UvA Econometrics

Discussion Paper: 2006/05

Power of the Neyman smooth test for evaluating multivariate forecast densities

Jan G. de Gooijer

www.fee.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA Universiteit van Amsterdam

Power of the Neyman Smooth Test for Evaluating Multivariate Forecast Densities *

Jan G. De Gooijer

Department of Quantitative Economics
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
Telephone: +31–20–525 4244; Fax: +31–20–525 4349
e-mail: j.g.degooijer@uva.nl

SUMMARY *We compare and investigate Neyman's smooth test, its components, and the Kolmogorov-Smirnov (KS) goodness-of-fit test for testing uniformity of multivariate forecast densities. Simulations indicate that the KS test lacks power when the forecast distributions are misspecified, especially for correlated sequences of random variables. Neyman's smooth test and its components work well in samples of size typically available, although there sometimes are size distortions. The components provide directed diagnosis regarding the kind of departure from the null. For illustration, the tests are applied to forecasts densities obtained from a bivariate threshold model fitted to high-frequency financial data.*

Key Words: Goodness-of-fit; Multivariate density forecasts; Uniform distribution.

*Forthcoming in: *Journal of Applied Statistics*

1 Introduction

Standard forecast evaluation criteria, like the root mean squared forecast error, condense the relative forecast performance at all horizons down to a single number. By contrast, a more complete characterization of the degree of uncertainty can be obtained by reporting interval forecasts and density forecasts; see, e.g., Clements (2005, Ch. 5) for a good introduction. This applies to both time series and cross section data. The focus in this paper is on evaluating multivariate forecast densities, and in particular on testing the hypothesis that the conditional multivariate forecast density depicts the true conditional forecast density. The following framework addresses this testing problem for bivariate time series data. For cross section data an analogous framework holds.

Suppose, for ease of exposition, that we have a series of L 1–step ahead forecasts of a bivariate time series $Z_t = (Z_{1,t}, Z_{2,t})'$. Let $p_t(Z_{1,t}, Z_{2,t}|\Omega_{t-1})$ $(t = 1, \ldots, L)$ denote the joint forecast density, where the set Ω_{t-1} refers to the past history of $(Z_{1,t}, Z_{2,t})$. Further, suppose this density can be factorized into the product of the conditional (c) density and the marginal (m) density as, e.g., $p_t(Z_{1,t}, Z_{2,t}|\Omega_{t-1}) = p_t^c(Z_{1,t}|Z_{2,t}, \Omega_{t-1}) \times p_t^m(Z_{2,t}|\Omega_{t-1})$. Each element $(Z_{1,t}, Z_{2,t})$ can be transformed by its corresponding probability integral transformation (PIT) to give

$$U_{1|2,t}^c = \int_{-\infty}^{Z_{1|2,t+1}^c} p_t^c(u|Z_{2,t}, \Omega_{t-1})du, \quad U_{2,t}^m = \int_{-\infty}^{Z_{2,t+1}^m} p_t^m(u|\Omega_{t-1})du, \ (t = 1, \ldots, L), \tag{1}$$

where $Z_{1|2,t+1}^c$ and $Z_{2,t+1}^m$ are respectively the conditional and marginal 1–step ahead forecasts. Under the null (H_0) that the model forecast density corresponds to the true forecast density, given by the data generating process (DGP) which is denoted by $f_t(Z_{1,t}, Z_{2,t}|\Omega_{t-1})$, that is $p_t(Z_{1,t}, Z_{2,t}|\Omega_{t-1}) = f_t(Z_{1,t}, Z_{2,t}|\Omega_{t-1})$, the two sequences of random variables $\{U_{1|2,t}^c\}$ and $\{U_{2,t}^m\}$ $(t = 1, \ldots, L)$ will each be *i.i.d.* $U(0, 1)$ distributed (Rosenblatt, 1952). Moreover, the two sequences of PITs will themselves be independent.

Various approaches can be used to assess whether a particular sequence of PITs, say $\{U_t\}_{t=1}^L$, is *i.i.d.* $U(0, 1)$. Diebold *et al.* (1998) checked uniformity and serial independence of the U_t's graphically via the histogram and the sample correlogram. They also advocated the use of the Kolmogorov-Smirnov (KS) goodness-of-fit test as a more formal way of testing the uniformity part. However, it is well-known that the power of the KS statistic is rather low; see, e.g., Stephens (1974). Alternative tests for *i.i.d.* uniformity, which are often equal in power to the KS statistic, are for instance the Crámer-von Mises test and the Kuiper test.

The above methods are sometimes referred to as omnibus tests, i.e. they are sensitive to almost all alternatives to the null. In the present context, this property implies that when an

omnibus test fails to reject the H_0, we can conclude that there is not enough evidence that the time series is not generated from the joint forecasting density. On the other hand, a rejection would not provide any information about the form of the density. Test statistics that can be decomposed into interpretable components may be a solution. Such a test is Neyman's (1937) smooth test for testing uniformity. The test can be viewed as a compromise between omnibus tests and tests whose power is focused in the direction of a specific alternative. Successive components of the smooth test can be directly related to changes in mean, variance, skewness, and kurtosis (Section 2). Surveys of the works on Neyman's smooth test are provided by Rayner & Best (1989) and Bera & Ghosh (2001). Interestingly, in reviewing several goodness-of-fit tests, Rayner & Best (1990) concluded, "Don't use those other methods – use a smooth test!".

