Evaluating Multidimensional Value at Risk

Arnold Polanski, *Queen's University Belfast*

Evarist Stoja, *University of Bristol*

February 2011

Abstract

We propose two simple evaluation methods for time-varying density forecasts of continuous higher-dimensional random variables. Both methods are based on the probability integral transformation for unidimensional forecasts. The first method tests multinormal densities and relies on the rotation of the coordinate system. The advantage of the second method is not only its applicability to arbitrary continuous distributions but also the evaluation of the forecast accuracy in specific regions of its domain as defined by the user's interest. We show that the latter property is particularly useful for evaluating a multidimensional generalization of the Value at Risk. In simulations and in an empirical study, we examine the performance of both tests.

Keywords: Multivariate Density Forecast Evaluation; Probability Integral Transformation; Multidimensional Value at Risk; Monte Carlo Simulations.

JEL classification: C52; C53.

Address for correspondence: *Arnold Polanski*, Queen's University Management School, 25 University Square, Queen's University Belfast, Belfast, BT7 1NN, UK. Email: a.polanski@qub.ac.uk. *Evarist Stoja*, School of Economics, Finance and Management, University of Bristol, 8 Woodland Road, Bristol, BS8 1TN, UK. Email: e.stoja@bristol.ac.uk.

1. Introduction

Evaluation of the accuracy of forecasts occupies a prominent place in the finance and economics literature. However, most of this literature (e.g., Diebold and Lopez, 1996) focuses on the evaluation of point forecasts as opposed to interval or density forecasts. The driving force for this over-focus is that, until recently, point forecasts appeared to serve well the requirements of the forecast users. However, there is increasing evidence that a more comprehensive approach is needed. One example is Value at Risk (VaR) which is defined as the maximum loss on a portfolio over a certain period of time that can be expected with a certain probability. When returns are normally distributed, the VaR of a portfolio is a simple function of the variance of the portfolio.[1] In this case, normality justifies the use of point forecasts for the variance. However, when the return distribution is non-normal, as is now the general consensus, the VaR of a portfolio is determined not just by the portfolio variance but by the entire conditional distribution of returns. More generally, decision making under uncertainty with asymmetric loss function and non-Gaussian variables involves density forecasts (see Tay and Wallis, 2000; and Guidolin and Timmermann, 2005, for a survey and discussion of density forecasting applications in finance and economics).

The increasing importance of forecasts of the entire (conditional) density naturally raises the issue of forecast evaluation. The relevant literature, although developing at a fast pace, is still in its infancy. This is somewhat surprising considering that the crucial tools employed date back a few decades. Indeed, a key contribution by Diebold et al. (1998)

[1] When the mean return on an asset is assumed to be zero, as is commonly the case in practice when dealing with short-horizon returns, the VaR of a portfolio is simply a constant multiple of the square root of variance of the portfolio.

relies on the probability integral transformation (PIT) result in Rosenblatt (1952). Diebold et al. point out that the correct density is weakly superior to all forecasts. This suggests that forecasts should be evaluated in terms of their correctness as this is independent of the loss function. To this end, Diebold et al. (1998) employ the PIT of the univariate density forecasts which, if accurate, are $i.i.d.$ standard uniform. They measure the forecast accuracy by the distance between the empirical distribution of the PITs and the 45° line and argue that the visual inspection of this distance may provide valuable insights into the deficiencies of the model and ways of improving it. Obviously, standard goodness-of-fit tests can be directly applied to the PITs (see Noceti et al., 2003 for a comparison of the existing goodness-of-fit tests). Additional tests have been proposed by Anderson et al. (1994), Li (1996), Granger and Pesaran (1999), Berkowitz (2001), Li and Tkacz (2001), Hong (2001), Hong and Li (2003), Bai (2003) and Hong et al. (2007).

The existing evaluation methods of the multidimensional density forecasts (MDF) rely on the advances made in the univariate case. Diebold et al. (1999) extend the PIT idea to the multivariate forecasts by factoring the multivariate probability density function (PDF) into its conditionals and computing the PIT for each conditional. As in the univariate case, the PIT of these forecasts is $i.i.d.$ uniform if the sequence of forecasts is correct. Clements and Smith (2000, 2002) extend Diebold et al.'s (1999) idea and propose two tests based on the product and ratio of the conditionals and marginals. While the latter tests perform well when there is correlation misspecification, they underperform the original test by Diebold et al. (1999) when such misspecification is absent. However, both approaches rely on the factorization of each period forecasts into their conditionals, which may be impractical for some applications (e.g., for numerical approximations of density forecasts). Moreover, these approaches assume that the forecasting model is correct under the null hypothesis. This assumption has important implications for the evaluation tools employed, particularly in relation to parameter estimation uncertainty. Recognising this issue, another strand of MDF evaluation literature has recently gained momentum. This literature allows for dynamic misspecification and/or parameter estimation uncertainty and includes important contributions by Corradi and Swanson (2006b), Bai and Chen (2008), Chen and Hong (2009), *inter alia*. Corradi and Swanson (2006b) construct Kolmogorov-type conditional distribution tests in the presence of both dynamic misspecification and parameter estimation uncertainty. While their testing framework is flexible, it suffers from the fact that the limiting distribution is not nuisance parameters free and bootstrapping is needed to obtain valid critical values. Bai and Chen (2008) and Chen and Hong (2009) propose MDF evaluation tests that, under certain conditions, deal with the parameter estimation uncertainty. For example, Bai and Chen (2008) employ the *K*-transformation of Khmaladze (1981) to remove the effect of parameter estimation, so that a distribution-free test can be constructed. However, they still rely on the factorization of the joint density and apply this procedure only to multivariate normal and multivariate-*t* distributions, in which case they obtain closed-form results. We discuss these issues in more detail in Section 3 and refer the interested reader to Mecklin and Mundfrom (2004) and Corradi and Swanson (2006a) for further insights into density forecast evaluation.

Broadly speaking, this paper belongs to the literature established by Diebold et al. (1998, 1999) and Clements and Smith (2000, 2002) which does not account for parameter estimation uncertainty. This approach also dominates in the parametric-VaR area of the risk management literature, in which we are mainly interested (see, for example, Gourieroux and Jasiak, 2010). Thus, in simulations and empirical examples, we ignore parameter estimation uncertainty and potential dynamic misspecification but we acknowledge that these could be important. Finally, we stress that forecasts may vary over time making parameter estimation and forecast evaluation based on the laws of large numbers unfeasible.

This paper makes two important contributions. Firstly, it proposes two new tests to evaluate multidimensional, time-varying density forecasts which although – similarly to its counterparts – may suffer from parameter estimation error and dynamic misspecification, are nevertheless simpler and more flexible. Secondly, to the best of our knowledge it is the first to formalise and propose a theoretical framework to testing the accuracy of multidimensional VaR (MVaR). This framework is particularly important for examining multiple sources of tail risk.

The outline of the remainder of this paper is as follows. In Section 2, we discuss an evaluation procedure for multinormal density forecasts. Section 3 presents a test for arbitrary continuous densities while Section 4 discusses the results of Monte Carlo simulations and an empirical application for the newly proposed tests. Finally, Section 5 concludes.

