of parameter uncertainty tends to diminish as n gets larger, and is small for any "reasonable" sample size.[24] The effect of parameter uncertainty is particularly small on the even moments, and this strongly suggests that tests based on the even moment predictions are very robust to parameter uncertainty. The effect of parameter uncertainty is somewhat larger on the odd moment predictions, but it is still fairly small for "reasonable" sample sizes. The overall impression is, therefore, that parameter uncertainty seems to have little quantitative effect on the distribution of z_t provided we have good parameter estimators and a "reasonable" sample size. Under such circumstances our earlier iid $N(0, 1)$ testing strategy is then still valid even in the presence of parameter uncertainty.[25]

7.2 Impact of parameter uncertainty with less-than-"best" parameter estimators

In circumstances where we are less confident in the "optimality" of our parameter forecasts, we can sometimes resort to simulated z_t distributions. For example, we might use Monte Carlo to simulate the distribution of the parameters in the "underlying" model,[26] and then obtain simulated values using simulated values of these parameters. We might then estimate the moments of the simulated z_t distribution and test whether the moments of the empirical z_t distribution are consistent with the simulated moments.

Even where parameter uncertainty might have a "significant" effect on the distribution, we can sometimes still apply our iid $N(0, 1)$ strategy provided (as in the paragraph above) that we have some means of estimating the predicted moments of this distribution under conditions of parameter uncertainty. To illustrate, suppose that we have estimated these moments to be $\hat{\mu}, \hat{\sigma}$, and so on, and we believe that we can ignore the higher moments, ie, so we assume that the distribution is still normal or near normal. This means that the moment-matched series

$$\hat{z}_t = (z_t - \hat{\mu})/\hat{\sigma} \tag{17}$$

[23] These moments are calculated using Monte Carlo simulation: for each of our chosen values of n, we first simulate values for m and s using the relevant distributions; we then simulate corresponding values of X_t from a standard normal, map each one to the value of a normal CDF with mean m and standard deviation s, then take the Berkowitz transformation and estimate the sample moments.

[24] It should be borne in mind here that most risk practitioners would be working with sample sizes of at least one or two years' daily trading returns, ie, with samples of at least 250 or 500 observations: for such sample sizes, the predicted moments are not much different from those we would get under standard normality.

[25] We can also respond to the possible impact of parameter uncertainty on backtesting results by carrying out stress-tests in which we shock key parameters and repeat our analysis using the shocked parameter values. For example, we might shock key parameters by a couple of standard errors, and check the robustness of our results to these shocks.

[26] Some examples are Dowd *et al* (2006), who simulate their parameters from a combination of Wishart and normal distributions, and Bams *et al* (2005), who simulate the parameters of GARCH models with various (eg, Student-t and stable Paretian) noise processes.

will then be distributed as iid $N(0, 1)$, provided no parameter estimation "noise" creates any temporal dependence. Thus, even the problem of validating models with "significant" parameter risk can – at least sometimes – be handled within our iid $N(0, 1)$ backtesting strategy.

One final point: if we have doubts about our parameter estimators (as is often likely to be the case in practice) it would sometimes be useful to compare the results of such an exercise with those obtained in Section 7.1 under the assumption that we have "best" parameter estimators. The reason for this is that any substantial differences can only be ascribed to weaknesses in the parameter estimation processes. So, if tests based on those in Sections 7.1 and 7.2 produce similar results, we can conclude that any problems in our density forecasts are not due to incorrect parameter forecasts (ie, and must therefore be due to incorrect distributional assumptions predicated on presumably correct parameters); conversely, if tests based on those in Sections 7.1 and 7.2 are significantly different, then we can conclude that there are inadequacies in our parameter forecasting methods, regardless of any problems that might exist with the distributional assumptions predicated on these parameter estimates. Tests like these therefore provide useful information, because they can tell the model evaluator where to direct his/her attention, ie, to the distributional assumptions, to the parameter forecasts or possibly to both.