In this paper we adopt Neyman's smooth test to assess departures from the H_0 with respect to specific features of the multivariate forecast distribution function, like their location, scale and skewness. We compare and investigate the size and power of Neyman smooth test, its components, and the KS test under a number of distributional assumptions for the DGP (Subsections 3.1–3.3). We also consider the properties of the tests in the presence of misspecified time series correlation structure of a bivariate VAR model, using PITs of 1– and 2–step ahead forecast error densities from estimated univariate AR models; Subsection 3.4. In Section 4 we apply the tests to forecasts densities obtained from a bivariate threshold model fitted to minute returns of S&P 500 index futures and prices. Section 5 concludes.

Because we are interested in misspecification in one of the first four moments, rather than the *i.i.d.* property, we do not consider the Berkowitz (2001) test. Also, in the simulations and the empirical example, we ignore the impact of various forms of model misspecification on the outcomes of the test statistics.

2 Neyman's smooth test

Suppose $X_1, \ldots, X_n$ are n independent observations on a random variable X with unknown distribution function $F(x)$. Let $f(x)$ denote the associated probability density function (pdf). To test whether the pdf of X is a uniform density function in the interval $[0, 1]$, Neyman (1937) postulated the (smooth) alternative hypothesis that the pdf of the X_i is given by

$$C(\theta) \exp\Big\{ \sum_{i=1}^{k} \theta_i h_i(x) \Big\}, \quad (k = 1, \ldots, n-1),$$

where $\theta = (\theta_1, \ldots, \theta_k)$ is a vector of k real parameters, the functions $h_i(\cdot)$ are the first normalized Legendre polynomials on $[0,1]$ of order i, θ_i are free parameters, and $C(\cdot)$ is a normalization function that ensures the pdf integrates to one. Using the multiparameter version of the generalized Neyman-Pearson lemma, the smooth test statistic is given by

$$\Psi_k^2 \quad = \quad \sum_{i=1}^{k} u_i^2 \quad \text{with} \quad u_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_i(y_j), \tag{2}$$

where $y_j = F(x_j) = \int_0^{x_j} f(u)du$ $(j = 1, \ldots, n)$. Under $H_0^* : \theta = 0$, Ψ_k^2 is asymptotically χ_k^2 distributed, and under H_1^* : at least one $\theta_i \neq 0$, the test statistic follows a non-central χ_k^2 distribution with non-centrality parameter $\lambda = \sum_{i=1}^{k} \theta_i^2$.

Many generalizations of Neyman's smooth test for uniformity have been proposed in the literature, including testing for different alternatives using different orthonormal polynomials; see, e.g., Rayner & Best (1989, 1990). One advantage of introducing the orthonormal Legendre functions is that the asymptotic null distribution of Ψ_k^2 turns out to be simple. Another advantage is that the components u_i^2 are easily interpretable. To see this in more detail, consider the following recursive relation for Legendre polynomials $P_i(z)$ $(i \geq 2)$ on $[-1,1]$: $(i+1)P_{i+1}(z) = (2i+1)zP_i(z) - iP_{i-1}(z)$, starting from $P_0(z) = 1$, $P_1(z) = z$. To render them orthogonal we introduce the normalized functions $h_i(z) = P_i(z)/s_i$ with $s_i = (2i+1)^{-1/2}$. Now, if we want to test whether $f(x)$ is $U(0,1)$ distributed, we define $z = 2x - 1$. This implies that we test if z is a uniform variable in $[-1,1]$. Then, it is easy to verify, that the components u_i^2 $(i = 1, \ldots, 4)$ are given by

$$u_1^2 = 3n(\hat{\mu}_1)^2,\ u_2^2 = 45n(\hat{\mu}_2 - 1/3)^2/4,\ u_3^2 = 7n(5\hat{\mu}_3 - 3\hat{\mu}_1)^2/4,$$

$$u_4^2 = 9n\{35(\hat{\mu}_4 - 1/5) - 30(\hat{\mu}_2 - 1/3)\}^2/64,$$

where $\hat{\mu}_i = (\sum_{j=1}^{n} z_j^i)/n$.

3 Monte Carlo simulations

Here we present some Monte Carlo power results for Neyman's smooth test Ψ_4^2, its components u_i^2 $(i = 1, \ldots, 4)$, and the KS test, for a number of bivariate distributions. The choice $k = 4$ was motivated by Monte Carlo results obtained by Rayner & Rayner (2001). They noted that fewer components result in more powerful smooth tests for uniformity. As to the choice of PITs, Clements & Smith (2002) showed that the KS test of uniformity has the highest power for

both the 'product' (p) and the 'ratio' (r) of PITs, having typical elements $\{U_t^p = U_{1|2,t}^c \times U_{2,t}^m\}$ and $\{U_t^r = U_{1|2,t}^c / U_{2,t}^m\}$ respectively. So we decided to restrict the simulations to these two combinations of PITs. The associated distribution functions can be easily obtained; see, e.g. Clements & Smith (2002, Appendix). Specifically, let U_1 and U_2 be two independent random variables each $U(0,1)$ distributed. Then the random variable $U^p = U_1 \times U_2$ has a distribution function given by $F_{U^p}(x) = x - x\ln(x)$, $0 < x < 1$. Further, it can be shown the distribution function of the ratio of U_1 and U_2, say $U^r = U_1/U_2$, is given by $F_{U^r}(x) = x/2$ if $0 < x < 1$, and $F_{U^r}(x) = 1 - 0.5 \times x^{-1}$ if $1 < x < \infty$.

In all experiments the nominal significance level was set at 0.05. The number of replications was fixed at 1000.