2. Evaluation Procedure for Multinormal Density Forecasts

Rosenblatt (1952) showed that for the cumulative distribution function (CDF) $\widehat{F}_t$ (PDF $\widehat{f}_t$), which correctly forecasts the true data generating process (DGP) F_t of the observation x_t, i.e., for which $\widehat{F}_t(x_t)= F_t(x_t)$, the PIT

$$z_t = \int_{-\infty}^{x_t} \widehat{f}_t(u)du = \widehat{F}_t(x_t)$$

is $i.i.d.$ according to $U[0,1]$. Therefore, the adequacy of forecasts can be easily evaluated by examining the z_t series for violations of independence and uniformity.

The PIT idea is extended to the multivariate case by Diebold et al. (1999). Their test procedure (D-test hereafter) factors each period MDF into the product of the conditionals

$$\widehat{f}_{t-1}(x_{1t}, x_{2t},..., x_{Nt}) = \widehat{f}_{t-1}(x_{Nt} \mid x_{1t}, x_{2t},..., x_{N-1,t}) \cdots \widehat{f}_{t-1}(x_{2t} \mid x_{1t}) \cdot \widehat{f}_{t-1}(x_{1t})$$

and obtain the PIT for each conditional distribution, producing a set of N z_t-series, which are $i.i.d.$ $U[0,1]$ individually and as a whole whenever the MDF is correct.[2] Rejecting the null of $i.i.d.$ $U[0,1]$ for any, as well as the combined z_t series implies that the MDF is misspecified. Clements and Smith (2000, 2002) propose two tests (CS-tests hereafter)

---

[2] There are N! different ways to factor the MDF $\widehat{f}_t(x_{1,t-1},..., x_{N,t-1})$, giving us a wealth of z series with which to evaluate the forecast.

based on the product (CS1) and the ratio (CS2) of PITs for the conditionals and marginals, where the N-dimensional vector of scores has typical elements $z_t^j = z_{2|1,t}^c \cdot z_{1,t}^m$ and $z_t^j = z_{2|1,t}^c / z_{1,t}^m$ respectively.

For a multinormal density forecast, we describe below a test (MN-test hereafter) that avoids the possibly cumbersome factorization of the MDF. Instead, we transform the coordinate system according to a linear transformation composed of a translation and a rotation and compute the PITs for each marginal distribution. Note that the standard multinormality tests (e.g., Cox and Small, 1978; Smith and Jain, 1988) do not apply for time-varying distributions.

Specifically, let $X_t = (X_{1,t},..., X_{N,t})$ represent an N-dimensional multinormal random variable with mean μ_t and the variance-covariance matrix Σ_t. The null hypothesis assumes that the MDF $\widehat{F}_{t-1}$ is the same as the true distribution F_t of X_t and we do not distinguish between these functions in what follows,

$$H_0: \quad \{\widehat{F}_{t-1}\}_{t=1}^T \text{ is the true DGP,} \qquad H_A: \{\widehat{F}_{t-1}\}_{t=1}^T \text{ is not the true DGP.}$$

It is well known that the random variable $\widetilde{X}_t = R_t(X_t - \mu_t)$, where R_t is the matrix of eigenvectors of Σ_t, is multinormal with mean zero and a diagonal variance-covariance matrix $\widetilde{\Sigma}_t = R_t \Sigma_t R_t^T$. Since X_t is multinormal, $\widetilde{X}_t$ is a collection of independent univariate variables with marginal distributions $\tilde{F}_{1,t},...,\tilde{F}_{N,t}$, $\tilde{F}_{i,t}$ ~ $N(0, \tilde{\Sigma}_t(i,i))$. Moreover, the null hypothesis that the observations x_t are drawn from X_t is equivalent to the hypothesis that the transformed observations $\tilde{x}_t = R_t(x_t - \mu_t)$ are drawn from $\tilde{X}_t$. From the results in Rosenblatt (1952) and by the independence of the components of $\tilde{X}_t$ follows that under the null, the scores $\tilde{z}_{i,t} = \tilde{F}_{i,t}(\tilde{x}_{i,t})$, $i = 1,...,N$, are independently and uniformly[3] distributed on $[0,1]^N$ individually and as a whole. The null can then be tested by the standard tests of uniformity (see Noceti et al., 2003) and independence (see Brock et al., 1991). In the next section, we define the test statistic that we use in our simulations and empirical studies and show also that linear transformations re-emerge as a useful tool in a test that does not rely on the normality of the forecasts.

3. Evaluation Procedure for Arbitrary Continuous MDFs

The test introduced in this section (Q-test hereafter) fulfils two purposes. On the one hand, it is a simple, ready-to-use procedure to evaluate arbitrary continuous MDFs. On the other hand, it allows for focusing on a specific region of the MDF instead of examining it over its entire domain. As we shall explain later in this section, existing tests can then be used to verify the region-specific accuracy of the forecasts. The latter application is particularly interesting from a risk-management perspective. Risk managers and regulators are interested, generally, in the likelihood of large losses, i.e. in a specific tail of the distribution. If this is the case, then, a model superior in forecasting the central part of the distribution will be eschewed in favour of another model which accurately forecasts the

[3] This will be the case when all variables in $\tilde{X}_t$ are not degenerated. Otherwise, we use only variables with positive variance to compute the scores.

tails. This objective motivates the censored likelihood test of Berkowitz (2001), in which the observations not falling into the negative tail of the distribution (with the cut-off point being decided by the user's requirements) are truncated.

As the Q-test is based on the PIT computation, we show first in a simple example that for a correct MDF $\widehat{F}_{t-1}$, the PITs $\widehat{F}_{t-1}(x_t)$ are not necessarily uniformly distributed. For the standard binormal $\widehat{F}_{t-1}(x_t)$, it is straightforward to compute that the probability mass of the contour area $\{y \in R^2 : \widehat{F}_{t-1}(y) < 0.025\}$ is 0.117. Thus, under this distribution, the probability of obtaining a score $z_t = \widehat{F}_{t-1}(x_t)$ < 0.025 is 0.117 rather than 0.025 as would be the case if z_t were uniformly distributed. It follows that, generally, the multidimensional extension of the PIT does not produce uniformly distributed scores. However, a simple modification in the PIT computation restores the uniformity. First, we transform the series $x = \{x_t\}_{t=1}^{T}$ into $x_t^M = Max\{x_{1,t},....,x_{N,t}\} \cdot (1,....,1)$ and then compute the scores $z_t^M = \widehat{F}_{t-1}(x_t^M)$. Instead of the original observation x_t, we use for the computation of the PIT the projection of the largest coordinate of x_t on the main diagonal along the vector perpendicular to the corresponding axis (see Figure 1). Note that for unidimensional forecasts, our procedure reduces to the traditional PIT. In the appendix, we prove the following result.