8 Conclusions

This paper extends the basic Berkowitz backtesting strategy to encompass a variety of important complications that often arise in real-world backtesting problems. Whereas his original framework was couched in the simple context of a one-step ahead univariate density forecasting model with no parameter uncertainty, many backtesting problems involve complications such as multi-step or multivariate density forecasts, and the additional problems created by risk aggregation and possible parameter uncertainty. Our discussion suggests that the Berkowitz backtesting strategy – that is, the strategy of transforming the data so that they become iid standard normal under the null and then testing them for iid standard normality – can still be applied in the presence of such complicating factors, and is therefore more powerful and more useful than has hitherto been appreciated. The transformations become more involved, but the strategy itself remains the same. This suggests that backtesting is not as difficult as it sometimes appears to be: sometimes the "trick" to backtesting is not to develop sophisticated new tests, as such, but to look for ways to simplify backtesting problems so that we can apply simple tests to them, and the simplest tests of all are those of iid $N(0, 1)$.[27]

[27] Naturally, the extensions suggested here are only indicative, and one suspects that the same strategy could be made to work in other backtesting situations. Perhaps the most obvious extensions are to handle the additional complexities associated with liquidity risks, credit risks, and operational risks. More work needs to be done to show how the strategy can be extended to these types of backtesting problem, and also to establish the limits of this strategy, ie, can we establish the circumstances where it will not work?

REFERENCES

Bams, D., Lehnert, T., and Wolf, C. P. P. (2005). An evaluation framework for alternative VAR models. *Journal of International Money and Finance* **24**, 944–58.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics* **19**, 465–74.

Blix, M., and Sellin, P. (2000). A bivariate distribution for inflation and output forecasts. Sveriges Riksbank Working paper no. 102.

Brock, W. A., Dechert, W. D., and Scheinkman, J. (1987). A test for independence based on the correlation dimension. University of Wisconsin-Madison SSRI Workshop Paper 8702.

Christoffersen, P. F., and. Pelletier, D. (2004). Backtesting value-at-risk: a duration-based approach. *Journal of Financial Econometrics* **2**, 84–108.

Crnkovic, C., and Drachman, J. (1996). Quality control. *Risk* **9** (September), 139–43.

David, F. N. (1947). A power function for tests of randomness in a sequence of alternatives. *Biometrika* **34**, 335–9.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**, 863–83.

Diebold, F. X. Hahn, J., and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics* **81**, 661–873.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253–63.

Dobrić, J., and Schmid, F. (2005). Testing goodness of fit for parametric families of copulas – application to financial data. *Communications in Statistics: Simulation and Computation* **34**, 1053–68

Dowd, K. (2005a) *Measuring Market Risk*. Second edition. Chichester: John Wiley and Sons.

Dowd, K. (2005b). Validating multiple-period density forecasting models. *Journal of Forecasting*. Forthcoming.

Dowd, K. (2006). A framework for the evaluation of fitted copulas. Working paper, Centre for Risk and Insurance Studies, Nottingham University Business School.

Dowd, K., Cairns, A. J. G., and Blake, D. (2006). Mortality-dependent measures of financial risk. *Insurance: Mathematics and Economics* **38**, 427–40.

Durlauf, S. N. (1991). Spectral based testing of the martingale hypothesis. *Journal of Econometrics* **50**, 355–76.

Engle, R. F., and Manganelli, S. (2004). CAViaR: Conditional autoregressive Value-at-Risk by regression quantiles. *Journal of Business and Economic Statistics* **22**, 367–81.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis* **95**, 119–52.

Patton, A. J. (2005). Modelling asymmetric exchange rate dependence. *International Economic Review* **47**, 527–56.

Savu, C., and Trede, M. (2004). Goodness-of-fit tests for parametric families of Archimedean copulas. Working paper, University of Münster.

Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leur merges. *Publ. Inst. Stat. Univ. Paris* **8**, 229–31.

Tsay, R. S. (2005) *Analysis of Financial Time Series*. Second edition. Wiley Interscience, Hoboken, NJ.

West, K. D. (1995). Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–84.

West, K. D., and McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review* **39**, 817–40.

Wolfowitz, J. (1943). On the theory of runs with some applications to quality control. *Annals of Mathematical Statistics* **14**, 280–8.