3.1 Standardized bivariate normal

When there are just two variables, X_1 and X_2, it is well-known that the standardized bivariate normal distribution is given by

$$p(x_1, x_2; \rho) = Z(x_1)Z\Big(\frac{x_2 - \rho x_1}{\sqrt{1-\rho^2}}\Big)\Big/\sqrt{1-\rho^2} = Z(x_2)Z\Big(\frac{x_1 - \rho x_2}{\sqrt{1-\rho^2}}\Big)\Big/\sqrt{1-\rho^2}, \tag{3}$$

where $Z(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$, and ρ the correlation coefficient. Thus, the conditional distribution of X_1, given X_2, is normal with expected value ρX_2 and variance $(1-\rho^2)$. In generating the PITs from (3) it is effectively assumed that the conditional distribution for X_1, given X_2, is equal to the marginal for X_1. This is an example of misspecification that affects only the correlation. To make the rejection rates comparable across statistics, the estimated rejection rates are size-adjusted, i.e. the size being common for all test statistics when $\rho \neq 0$.

Table 1 shows sizes and size-adjusted power results for $n = 50$, and 100 with correlation values ranging from -0.8 to 0.8 in steps of 0.4. The simulations permit several observations.

Table 1 about here

Table 2 about here

- The $u_2^2(\cdot)$ tests are the most powerful for both $n = 50$ and $n = 100$. The second highest power is obtained by the $\Psi_4^2(\cdot)$ tests.
- The power of the $u_2^2(p)$ tests is markedly better for $\rho < 0$ than for $\rho > 0$. The $u_2^2(r)$ tests show a similar asymmetry in power, but now the performance of both tests is better for $\rho > 0$ than for $\rho < 0$.

- The KS$(\cdot)$ tests have the lowest power as compared to the $u_2^2(\cdot)$, and $\Psi_4^2(\cdot)$ tests. In particular this applies to values of ρ in the range $[-0.4, 0.4]$.

3.2 Marginal Student's t with Gaussian copula

Copulae provide a general approach to modelling dependence between random variables. Similar as Clements & Smith (2002), we use the bivariate Gaussian copula function to generate alternative bivariate distributions. In particular, given the standardized bivariate vector of random variables (X_1, X_2) with correlation ρ, we construct correlated uniform random variables as $U_i = \Phi(X_i)$ $(i = 1, 2)$. Then, applying the inverse of the Student-t distribution to the uniform random variables, we obtain random drawings from bivariate distributions with correlated marginal Student-t distributions.

Table 2 shows sizes and size-adjusted power results for $n = 50$ and 100, with marginal Student-t distributions having 5 and 10 degrees of freedom (df). Some observations are in order:

- When $\rho < 0$, $u_2^2(p)$ has the highest power for all values of n and df. On the other hand, when $\rho > 0$, the $u_2^2(r)$ test is more powerful.
- When $\rho < 0$, the $\Psi_4^2(p)$ test displays power nearly as high as the $u_2^2(p)$ test. Also, when $\rho > 0$, the $\Psi_4^2(r)$ test performs slightly worse than the $u_2^2(r)$ test. This suggests that a reasonable power of Neyman's smooth test can already be obtained from using a few components like $u_2^2(\cdot)$, $u_4^2(\cdot)$, or $u_2^2(\cdot) + u_4^2(\cdot)$.
- All tests show increasing power as n and/or df increase.
- The size properties of all tests are similar. In almost all cases the empirical sizes are higher than the nominal size.

3.3 Marginal generalized λ distribution with Gaussian copula

The generalized lambda distribution (GLD) is a flexible four-parameter family of curves which includes several standard distributions as special cases; see, e.g., Karian & Dudewicz (2000). The GLD is defined by its quantile function

$$F^{-1}(p) \equiv Q_{\lambda_1,\lambda_2,\lambda_3,\lambda_4}(p) = \lambda_1 - \frac{1}{\lambda_2\lambda_3} + \frac{1}{\lambda_2\lambda_4} + \frac{1}{\lambda_2}\Big(\frac{p^{\lambda_3}}{\lambda_3} - \frac{(1-p)^{\lambda_4}}{\lambda_4}\Big), \quad (0 \leqslant p \leqslant 1),$$

where λ_1 denotes the location parameter, λ_2 is the scale parameter, and λ_3 and λ_4 are the shape parameters. If $\lambda_3 = \lambda_4$ the distribution is symmetric. The condition $\min(\lambda_3, \lambda_4) > -1/4$ ensures

that fourth and lower moments exist. In this case, the skewness and kurtosis are well-defined functions of λ_3 and λ_4; see, e.g., Ramberg *et al.* (1979). Inverting the first derivative of the quantile function yields the density function $f(Q(p))$:

$$f(Q(p)) = \lambda_2/(\lambda_3 p^{\lambda_3 - 1} + \lambda_4(1-p)^{\lambda_4 - 1}).$$

The condition $\lambda_3 \times \lambda_4 > 0$, along with the appropriate scaling function, ensures that this density function is non-negative on the probability measure p.

We consider three GLDs, each scaled to have a 0 mean and unit variance. The parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are set at $(-0.494, 0.2000, 0.0782, 0.2069)$, $(-0.2220, 0.0485, 0.0223, 0.0337)$, and $(-0.049, 0.0276, 0.0149, 0.0163)$. The corresponding skewness and kurtosis values are, respectively, i) (0.4,3), ii) (0.4,4), and iii) (0.1,4). Figures 1.a)–1.f) give plots of the size-adjusted estimated power functions of the test statistics KS$(\cdot)$, $u_2^2(\cdot)$, and $\Psi_4^2(\cdot)$. The $u_2^2(\cdot)$ test was selected because its estimated power was the highest among the $u_i^2(\cdot)$ $(i = 1, \ldots, 4)$ tests. The sample size was fixed at $n = 100$. We also looked at random samples of $n = 50$. The results presented here are typical.