Proposition 1: If $\{\widehat{F}_{t-1}\}_{t=1}^{T}$ is the true DGP for the sequence $\{x_t\}_{t=1}^{T}$, then $\{z_t^M = \widehat{F}_{t-1}(x_t^M)\}_{t=1}^{T}$, $x_t^M = Max\{x_{1,t},....,x_{N,t}\} \cdot (1,....,1)$, is $i.i.d.$ according to the uniform distribution $U[0,1]$.

For an intuition of the proof of Proposition 1, we focus on two-dimensional orthants (quadrants) $Q((v,v)) = \{y \in R^2 : y \le (v,v)\}$, $v \in R$, as illustrated by the dark gray rectangle in Figure 1.[4] The crucial observation is that for any point x_t inside (outside) of the quadrant $Q((v,v))$, x_t^M lies also inside (outside) of $Q((v,v))$. In other words, $x_t \le (v,v)$ implies $x_t^M \le (v,v)$ and $x_{i,t} > v$ for at least one i implies $x_t^M > (v,v)$. As a consequence, the probability of obtaining a score $z_t^M = \widehat{F}_{t-1}(x_t^M)$ below $\widehat{F}_{t-1}((v,v))$ is equivalent to the probability of x_t lying in $Q((v,v))$, i.e., it is equal to $\widehat{F}_{t-1}((v,v))$.

[Figure 1]

Importantly, the proposed procedure effectively transforms a MDF $\widehat{F}_{t-1}$ into a unidimensional random variable $Z_t^M = \widehat{F}_{t-1}(X_t^M)$, $X_t^M = Max\{X_{1,t},....,X_{N,t}\} \cdot (1,....,1)$. Due to the Max{.} operator, each realization z_t^M of Z_t^M exploits the information in the entire multidimensional observation x_t. The forecast $\widehat{F}_{t-1}$ is then deemed correct whenever the proportion of observations that fall into each orthant $Q((v,...,v))$ approximates the probability of this orthant under $\widehat{F}_{t-1}$. In particular, the Q-test allows for assessing the accuracy of the forecasts in the "negative tail" of the distribution, as illustrated in the application to risk management later in this section.

[4] Strictly speaking, the set $Q((v,...,v)) = \{y \in R^N : y \le (v,...,v)\}$ is an orthant in the coordinate system centred at $(v,...,v)$. Due to the importance of orthants (quadrants), we call our procedure the Q-test**.**

Hypotheses

Proposition 1 leads to a testable null hypothesis that the series $\{\widehat{F}_{t-1}\}_{t=1}^{T}$ is the actual DGP against the alternative that it is not.

H_0: $\{\widehat{F}_{t-1}\}_{t=1}^{T}$ is the true DGP, H_A: $\{\widehat{F}_{t-1}\}_{t=1}^{T}$ is not the true DGP.

As the z_t^M-scores are $i.i.d.$ uniform under the null, the standard tests of uniformity apply (see Noceti et al., 2003). The assumption that z_t^M-scores are $i.i.d.$ can be tested by simply examining correlograms of various powers of z_t^M (see Brock et al., 1991 for other, formal tests). However, when applied to the dynamic estimation of the MVaR, standard tests from the VaR literature on risk management (Kupiec, 1995 and Christoffersen, 1998) appear to be the suitable choice. In most of our examples, we focus on testing uniformity of the scores with the Pearson's X^2-statistic

$$X^2 = \sum_{k=1}^{K} (n_i - T/K)^2 / (T/K)$$

where n_i is the number of z_t^M-scores in the interval [(*i*-1)/*K*, *i/K*], *i* = 1,…,*K*. By the Pearson-Fisher theorem, X^2 is distributed asymptotically as χ^2 with *K*-1 degrees of freedom when no parameters are estimated from the data. We discuss the case with estimated parameters below. In any case, the rejection of either uniformity or independence, or both, provides statistical evidence that $\{\widehat{F}_{t-1}\}_{t=1}^{T}$ is not the true DGP. On the other hand, failure to reject the null implies that we cannot dismiss the hypothesis that the MDF model under examination is correct.

We note here that our test is not fully consistent as it cannot detect specific departures from the null.[5] For example, a bidimensional density that mirrors the density under the null across the main diagonal will generate the same z_t^M -scores. However, this inconsistency is not unique to our test (see, for example, Bai and Chen, 2008).[6] Moreover, although clearly important for MDF evaluation, the lack of full consistency is less important for MVaR evaluation. The "mirror" bidimensional density in the example above will generate identical z_t^M -scores as the true MDF and will lead to identical conclusions on MVaR accuracy that are correct for both, the true and the "mirror" distribution.

Unlike other tests for time-varying MDFs (e.g. Diebold et al., 1999; Clements and Smith, 2002; Bai and Chen, 2008), the Q-test generates only one score for each observation $x_t = (x_{1,t},...,x_{N,t})$. Although this parsimony may lead to loss of information, we illustrate in one of our simulations (Table 2) the opposite effect of concentrating the evidence from the sample. Importantly also, a single score helps to circumvent the potential problems related to interdependence of multiple scores that are computed from the same observation (see, for example, Bai and Chen, 2008).

[5] We thank an anonymous referee for pointing this out.

[6] As we noted earlier, under a rotation of the coordinate system, the z_t^M -scores remain uniform for the true MDF but it is unlikely that they will remain uniform for an alternative distribution. Therefore, computing scores under sufficiently many rotations can re-establish consistency. Our Monte Carlo simulations suggest that at most *N* rotations, where *N* is the number of dimensions of the MDF, are generally sufficient to obtain the correct answer.

Multidimensional Value at Risk

In a market with N assets, an investor is interested in the event E that the random return of each asset falls below a certain value v. Equipped with the forecast $\widehat{F}_{t-1}$, the investor can compute v_t such that $\widehat{F}_{t-1}((v_t,...,v_t)) = \alpha$, i.e., such that the event E is expected to occur with probability α. If the value of v_t is negative, the investor can compute the loss due to the event E for any portfolio of long positions.

The rationale in this example lies at the heart of the concept of Value at Risk (VaR) which is now one of the most widely used risk measures among practitioners, largely due to its adoption by the Basel Committee on Banking Regulation (1996) for the assessment of the risk of the proprietary trading books of banks and its use in setting risk capital requirements (see Jorion, 2000). For the unidimensional CDF $\widehat{F}_{t-1}$, the VaR at the coverage level 1-α is the quantile v_t for which $\widehat{F}_{t-1}(v_t) = \alpha$. Generalizing this definition to the MDF $\widehat{F}_{t-1}$, we require that the MVaR $(v_t,...,v_t)$ satisfies the condition $\widehat{F}_{t-1}((v_t,...,v_t)) = \alpha$.[7] From the definition $z_t^M = \widehat{F}_{t-1}(x_t^M)$ follows immediately that z_t^M is less than α whenever all components of the observation $x_t = (x_{1,t},...,x_{N,t})$ fall below (exceed) the critical value v_t,

$$z_t^M < \alpha \Leftrightarrow x_{i,t} < v_t \text{ for all } i = 1,\ldots,N$$

[7] Asymmetric specifications of MVaR, $\widehat{F}_{t-1}((v_{1,t},...,v_{N,t})) = \alpha$, where $v_{1,t} \neq ... \neq v_{N,t}$, are also possible. Investigating these alternatives is beyond the scope of this paper.