- From Figures 1.a), 1.c), and 1.e) we see that the $u_2^2(p)$ test (dashed-dotted-dotted lines) has the highest power for all skewness and kurtosis values, for all values of ρ. The lowest power is obtained with the KS(p) test (solid lines).

- From Figures 1.b), 1.d), and 1.f) we see that $u_2^2(r)$ has the highest power, followed by the $\Psi_4^2(r)$ test (medium-dashed lines) for all values of ρ, skewness and kurtosis. The powers of both tests increase in the direction of a symmetric distribution when skewness and kurtosis move from (0.4,4) to (0.1,4). Also the power increases when the distribution is more skewed and less peaked; compare Figure 1.d) and Figure 1.f).

Figure 1 about here

3.4 VAR model

In this subsection we compare the relative merits of the tests in a dynamic out-of-sample forecasting experiment. To this end, we consider the stationary bivariate VAR(1) process:

$$Z_t = \begin{bmatrix} 5 \\ 10 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} Z_{t-1} + \varepsilon_t, \tag{4}$$

where $\{\varepsilon_t\}$ is a sequence of random shock vectors normally, *i.i.d.* with zero mean and with positive definite covariance matrix $\Omega = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. With this process, we generated bivariate time series of length $T = 200$, plus some pre-sample values to avoid possible start-up problems, for values of ρ ranging from -0.8 to 0.8 in steps of 0.4. However, we incorrectly assume that both series $Z_{1,t}$ and $Z_{2,t}$ can be well-fitted by univariate AR(2) models.

To generate the forecasts, the rolling forecasting method is used. That is: let T be the total number of observations. Let m be the last in-sample observation. Then, for this particular origin, $T-m$ observations are retained as the subsample for evaluating the forecasts of the time series models. By 'rolling' it is meant that time-index t extends as far as $T-S$ where S is the maximum forecast horizon under consideration. At each time point t the parameters estimates of the AR(2) models are re-estimated as new observations become available in the subsample. Thus, the method gives rise to $T-m$ 1–step ahead forecasts and associated forecast errors, $T-m-1$ 2–step ahead forecasts and associated forecast errors, etc.

Note that all tests discussed in this paper can be applied to $h > 1$–step ahead forecasts provided the following simple provision for the usual $(h-1)$–order dependence in the optimal successive forecasts is made. That is, we divide the forecasts into sets of independent forecasts, taking the first, the $h+1$, the $2h+1$ etc. for set 1, and the second, the $h+2$, the $2h+2$ etc. for the second set, and so on; see Diebold *et al.* (1999). Thus, if $h = 2$, each of the sub-series of PITs $\{U_1^p, U_3^p, U_5^p, \ldots\}$, $\{U_2^p, U_4^p, U_6^p, \ldots\}$ ... should be *i.i.d.* $U(0,1)$. In a similar way sub-series of PITs can be obtained from the sequences $\{U_t^r\}_{t=1}^L$. Tables 3 and 4 show power results for, respectively, a maximum forecast horizon $S = 1$ (so $h = 1$) and $S = 2$. From Table 3 we make the following observations:

Table 3 about here

Table 4 about here

- When $\rho = 0$, the highest power is obtained with the $u_2^2(\cdot)$ test. Note that the actual marginal models of a bivariate VAR(1) model is univariate ARMA(2,1). Thus, the fitted univariate AR(2) models are misspecified. It is interesting to see the test detects this neglected dynamic structure of the data through the PITs of the forecast error densities.

- When $\rho = -0.4$, $u_2^2(p)$ is more powerful than $u_2^2(r)$. On the other hand, when $\rho = 0.4$, $u_2^2(r)$ is to be preferred over $u_2^2(p)$. This was also observed in Subsection 3.1. Finally, as

expected, all powers increase when the concurrent correlation between the two series is increasing.

The results in Table 4 permit the following conclusions:

- When $\rho = 0$, the highest power resulted from the $u_2^2(r)$ test. There is no power difference for $h = 1$ and for $h = 2$. This also applies to the case $\rho \neq 0$. As compared to the case $\rho = 0$ and $h = S = 1$ (Table 3), all tests display lower power results. This suggests that for out-of-sample forecasting the tests are sensitive to the choice of the maximum forecast horizon S.

- When $\rho > 0$, the powers of all tests are markedly better for the product of PITs than for the ratio of PITs. However, when $\rho < 0$, the performance of the tests is better for the ratio of PITs than for the product of PITs.

The above observations are typical for forecasts made from other AR models, and different values of T and m.

4 An application

As an illustration, we evaluate density forecasts from a two-regime bivariate threshold model. The data under study are the transactions for the S&P500 stock index in May 1993 and its June futures contract traded at the Chicago Mercantile Exchange; see Forbes *et al.* (1999), who used to construct the time series $f_{t,\ell}$, the log price of the index futures at maturity ℓ, and the series s_t, the log of a security index price. The time interval is 1–minute (intraday). Several authors used this data to study index futures arbitrage. Let $Z_{1,t} = f_{t,\ell} - f_{t-1,\ell}$ and $Z_{2,t} = s_t - s_{t-1}$ denote the first differences of the series. Similar to Tsay (1998), we replaced 10 extreme values in the series $Z_{1,t}$ and $Z_{2,t}$ by the simple average of their two nearest neighbors. This step may affect the conditional heteroskedasticity in the data. However, it is not the intention of this paper to specify any type of parametric model to take care of (G)ARCH-type effects. Then, using $T = 7060$ observations, Tsay (1998) fitted a three-regime bivariate threshold model to $Z_{1,t}$ and $Z_{2,t}$, with a third (exogenous) variable X_{t-1}^* controlling the switching dynamics. The variable X_t^* is assumed to be weakly stationary and have a continuous distribution. Its values follow from a version of the so-called *cost-of-carry model*; see, e.g., Tsay (1998). Time plots of the three series are provided by Tsay (1998, 2002).

Let $Z_t = (Z_{1,t}, Z_{2,t})'$. Then the two-regime bivariate threshold model of order p is given by

$$Z_t = \begin{cases} c_1 + \sum_{i=1}^{p} \Phi_i^{(1)} Z_{t-i} + \beta_1 X_{t-1} + \varepsilon_t^{(1)} & \text{if } X_{t-1} \leq \gamma \\ c_2 + \sum_{i=1}^{p} \Phi_i^{(2)} Z_{t-i} + \beta_2 X_{t-1} + \varepsilon_t^{(2)} & \text{if } X_{t-1} > \gamma, \end{cases} \tag{5}$$

where $X_t = 100 \times X_t^*$, γ is a real number, $\Phi_i^{(1)}$ $(i = 1, \ldots, p)$ are 2×2 matrices of coefficients, c_j are constants, and β_j $(j = 1, 2)$ are unknown parameters. The innovations $\{\varepsilon_t^{(j)}\}$ satisfy $\varepsilon_t^{(j)} = \Sigma_j^{1/2} a_t$, where $\Sigma_j^{1/2}$ are symmetric positive definite matrices, and $\{a_t\}$ is a sequence of *i.i.d.* random variates with mean zero and covariance matrix I, the identity matrix. Model (5) is a special case of Tsay's (1998) three-regime bivariate model. But, since in his model the series $Z_{1,t}$ and $Z_{2,t}$ do not depend on the variable X_t in the middle regime, it is reasonable to evaluate the prediction performance of (5).

Table 5 about here

Using the notation introduced in Subsection 3.4, we fixed $S = 1$. The following three values were selected for the last in-sample observation $m = 6960$, 6860, and 6760. Hence, 1–step ahead forecast densities will be based on respectively, $L = 100$, 200, and 300 1–step ahead forecasts. We set the maximum value of the lag order p at 8. In each step of the 'rolling' forecasting method, the minimum AIC criterion was employed to select the threshold value γ, using a grid search method with 300 points. The models' innovations $\{\varepsilon_t^{(j)}\}$ are assumed to be Gaussian distributed. Note that within this setup, both the model's order and the model's parameter estimates are the same as the population values for generating the forecasts. Thus, we neglect possible sensitivity of the tests with respect to various forms of model misspecification. Also the imposition of normality of the error process may affect the outcomes of the tests.

Table 5 presents the results of the KS, the smooth test and its components. Values of $\epsilon(\alpha, n)$ for which $\alpha = \text{Prob.}(\text{KS} \geq \epsilon(\alpha, n))$ are tabulated by Miller (1956) for various levels α, and sample sizes $n = 1, 2, \ldots, 100$. At $n = 100$, $\alpha = 0.01$, and 0.05 the tabular values read 0.14987 and 0.12067, respectively. For $n > 100$, critical values can be obtained by the asymptotic formula $\tilde{\epsilon}(\alpha, n) = \sqrt{\ln(1/\alpha)/2n}$. At $n = 200$ $(n = 300)$, this formula gives 0.10730 (0.08761) $(\alpha = 0.01)$ and 0.08654 (0.07066) $(\alpha = 0.05)$.

We see that for $L = 100$, 200, and 300 almost all test statistics based on the product of PITs suggest that there is no evidence against the usefulness of the threshold model to predict the density of future realizations of minute returns of S&P 500 index futures and prices at the 5% level. One exception is the first component $u_1^2(p)$ for $L = 300$. For the ratio of PITs both the

second component $u_2^2(r)$ and the overall $\Psi_4^2(r)$ statistic are highly significant at the 5% level for $L = 200$ and 300. From this we can infer that some variability in the second moments are not accounted for by the fitted threshold models. The KS test does not provide any indication of model misspecification.

5 Concluding remarks

We have compared and evaluated the power and size properties of the classical KS goodness-of-fit test with Neyman's smooth test and its components for testing uniformity of multivariate forecast densities. Both simulations and empirical results indicate that the latter test and its components outperform the KS test for various alternative bivariate distributions, and forecast horizons. Moreover, Neyman's smooth test and its components are directional in detecting model misspecifications. Hence, the use of Neyman's smooth test is recommended as a formal test for evaluating multivariate forecast densities.

This paper makes one simplifying assumption: the forecast model is completely known. This assumption was adopted in order not to complicate the comparison of the tests. A pure non-parametric estimation of the model and its empirical forecast density may be used as a sensible alternative approach. We are currently working in this direction.

REFERENCES

Bera, A.K. & Ghosh, A. (2001) Neyman's smooth test and its applications in econometrics. *In:* A. Ullah, A. Wan & A. Chaturvedi (Eds) *Handbook of Applied Econometrics and Statistical Inference* (New York: Marcel Dekker), pp. 177–230.