The latter property has important consequences when assessing the MVaR forecasts (the density forecasts for an orthant $Q((v_t,...,v_t))$). For a sufficiently large number of observations, we can compute the proportion of scores that exceed the MVaR (the proportion of observations that fall into $Q((v_t,...,v_t))$), and compare this number to the nominal significance level α. We refer to this procedure as unconditional accuracy. On the other hand, the conditional accuracy requires that the number of scores that exceed the MVaR forecast should be unpredictable when conditioned on the available information (i.e., the MVaR violations should be serially uncorrelated). To assess both types of accuracy, we can resort to the unconditional accuracy test of Kupiec (1995) and the conditional accuracy test of Christoffersen (1998). Although both tests are designed for testing the univariate VaR accuracy, they still apply for our purposes because the Q-test effectively converts a MDF into a univariate score variable.

In the context of the last example, the MVaR is a suitable instrument of risk measurement for situations of joint losses incurred by long positions in N assets. If, however, the investor contemplates also (some) short positions, she will be interested in the joint risk of positive (and negative) returns. In other words, the investor will be interested in the appropriate orthant which combines negative returns for the long positions and positive returns for the short positions. The accuracy of the density forecasts for areas other than the "negative orthant" can be assessed by transforming the canonical coordinate system. In order to compute the z_t^M-scores in the transformed system, we have to express the observations x_t and the arguments in the MDF $\widehat{F}_{t-1}$ in the new coordinates. Specifically, for a translation vector μ_t and a rotation matrix R_t, we compute $\tilde{x}_t = R_t(x_t - \mu_t)$, $\tilde{x}_t^M =$

$Max(\tilde{x}_{1,t},...,\tilde{x}_{N,t})\cdot(1,...,1)$ and $\tilde{z}_t^M = \tilde{F}_{t-1}(\tilde{x}_t^M) = \hat{F}_{t-1}(R_t^{-1}\tilde{x}_t^M + \mu_t)$. Note that under this transformation, $\tilde{F}_{t-1}$ is a CDF and the $\tilde{z}_t^M$-scores are *i.i.d.* $U[0,1]$ when $\hat{F}_{t-1}$ is the true DGP. The orthant $Q((v_t,...,v_t))$ in the transformed system corresponds then to a different area of the original $\hat{F}_{t-1}$ domain and the accuracy of the $\hat{F}_{t-1}$ in this area can be tested by the same means as in the canonical system. Figure 2 shows the example of $N = 2$ assets with means zero and MDF $\hat{F}_{t-1}$. The rotation of the coordinates clockwise by 90° relocates the south-east orthant (a positive and a negative return) in the canonical coordinates to the south-west orthant (two negative returns). The investor can, consequently, assess the MVaR under $\hat{F}_{t-1}$ for a portfolio composed of a short position in the first asset and a long position in the second asset.

[Figure 2]

The possibility of generating scores in different coordinate systems allows, potentially, for gathering abundant information on the tested MDF. Unlike the D-test and CS-tests, where various independent score series can be generated, the scores in the Q-test are not independent across transformations. Figure 3 shows the scatter plot of the scores computed under the standard binormal in the canonical (x-axis) and in the 90°-rotated system (y-axis) are dependent. For example, the canonical and the rotated scores are not less than 0.2 simultaneously.

[Figure 3]

On the other hand, the use of only one score series raises the question of the transformation that maximizes the power of the test. A simple transformation that, arguably, comes closest to this goal, projects the largest component from the principal component analysis of the covariance matrix Σ_t of $\widehat{F}_{t-1}$ on the main diagonal. This transformation can be constructed by rotating the demeaned $\widehat{F}_{t-1}$ firstly by the matrix of eigenvectors of Σ_t, and then by the matrix that rotates the axis with the largest variance to the main diagonal.

Parameter Estimation Uncertainty

It is well known that the presence of estimated parameters may complicate test inference. For example, the Kolmogorov test can be difficult to apply in the presence of estimated parameters, particularly for multivariate data with many parameters (see, for example, Bai and Chen, 2008). Following other scholars (Diebold and Mariano, 1995; Christoffersen, 1998; Diebold et al. 1998, 1999; Clements and Smith, 2000, 2002), we consider the forecasts as primitives and ignore the method employed to obtain them. In many situations this may be an acceptable practice. Firstly, many density forecasts are not based on estimated models. For example, the large-scale market risk models at many financial institutions combine estimated parameters, calibrated parameters and ad-hoc modifications that reflect the judgment of management. Another example is the density forecasts of inflation of the Survey of Professional Forecasters (see Diebold et al., 1998). Moreover, previous research suggests that parameter estimation uncertainty is of second-order importance when compared to other sources of inaccuracies such as model misspecification (Chatfield, 1993). Further, Diebold et al. (1998) find that the effects of parameter estimation uncertainty are immaterial in simulation studies geared toward the relatively large sample sizes employed in financial studies such as the present one.

When parameter estimation cannot be ignored, suitable estimators can often be found that lead to pivotal test statistics (e.g., the "super-efficient" estimators in Watson, 1958). Particularly, for time-invariant forecasts $\widehat{F}_{t-1}(x_t \mid \theta_{t-1}) = F(x_t \mid \theta)$, a suitable k-dimensional ML estimator $\widehat{\theta}$ in the Pearson's-χ^2 goodness-of-fit test is obtained by maximizing with respect to θ the multinomial distribution

$$\varphi(n_1,..., n_K \mid \theta) = \frac{T!}{n_1!...n_K!} p_1^{n_1}(\theta)... p_K^{n_K}(\theta)$$

where $p_i(\theta) = \int_{B_i} f(\omega \mid \theta) d\omega$ is the probability of the fixed hyperbox $B_i \in R^n$, i =1,…,K, under $f(x_t \mid \theta)$ and n_i is the number of observations in this hyperbox. By the Pearson-Fisher theorem, the statistic

$$X^2(\widehat{\theta}) = \sum_{k=1}^{K} (n_i - p_i(\widehat{\theta})T)^2 / (p_i(\widehat{\theta})T)$$

is distributed asymptotically as χ^2 with K-k-1 degrees of freedom (see Watson, 1958; Birch, 1964). It is straightforward to show that this test is equivalent to the test of uniformity of the scores $z_t^M = F(x_t^M \mid \widehat{\theta})$. Therefore, the X^2 test statistic which we use extensively in this work is pivotal for time-invariant densities, when the parameters are estimated from the multinomial PDF.

An important class of models comprises a time-varying hypothesised distribution with a well-defined structure on the co-evolution of the variables (e.g. VAR and GARCH models). In this case, one way of accounting for parameter estimation uncertainty is to apply the *K*-transformation (Khmaladze, 1981), which allows for the construction of a distribution-free test statistic. In principle, the *K*-transformation can be applied to the MN and Q-test along the lines of the *V*-test in Bai (2003) and Bai and Chen (2008). Its computation, however, may be cumbersome for non-standard MDFs.