Berkowitz, J. (2001) Testing density forecasts, with applications to risk management, *Journal of Business & Economic Statistics*, 19, pp. 465–474.

Clements, M.P. & Smith, J. (2002) Evaluating multivariate forecast densities: a comparison of two approaches, *International Journal of Forecasting*, 18, pp. 397–407.

Clements, M.P. (2005) *Evaluating Econometric Forecasts of Economic and Financial Variables*, (New York: Palgrave Macmillan).

Diebold, F.X., Gunther, T.A. & A.S. Tay (1998) Evaluating density forecasts with applications to financial risk management, *International Economic Review*, 39, pp. 863–883.

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

Economics and Statistics, 81, pp. 661–673.

Forbes, C.S., Kalb, G.R.J. & Kofman, P. (1999) Bayesian arbitrage threshold analysis, *Journal of Business & Economic Statistics*, 17, pp. 364–372.

Karian, Z. & Dudewicz (2000) *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods* (Boca Raton: CRC Press).

Miller, L.H. (1956) Table of percentage points of Kolmogorov statistics, *Journal of the American Statistical Association*, 51, pp. 111–121.

Neyman, J. (1937) Smooth' test for goodness of fit, *Skandinavisk Aktuarietidskrift*, 20, pp. 150–199.

Ramber, J.S, Dudewicz, E.J., Tadikamalla, P.R., & Mykytka, E.F. (1979) A probability distribution and its uses in fitting data, *Technometrics*, 21, pp. 201–214.

Rayner, J.C.W. & D.J. Best (1989) *Smooth Tests of Goodness of Fit* (Oxford: Oxford University Press).

Rayner, J.C.W. & D.J. Best (1990) Smooth tests of goodness of fit; an overview, *International Statistical Review*, 58, pp. 9–17.

Rayner, G.D. & J.C.W. Rayner (2001) Power of the Neyman smooth test for the uniform distribution, *Journal of Applied Mathematics and Decision Sciences*, 5, pp. 1–11.

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

Stephens, M.A. (1974) EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, 69, pp. 730–737.

Tsay, R.S. (1998) Testing and modeling multivariate threshold models, *Journal of the American Statistical Association*, 93, pp. 1188–1202.

Tsay, R.S. (2002) *Analysis of Financial Time Series* (New York: Wiley).

Table 1: Sizes and power of KS$(\cdot)$, $\Psi_4^2(\cdot)$, and $u_i^2(\cdot)$ $(i = 1, \ldots, 4)$ tests; normal distribution.

	Product (p)						Ratio (r)					
		Smooth tests						Smooth tests				
ρ	KS(p)	$u_1^2(p)$	$u_2^2(p)$	$u_3^2(p)$	$u_4^2(p)$	$\Psi_4^2(p)$	KS(r)	$u_1^2(r)$	$u_2^2(r)$	$u_3^2(r)$	$u_4^2(r)$	$\Psi_4^2(r)$
						$n = 50$						
-0.8	0.781	0.117	0.995	0.023	0.116	0.978	0.219	0.094	0.628	0.066	0.070	0.451
-0.4	0.105	0.061	0.369	0.033	0.033	0.149	0.089	0.054	0.263	0.071	0.068	0.186
0.0	0.056	0.048	0.045	0.056	0.052	0.045	0.040	0.036	0.049	0.048	0.056	0.052
0.4	0.118	0.103	0.295	0.081	0.135	0.287	0.083	0.032	0.393	0.034	0.041	0.164
0.8	0.321	0.237	0.748	0.163	0.444	0.754	0.737	0.004	1.000	0.011	0.377	0.989
						$n = 100$						
-0.8	0.997	0.299	1.000	0.028	0.279	1.000	0.418	0.106	0.918	0.070	0.095	0.785
-0.4	0.190	0.095	0.709	0.030	0.056	0.499	0.142	0.074	0.457	0.059	0.067	0.302
0.0	0.040	0.043	0.036	0.030	0.046	0.040	0.029	0.038	0.045	0.058	0.045	0.044
0.4	0.195	0.161	0.526	0.088	0.252	0.511	0.152	0.030	0.692	0.033	0.046	0.415
0.8	0.664	0.396	0.957	0.219	0.693	0.967	0.996	0.006	1.000	0.011	0.744	1.000

Table 2: Size and power of KS$(\cdot)$, $\Psi_4^2(\cdot)$, and $u_i^2(\cdot)$ $(i = 1, \ldots, 4)$ tests; Student'-t (Gaussian copula) distribution.