Finally, in the case of arbitrary time-varying MDFs – for which our general model is particularly suited – parameter estimation is infeasible as only one observation is drawn from the MDF at each date. As such, the only practical solution is to assume that the hypothesised model is correct under the null.

4. Monte Carlo Simulations and Empirical Results

Although a comprehensive study of the statistical properties of the proposed tests is beyond the scope of this work, we performed Monte Carlo simulations, in which we compared the performance of four test procedures (D-test, CS-tests, Q-test and MN-test).

In the first experiment, we generated observations according to a mixture of two binormal distributions, i.e., at each time t, an observation was drawn from one of the distributions according to the probability weights in the mixture. Note that this experiment can be interpreted as emulating a time-varying DGP that is forecasted correctly by time-varying densities. Specifically, we used two mixtures, $\mathrm{M}_1 = ½\mathrm{N}((\text{-}\delta,\text{-}\delta),\mathrm{I}_\mathrm{N}) + ½\mathrm{N}((\delta,\delta),\mathrm{I}_\mathrm{N})$ and $\mathrm{M}_2 = ½\mathrm{N}((0,0),((1,\text{-}\delta/2),(\text{-}\delta/2,1)))$ + ½ $\mathrm{N}((0,0),((1,\delta/2),(\delta/2,1)$, where δ is interpreted as the deviation from the null hypothesis and I_N is the *N*-dimensional identity matrix The scatter plots of the representative data are reproduced in Figures 4 and 5, respectively.

[Figures 4 and 5]

For both mixtures, we tested the null hypothesis that the observations came from a binormal with mean and variance obtained from the relevant mixture,

$$\mu_1 = (0,0),\, \Sigma_1 = \begin{pmatrix} 1+\delta^2 & \delta^2 \\ \delta^2 & 1+\delta^2 \end{pmatrix}, \qquad \mu_2 = (0,0),\, \Sigma_2 = I_2$$

In order to compute the test statistic in the D-test and the CS-tests, we factor the multinormal PDF $f(x;\mu,\Sigma)$ into a product of two multinormal PDFs (a conditional and a marginal),

$$f(x;\mu,\Sigma) = f(x_1;\mu_{x_1|x_2},\Sigma_{x_1|x_2})\, f(x_2;\mu_2,\Sigma_{22}) \tag{1}$$

where

$$x = (x_1, x_2), \quad \mu = (\mu_1, \mu_2), \quad \Sigma = \begin{pmatrix} \Sigma_{11} \; \Sigma_{12} \\ \Sigma_{21} \; \Sigma_{22} \end{pmatrix},$$
$$\mu_{x_1|x_2} = \mu_{x_2} + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - x_1), \quad \Sigma_{x_1|x_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

In our bivariate case, we computed one score for the marginal $f(x_2;\mu_2,\Sigma_{22})$ and another for the conditional $f(x_1;\mu_{x_1|x_2},\Sigma_{x_1|x_2})$ PDF for each observation $(x_{1,t}, x_{2,t})$. When the null is true, these scores are *i.i.d.* $U[0,1]$ (Diebold et al., 1999). Two mutually independent scores can be also obtained from another factorization, in which x_1 and x_2 are swapped but they are not independent from the scores obtained in the first factorization. Therefore, we use one pair of independent scores per observation in the evaluation of the D-test and the CS-tests. For the Q-test, only one independent score series can be generated. For the reasons discussed at the end of Section 3, we compute the scores under the transformation that projects the largest component from the principal component analysis of the covariance matrix Σ on the main diagonal. Finally, the MN-test produces, by construction, two independent score series.

Table 1 reports the results of the experiment for two data generating processes (mixtures M_1 and M_2) and different values of the parameter δ.

[Table 1]

The performance of all tests, with the exception of the CS2-test and – to a lesser extent – the D-test, is comparable for the first mixture despite the fact that the Q-test uses only half of the scores relative to the other tests. For the second mixture, however, the Q-test and CS-tests clearly outperform their competitors.[8] The comparative disadvantage of the latter is due to the fact that the covariance matrices, estimated from the samples, are close to the identity matrix. In this case, the null hypothesis takes the form of the standard binormal. The D-test and the MN-test verify then, whether the marginal distributions follow the univariate standard normal and ignore the correlation between the variables. The Q-test and the CS-tests, on the contrary, combine the information from both variables, which allows for a sharper detection of a deviation from binormality. Furthermore, we found in this experiment that the performance of the Q-test does not deteriorate essentially in the canonical coordinate system.

Regarding the effect of the dimension N on the power of the tests, we investigated in another simulation the extent to which the tests suffer from the curse of dimensionality. For this purpose, we generalized the mixture M_1 from the previous example to ½N($(-\delta/\sqrt{N},...,-\delta/\sqrt{N}), I_N)$ + ½N($(\delta/\sqrt{N},...,\delta/\sqrt{N}), I_N)$. In this mixture, δ is the Euclidean distance between the origin of the coordinates and the means $(\pm\delta/\sqrt{N},...,\pm\delta/\sqrt{N})$ of the DGP. This distance remains constant for all dimensions N which makes the test results comparable across dimensions,

$$d((\delta/\sqrt{N},\ldots,\delta/\sqrt{N}),(0,..,0)) = d((-\delta/\sqrt{N},\ldots,-\delta/\sqrt{N}),(0,..,0)) = \sqrt{(\delta/\sqrt{N})^2 N} = \delta$$

[8] These results confirm the findings in Clements and Smith (2002) for the CS-tests and the D-test.

As in the previous experiment, the scores were computed under the null of multinormality with mean and covariance matrix obtained from the mixture, $\mu = (0,...,0)$, $\Sigma_{ii} = (N + \delta^2)/N$, $\Sigma_{ij} = \delta^2 / N$. For reasons of computational efficiency, the scores in the Q-test were obtained in the coordinate system rotated by the matrix of the eigenvectors of the estimated covariance matrix. As the hypothesised function becomes then a product of *N* marginal PDFs, the computation simplifies to the multiplication of *N*-PITs of these marginals. This operation can be performed efficiently in higher dimensions. For the evaluation of the MN-test, we stacked the *N*-dimensional scores into a single vector. Additionally, in a unidimensional version of the MN-test (MN1-test hereafter), we examined the vector of MN scores that corresponded to the rotated variable with the largest variance (the first principal component). The *N* vectors of scores in the D-test were obtained from the repeated application of the factorization (1) to the *N*-dimensional forecast. One score per observation $(x_{1,t},...,x_{N,t})$ was then computed for each of the independent factors. Table 2 reports the p-values of the Pearson's X^2-statistic for the tests Q/MN1/MN/D as computed from a sample of 2500 observations drawn from the above mixture for each value of δ and *N*.

[Table 2]

The MN1-test is by far the most powerful among the three contenders and seems to retain power in higher dimensions, at least for the parameter space under study. Interestingly, the tests MN and D are the worst performing ones in spite of exploiting *N*-1 additional independent score series relative to the tests MN1 and Q. Further analysis of the MN scores showed that the information on the true DGP is concentrated in the scores corresponding to the first principal component. The inclusion of other scores dilutes this information and leads to loss of power. For the D-test, none of the *N* individual score vectors is consistently superior to any other or to the stacked vectors. Finally, the Q-test performs worse than MN1-test but is clearly more powerful than the tests MN and D, although its power appears to decrease with higher *N*.