	Product (p)						Ratio (r)					
		Smooth test						Smooth test				
ρ	KS(p)	$u_1^2(p)$	$u_2^2(p)$	$u_3^2(p)$	$u_4^2(p)$	$\Psi_4^2(p)$	KS(r)	$u_1^2(r)$	$u_2^2(r)$	$u_3^2(r)$	$u_4^2(r)$	$\Psi_4^2(r)$
$n = 50$, $df = 10$												
-0.8	0.824	0.068	0.992	0.021	0.257	0.982	0.151	0.097	0.379	0.067	0.046	0.303
-0.4	0.125	0.036	0.476	0.035	0.042	0.247	0.079	0.074	0.102	0.065	0.029	0.107
0.0	0.046	0.047	0.060	0.068	0.059	0.049	0.065	0.047	0.079	0.057	0.066	0.074
0.4	0.156	0.137	0.212	0.052	0.132	0.260	0.108	0.025	0.402	0.037	0.055	0.230
0.8	0.323	0.273	0.590	0.103	0.389	0.676	0.791	0.003	1.000	0.011	0.470	0.988
$n = 100$, $df = 10$												
-0.8	0.997	0.149	1.000	0.027	0.476	1.000	0.376	0.091	0.730	0.079	0.073	0.593
-0.4	0.225	0.049	0.709	0.043	0.026	0.492	0.151	0.069	0.221	0.065	0.066	0.181
0.0	0.065	0.057	0.079	0.063	0.060	0.080	0.033	0.033	0.065	0.043	0.049	0.053
0.4	0.169	0.156	0.283	0.051	0.205	0.346	0.320	0.032	0.741	0.032	0.103	0.532
0.8	0.539	0.394	0.792	0.136	0.583	0.891	1.000	0.004	1.000	0.015	0.880	1.000
$n = 50$, $df = 5$												
-0.8	0.719	0.022	0.988	0.032	0.437	0.972	0.096	0.082	0.170	0.074	0.083	0.185
-0.4	0.078	0.017	0.481	0.052	0.044	0.285	0.070	0.075	0.022	0.059	0.070	0.067
0.0	0.097	0.053	0.120	0.093	0.061	0.111	0.064	0.033	0.111	0.062	0.052	0.075
0.4	0.084	0.129	0.053	0.042	0.098	0.101	0.184	0.028	0.501	0.028	0.120	0.401
0.8	0.187	0.273	0.242	0.070	0.295	0.418	0.921	0.004	1.000	0.012	0.756	0.998
$n = 100$, $df = 5$												
-0.8	0.986	0.062	1.000	0.042	0.824	1.000	0.136	0.077	0.212	0.070	0.053	0.224
-0.4	0.144	0.017	0.755	0.153	0.093	0.830	0.060	0.075	0.027	0.055	0.083	0.556
0.0	0.116	0.064	0.211	0.142	0.066	0.176	0.061	0.038	0.213	0.056	0.060	0.105
0.4	0.087	0.197	0.048	0.020	0.184	0.116	0.358	0.029	0.759	0.032	0.180	0.726
0.8	0.261	0.406	0.337	0.043	0.508	0.573	1.000	0.002	1.000	0.015	0.963	1.000

Table 3: Powers of the tests for uniformity using PITs of 1– and 2–step ahead forecast error densities. Data generating process VAR(1), and forecasts are generated from estimated AR(2) models; $T = 200$, $m = 100$.

		Product (p)						Ratio (r)					
			Smooth test						Smooth test				
ρ	h	KS(p)	$u_1^2(p)$	$u_2^2(p)$	$u_3^2(p)$	$u_4^2(p)$	$\Psi_4^2(p)$	KS(r)	$u_1^2(r)$	$u_2^2(r)$	$u_3^2(r)$	$u_4^2(r)$	$\Psi_4^2(r)$
-0.8	1	0.640	0.021	0.998	0.030	0.067	0.990	0.055	0.027	0.045	0.045	0.207	0.087
	2	0.632	0.028	1.000	0.067	0.060	0.994	0.048	0.031	0.059	0.032	0.197	0.097
-0.4	1	0.291	0.106	0.828	0.040	0.042	0.576	0.036	0.015	0.255	0.023	0.111	0.125
	2	0.291	0.098	0.832	0.042	0.025	0.597	0.031	0.011	0.240	0.021	0.117	0.105
0.0	1	0.205	0.180	0.317	0.032	0.048	0.196	0.087	0.010	0.725	0.019	0.043	0.360
	2	0.211	0.160	0.317	0.040	0.052	0.194	0.075	0.007	0.702	0.018	0.041	0.345
0.4	1	0.208	0.248	0.065	0.042	0.043	0.139	0.373	0.003	0.991	0.011	0.056	0.893
	2	0.208	0.246	0.053	0.053	0.059	0.140	0.332	0.003	0.985	0.011	0.048	0.871
0.8	1	0.267	0.340	0.082	0.087	0.069	0.236	0.985	0.002	1.000	0.019	0.844	1.000
	2	0.272	0.320	0.096	0.075	0.062	0.235	0.982	0.000	1.000	0.017	0.815	1.000

Table 4: Powers of the tests using PITs of 1–step ahead forecast error densities. Data generating process VAR(1), and forecasts are generated from estimated AR(2) models; $T = 200$, $m = 100$.

	Product (p)						Ratio (r)					
		Smooth test						Smooth test				
ρ	KS(p)	$u_1^2(p)$	$u_2^2(p)$	$u_3^2(p)$	$u_4^2(p)$	$\Psi_4(p)$	KS(r)	$u_1^2(r)$	$u_2^2(r)$	$u_3^2(r)$	$u_4^2(r)$	$\Psi_4(r)$
-0.8	1.000	0.073	1.000	0.217	0.805	1.000	0.030	0.010	0.053	0.034	0.502	0.233
-0.4	0.889	0.162	0.999	0.101	0.037	0.999	0.020	0.001	0.313	0.033	0.317	0.258
0.0	0.534	0.244	0.921	0.055	0.050	0.879	0.083	0.000	0.856	0.013	0.125	0.650
0.4	0.408	0.388	0.354	0.050	0.072	0.422	0.610	0.000	1.000	0.009	0.036	0.993
0.8	0.454	0.558	0.053	0.065	0.047	0.354	1.000	0.000	1.000	0.001	0.949	1.000

Table 5: Values of KS test, Neyman's smooth test and its components using PITs of 1–step ahead forecast densities. Forecasts are based on a two-regime bivariate threshold model fitted to the minute returns of S&P 500 index futures and prices and the associated threshold variable. p-values are in parenthesis.