Finally, in an empirical study, we tested the hypothesis of multinormal distribution for the daily returns of S&P500, Dow Jones and Nasdaq equity indices. Table 3 presents summary statistics for the continuously compounded daily return series of equity indices computed from the raw prices. The mean returns are almost identical for all series and close to zero. In line with previous evidence, the distribution of daily returns is heavily leptokurtic and the hypothesis of univariate normality is strongly rejected for each equity index.

[Table 3]

In light of the individual results for the three indices, it comes as no surprise that the null of multinormality, where the parameters are estimated from the sample, is strongly rejected by all three tests with the p-values of the Pearson's χ^2-test virtually equal to zero.[9] More interesting are the insights offered by the scores computed by the Q-test. As explained in Section 3, these scores allow for verifying the accuracy of the forecasted

[9] For brevity, the detailed results are not presented. They are available from the authors upon request.

density in specific areas. For the Q-test in the canonical system, the scores contain information on the forecast accuracy in the "negative orthants" of the distribution. Table 4 contains the proportion of scores that fell into the orthant $Q((v_t,...,v_t))$, where v_t is defined by $\widehat{F}_{t-1}((v_t,...,v_t)) = \alpha$ for the nominal significance levels $\alpha = 0.005$, 0.01, 0.015, 0.02 and 0.025. By the results presented in Section 3, this proportion is equal to the exceedence rate of the MVaR at the corresponding coverage level 1-α. These proportions (exceedence rates) are consistently higher than the nominal levels α which means that the number of observations far in the negative tails is higher than that implied by a multinormal distribution. The stylized fact of fat tails in financial time series seems to be valid also in the multidimensional context.

[Table 4]

5. Summary and Conclusion

The focus of the forecasting literature has recently shifted to interval and density forecasts. This shift has been motivated by applications in finance and economics as well as the realization that density and interval forecasts convey more information that point forecasts. Density forecasts naturally raise the question of evaluation. While efficient evaluation techniques for the univariate case have developed rapidly, the literature on multivariate density forecast evaluation remains limited. Indeed, the Diebold et al. (1999) PIT test remains the main reference with extensions proposed by Clements and Smith (2000, 2002). A drawback of these approaches is that they rely on the PDF factorization into conditionals and marginals which may prove challenging even for simple functions.

In this paper, we provide flexible and intuitive alternative tests of multivariate forecast accuracy that rely on the univariate PIT idea and avoid the cumbersome decomposition into conditionals and marginals. The framework is particularly important for examining multiple sources of tail risk encapsulated in MVaR. We performed Monte Carlo simulations and an empirical case study that exemplified the applications of both procedures. Finally, regarding the sources of forecast errors, we expect the parameter estimation uncertainty to be of second-order importance when compared to dynamic misspecification (Chatfield, 1993). However, shedding light on the power of the proposed test in the presence forecast inaccuracy requires formal investigation which may suggest a possible avenue for future research.

6. Appendix

Proof of Proposition 1:

For a series of T observations $x = \{x_t\}_{t=1}^T$, $x_t = (x_{1,t},...,x_{N,t})$ of random variables $\{X_t\}_{t=1}^T$ with continuous distributions $\{F_t\}_{t=1}^T$, we define the series of T transformed values $\{z_t^M = F_t(x_t^M)\}_{t=1}^T$, where $x_t^M = Max\{x_{1,t},....,x_{N,t}\}\cdot(1,...,1)$, and the corresponding random variables $Z_t^M = F_t(X_t^M\} = F_t(Max\{X_{1,t},....,X_{N,t}\}\cdot(1,...,1))$.

We observe that if x_t belongs to the orthant $Q((v,...,v)) = \{y \in R^N : y \leq (v,...,v)\}$, $v \in R$, then x_t^M also belongs to $Q((v,...,v))$. This follows from the fact that $x_{i,t} \leq v$ for i=1,…,N implies $Max\{x_{1,t},....,x_{N,t}\} \leq v$. On the other hand, if x_t does not belong to $Q((v,...,v))$ then there must exist $x_{i,t} > v$ and, hence, $x_t^M \notin Q((v,...,v))$. Therefore,

$$\forall x_t \in Q((v,...v)),\ F_t(x_t) \leq F_t(x_t^M) \leq F_t((v,...v)), \qquad \text{(A1)}$$

$$\forall x_t \notin Q((v,...v)), F_t(x_t^M) > F_t((v,...v)).$$

In order to prove that Z_t^M is uniformly distributed over U[0,1], we have to show that Pr($Z_t^M < \alpha$) = α. From (A1) follows that $z_t^M = F_t(x_t^M) \leq \alpha =: F_t((v,...,v))$ whenever $x_t \in Q((v,...,v))$. The probability of the latter event is equal to the density mass over $Q((v,...,v))$, i.e., equal to $F_t((v,...,v)) = \alpha$.

Finally, since $Z_t^M \sim U[0,1]$ for any CDF $\widehat{F}_t$, the distribution of Z_t^M is independent of the distribution of Z_s^M for any s≠t.

References

Anderson NH, P. Hall, and D.M. Titterington, (1994) “Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-based Density Functions”. *Journal of Multivariate Analysis* 50, 41–54.

Bai, J. (2003) “Testing Parametric Conditional Distributions of Dynamic Models”. *Review of Economics and Statistics,* 85, 531-549.

Bai, J. and Z. Chen, (2008) “Testing Multivariate Distributions in GARCH Models’’. *Journal of Econometrics,* 143, 19-36.

Basel Committee on Banking Supervision. (1996, January). Overview of the amendment to the capital accord to incorporate market risks.

Berkowitz, J. (2001) “Testing Density Forecasts with Applications to Risk Management”. *Journal of Business and Economic Statistics,* 19, 465-474.

Birch, M. W. (1964) “A New Proof of the Pearson-Fisher Theorem”. *Annals of Mathematical Statistics*, 35(2), 817-824.

Brock, W. A., D. A. Hsieh, and B. LeBaron, (1991). Nonlinear Dynamics, Chaos and Instability. MIT Press, Cambridge, MA.

Chatfield, C. (1993) "Calculating Interval Forecasts," *Journal of Business and Economic Statistics*, 11, 121-135.

Chen, B. and Y. Hong (2009) “Characteristic Function-based Testing for Multifactor Continuous-Time Markov Models via Nonparametric Regression’’. *Econometric Theory,* doi:10.1017/S026646660999048X, 1-65.

Cox, D.R., and N.J.H. Small, (1978) “Testing multivariate normality”, *Biometrika* 65, 263–272

Christoffersen, P. F., (1998) “Evaluating Interval Forecasts”, *International Economic Review*, 39, 841–862.

Clements, M. P., and J. Smith, (2000) “Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment”, *Journal of Forecasting* 19, 255–276.