Number of forecasts	Product (p)						Ratio (r)					
		Smooth test						Smooth test				
(L)	KS(p)	$u_1^2(p)$	$u_2^2(p)$	$u_3^2(p)$	$u_4^2(p)$	$\Psi_4^2(p)$	KS(r)	$u_1^2(r)$	$u_2^2(r)$	$u_3^2(r)$	$u_4^2(r)$	$\Psi_4^2(r)$
100	0.133	2.333	0.030	0.220	0.127	2.710	0.118	1.192	2.024	5.044	0.590	8.850
		(0.127)	(0.862)	(0.639)	(0.722)	(0.607)		(0.275)	(0.155)	(0.025)	(0.442)	(0.065)
200	0.095	3.577	1.001	0.080	0.118	4.776	0.100	1.881	6.389	1.142	1.518	10.930
		(0.059)	(0.317)	(0.777)	(0.731)	(0.311)		(0.170)	(0.011)	(0.285)	(0.218)	(0.027)
300	0.077	4.128	0.112	2.060	0.278	6.578	0.094	2.420	6.828	0.245	5.971	15.464
		(0.042)	(0.738)	(0.151)	(0.598)	(0.160)		(0.120)	(0.009)	(0.621)	(0.145)	(0.004)

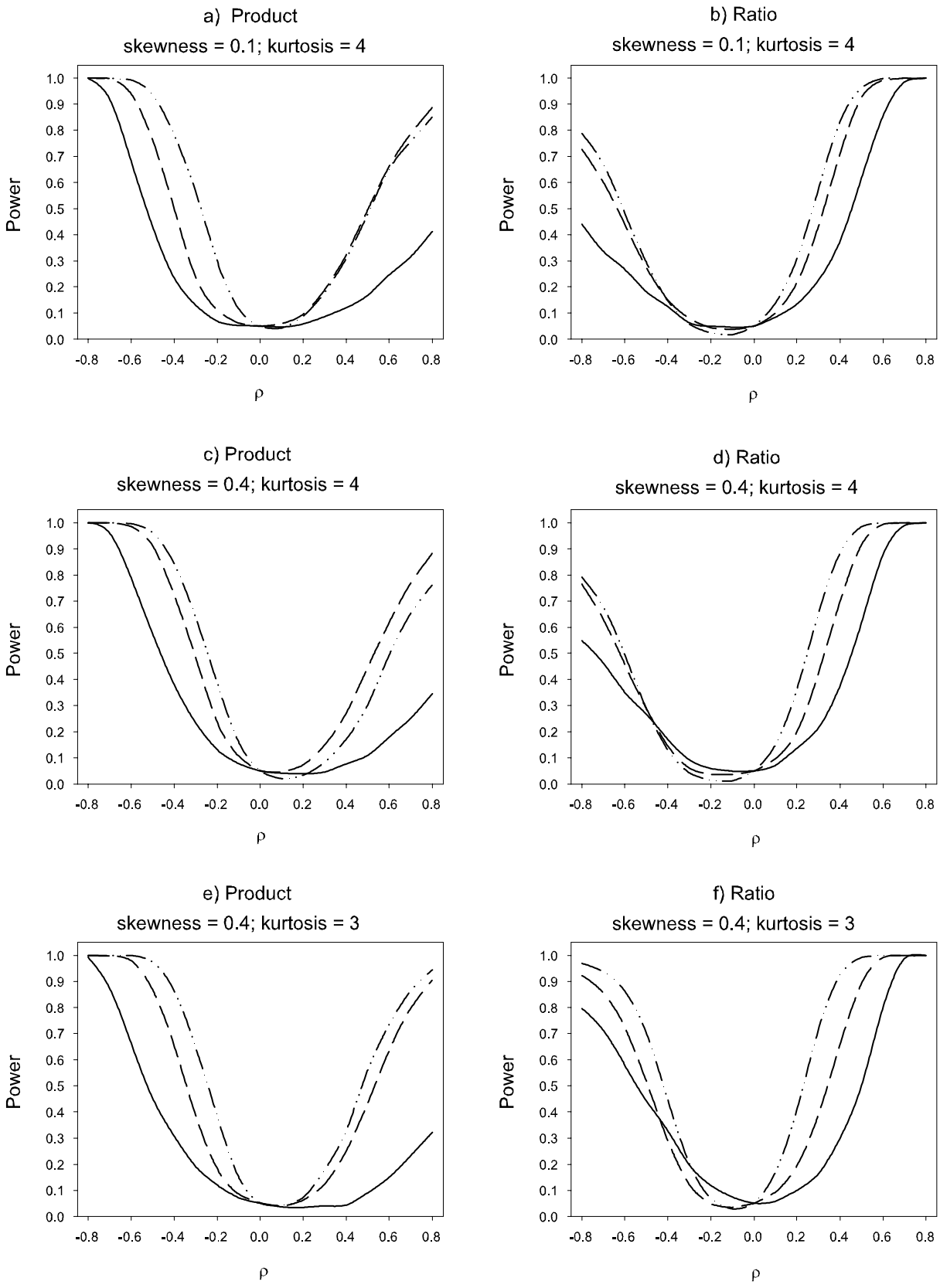

Figure 1: *Size-adjusted estimated power functions of the test statistics* $KS(\cdot)$ *(solid lines),* $\Psi_4^2(\cdot)$ *(medium-dashed lines), and* $u_2^2(\cdot)$ *(dashed-dotted-dotted lines);* $n = 100$.