Clements, M.P., and J. Smith, (2002) “Evaluating Multivariate Forecast Densities: A Comparison of Two Approaches”. *International Journal of Forecasting*, 18, 397-407.

Corradi, V., and N.R. Swanson, (2006a) “Predictive Density Evaluation”. In: Granger, C.W.J., Elliot, G., Timmerman, A. (Eds.), *Handbook of Economic Forecasting*. Elsevier, Amsterdam, 197–284.

Corradi, V., and N.R. Swanson, (2006b), “Bootstrap Conditional Distribution Tests In the Presence of Dynamic Misspecification”. *Journal of Econometrics*, 133, 779-806

Diebold, F.X., and J. Lopez, (1996) "Forecast Evaluation and Combination," in G.S. Maddala and C.R. Rao (eds.), *Statistical Methods in Finance* (Handbook of Statistics, Volume 14). Amsterdam: North-Holland, 241-268.

Diebold, F.X., and R.S. Mariano, (1995) “Comparing predictive accuracy”. *Journal of Business and Economic Statistics* 13, 253–263.

Diebold, F.X., T. Gunther and A.S. Tay, (1998) “Evaluating Density Forecasts with Applications to Finance and Management”. *Intern. Econ. Review*, 39, 863-883.

Diebold, F.X., J. Hahn and A.S. Tay, (1999) “Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High Frequency Returns on Foreign Exchange”. *Review of Economics and Statistics*, 81, 661-673.

Gourieroux, C., and J. Jasiak, (2010) “Value at Risk”, in Handbook of financial econometrics: Tools and techniques, Yacine Aït-Sahalia, Lars Peter Hansen (Eds.), Ch.10, 554 – 612.

Granger, C.W.J., and M.H. Pesaran, (1999) “A Decision Theoretic Approach to Forecast Evaluation”. In *Statistics and Finance: An Interface*, Chan WS, Lin WK, Tong, H (eds). Imperial College Press: London.

Guidolin, M., and A. Timmermann, (2005) “Term Structure of Risk Under Alternative Econometric Specifications”. *Journal of Econometrics*. 131, 285-308.

Hong, Y. (2001) “Evaluation of out-of-sample probability density forecasts with applications to S&P 500 stock prices”. *Working Paper*, Cornell University.

Hong, Y., and H. Li, (2003) “Nonparametric specification testing for continuous time models with applications to term structure of interest rates”. *Review of Financial Studies*, 18, 37–84.

Hong, Y., H. Li, and F. Zhao, (2007) "Can the random walk model be beaten in out-of-sample density forecasts? Evidence from intraday foreign exchange rates". *Journal of Econometrics*, 141, 736-776.

Jorion, P. (2000) Value at risk. New York: McGraw Hill.

Khmaladze, E.V. (1981) "Martingale approach in the theory of goodness-of-fit tests", in Theory of Probability and its Applications, 26, 240–257.

Kupiec, P. H., (1995) "Techniques for verifying the accuracy of risk measurement models", *Journal of Derivatives*, 3, 73–84.

Li Q. (1996) "Nonparametric Testing of Closeness between Two Unknown Distribution Functions". *Econometric Reviews*, 15, 261–274.

Li, F., and G. Tkacz, (2001) "A Consistent Bootstrap Test for Conditional Density Functions with Time-Dependent Data". *Bank of Canada*, Working Paper No. 2001–21.

Mecklin, C.J., and D.J., Mundfrom, (2004) "An appraisal and bibliography of tests for multivariate normality". *International Statistical Review*, 72, 123–138.

Noceti, P., J. Smith and S. Hodges, (2003) "An Evaluation of Tests of Distributional Forecasts", *Journal of Forecast.* 22, 447–455.

Rosenblatt, M. (1952) "Remarks on a multivariate transformation". *Annals of Mathematical Statistics*, 23, 470–472.

Smith, S.P. and A.K. Jain, (1988) "A test to determine the multivariate normality of a data set", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 10(5), 757– 761.

Tay, A.S., and K.F. Wallis, (2000) "Density Forecasting: A Survey", *Journal of Forecasting*, 19, 165-175.

Watson, G. S. (1958) "On Chi-Square Goodness-Of-Fit Tests for Continuous Distributions", *Journal of the Royal Statistical Society. Series B*, 20(1), 44-72

Table 1 The performance of Q/MN/D/CS1/CS2 in a Monte Carlo Simulation

δ	Mixture M_1 $\frac{1}{2}N((-\delta,-\delta),I_2)$ + $\frac{1}{2}N((\delta, \delta), I_2)$	Mixture M_2 $\frac{1}{2}N((0,0),((1, \delta/2),(\delta/2,1)))$ +$\frac{1}{2}$ $N((0,0),((1,-\delta/2),(-\delta/2,1)))$
0.80	.072/.006/.251/.092/.545	.632/.702/.481/.546/.723
1.00	.003/.002/.197/.000/.728	.181/.093/.199/.132/.943
1.20	.000/.000/.128/.000/.535	.017/.349/.284/.004/.204
1.40	.000/.000/.000/.000/.130	.000/.432/.391/.000/.009
1.60	.000/.000/.000/.000/.094	.000/.000/.432/.000/.000
1.80	.000/.000/.000/.000/.007	.000/.000/.153/.000/.000

Notes: The table reports the p-values of the Pearson's χ^2-test for the tests Q/MN/D/CS1/CS2, respectively, under the null $N(\mu_i, \Sigma_i)$, where μ_i and Σ_i were obtained from the corresponding mixture M_i. The test statistic X^2 was computed from 5000 stacked scores for MN/D/C1/C2 and from 2500 scores for the Q-test. The p-values were obtained from the χ_k^2-distribution with *k*=499 and *k*=249 degrees of freedom, respectively. The degrees of freedom parameter *k* was chosen in such a way that the expected number of observations in each of the *k*+1 subintervals of [0,1] was 10.

Table 2 The performance of Q/MN1/MN/D in a Monte Carlo Simulation

δ	N								
	2	3	4	5	6	7	8	9	10
0.8	.112	.301	.478	.729	.208	.345	.877	.921	.710
	.082	.131	.453	.260	.323	.111	.219	.342	.292
	.176	.354	.429	.455	.712	.399	.630	.415	.379
	.482	.537	.790	.223	.432	.519	.771	.816	.644
1.0	.025	.122	.071	.277	.423	.501	.514	.697	.329
	.004	.007	.233	.020	.197	.007	.021	.200	.010
	.073	.245	.779	.092	.774	.931	.707	.435	.217
	.604	.473	.329	.749	.231	.793	.893	.583	.438
1.2	.008	.015	.065	.245	.558	.321	.195	.078	.291
	.000	.000	.000	.002	.010	.000	.005	.000	.007
	.036	.065	.269	.129	.671	.727	.342	.812	.775
	.543	.139	.891	.393	.551	.173	.515	.741	.116
1.4	.000	.015	.295	.074	.039	.347	.412	.358	.060
	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.000	.001	.413	.708	.299	.387	.047	.551	.214
	.373	.569	.298	.905	.542	.259	.233	.972	.491
1.6	.000	.000	.000	.000	.002	.020	.139	.002	.098
	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.000	.000	.002	.312	.249	.551	.003	.191	.606
	.148	.631	.721	.337	.612	.638	.914	.285	.733
1.8	.000	.000	.000	.000	.000	.000	.000	.091	.037
	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.000	.000	.017	.064	.143	.194	.248	.124	.322
	.004	.120	.348	.573	.940	.341	.089	.777	.483

Notes: The table reports the p-values of the Pearson's χ^2-test for the tests Q/MN/D/CS1/CS2, respectively, under the null of multinormality, with mean and covariance matrix obtained from the mixture ½N($(-\delta/\sqrt{N},...,-\delta/\sqrt{N}),I_N$) + ½N($(\delta/\sqrt{N},...,\delta/\sqrt{N}),I_N$). The X^2-statistic was computed from 2500 scores for the tests Q and MN1 and from 2500**N* scores for the tests MN and D. The p-values were computed from the χ^2_k-distribution with k=249 and k=250*N-1 degrees of freedom, respectively. The degrees of freedom parameter *k* was chosen in such a way that the expected number of observations in each of the *k*+1 subintervals of [0,1] was 10.

Table 3 Summary Statistics

Statistics	S&P500	Dow Jones	Nasdaq
Mean (%)	0.0083	0.0147	0.0128
Stand Dev (%)	1.1389	1.0919	1.8163
Skewness	0.051	-0.064	0.116
Kurtosis	4.984	6.004	6.614
X^2-stat (df=249)	433.5(0)	378.1(0)	514.8(0)

Notes: The table reports the mean, standard deviation, skewness, kurtosis and the Pearson's X^2-statistic (p-values in parenthesis) under the null of normality for the log returns for S&P500, Dow Jones and Nasdaq for the sample period 25/09/1998 to 29/08/08 (2498 daily observations).

Table 4 MVaR Unconditional Forecast Accuracy for the Multinormal Density

Nominal Significance	%x	t_u
$\alpha = 0.5\%$	0.881	2.037
$\alpha = 1\%$	1.361	1.558
$\alpha = 1.5\%$	1.962	1.664
$\alpha = 2\%$	2.562	1.778
$\alpha = 2.5\%$	3.163	1.892

Notes: The table reports the percentage of exceptions out of 2498 daily observations (i.e., the proportion of times the forecasted MVaR is exceeded) and the Kupiec's *t*-statistic to test the null hypothesis of unconditional accuracy for different nominal significance levels.

Figure 1: The contour area $\{y \in R^2 : \widehat{F}_{t-1}(y) < 0.025\}$ (gray) and the quadrant $Q((-1,-1)) = \{y \in R^2 : y \leq (-1,-1)\}$ (dark gray) for the standard binormal $\widehat{F}_{t-1}$. For observations (black dots) lying inside (outside) of the quadrant $Q((-1,-1))$, the "highest" of the projections on the main diagonal along the axes lies also inside (outside) of $Q((-1,-1))$.

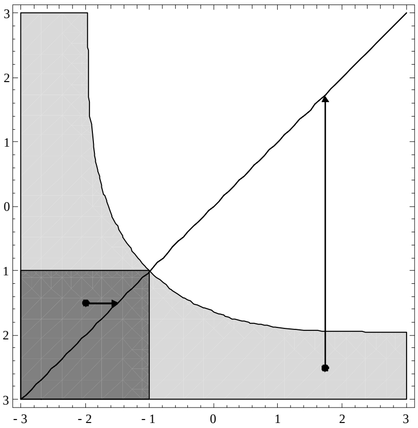

Figure 2: After the rotation of the canonical system clockwise by 90°, the south-east orthant Qse moves to the south-west position Qsw. The dashed lines are the main diagonals in the original and the rotated system while the shaded ellipse is the contour area of $\widehat{F}_{t-1}$.

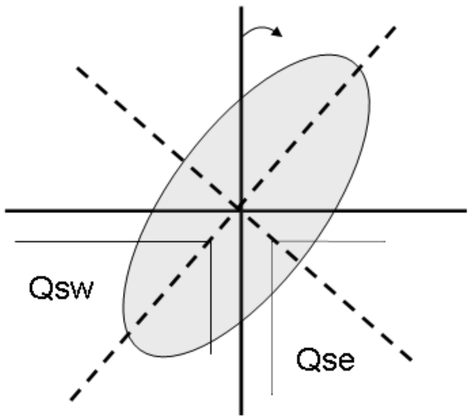

Figure 3: A scatter plot of scores generated from 1000 standard binormal observations under the null N((0,0),I). The x-axis (y-axis) corresponds to the scores computed in the canonical (90°-rotated) system.

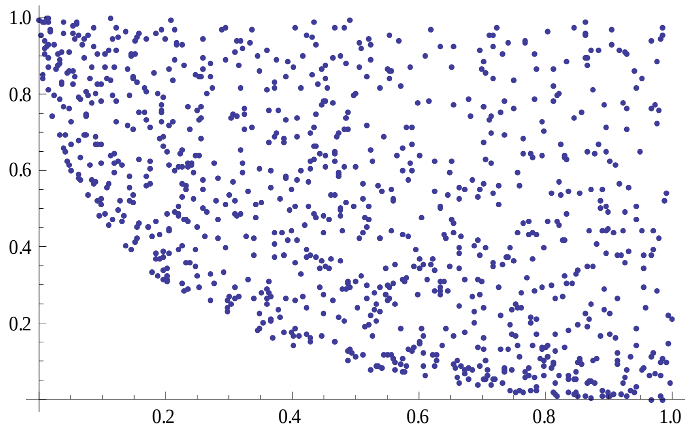

Figure 4: A sample of 1000 observations from the mixture 1: ½N((-δ,-δ),I_2) + ½N((δ,δ),I_2)) for δ=1.4.

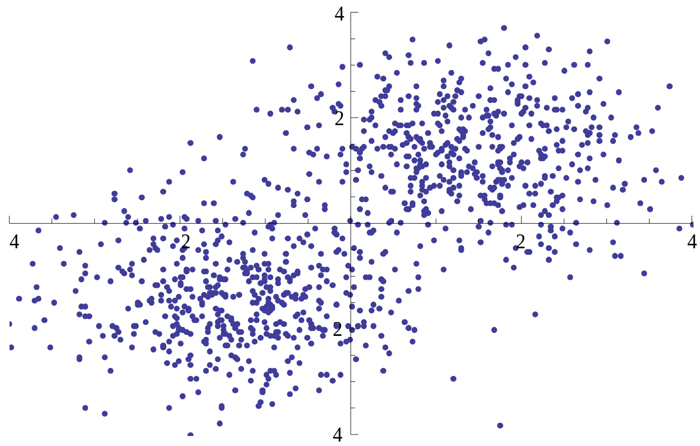

Figure 5: A sample of 1000 observations from the mixture 2: ½N((0,0),((1,-δ/2),(-δ/2,1))) + ½N((0,0), ((1, δ/2),(δ/2,1))) for δ=1.