# Robust Evaluation of Multivariate Density Forecasts*

Jonas Dovern[1,2] and Hans Manner[†3]

[1] Alfred-Weber-Institute for Economics, Heidelberg University
[2] RWTH Aachen University
[3] Institute for Econometrics and Statistics, University of Cologne

February 9, 2016

**Abstract**

We derive new tests for proper calibration of multivariate density forecasts based on Rosenblatt probability integral transforms. These tests have the advantage that they i) do not depend on the ordering of variables in the forecasting model, ii) are applicable to densities of arbitrary dimensions, and iii) have superior power relative to existing approaches. We furthermore develop adjusted tests that allow for estimated parameters and, consequently, can be used as in-sample specification tests. We demonstrate the problems of existing tests and how our new approaches can overcome those using two applications based on multivariate GARCH-based models for stock market returns and on a macroeconomic Bayesian vectorautoregressive model.

*JEL Classification: C12, C32, C52, C53*

*Keywords: Density forecast, multivariate density, Rosenblatt transformation, density calibration*

## 1 Introduction

The use of density forecasts has recently become common in many areas of economics. Density forecasts, which have been used in meteorology for a long time, are increasingly used, for instance, in the fields of energy economics (Huurman et al., 2012), demand management (Taylor, 2012), finance (Shackleton et al., 2010; Kitsul and Wright, 2013; Hallam and Olmo, 2014; Ghosh and Bera, 2015), and macroeconomics (Clark, 2011; Herbst and Schorfheide, 2012; Aastveit et al., 2014; Wolters, 2015). A number of approaches have been suggested to evaluate univariate density forecasts (e. g., Diebold et al., 1998; Berkowitz, 2001; Bai, 2003). Many tasks, such as the computation of Value-at-Risk measures for portfolios containing multiple assets or the planning

[*] We would like to thank the participants of the seminar at Heidelberg University for helpful comments on an early draft of the paper. All errors are our own responsibility.

[†] Corresponding author: Meister-Ekkehart-Str. 9, 50937 Köln, Germany. +49-221-470-4130. manner@statistik.uni-koeln.de

of production for a firm that serves many markets from one central production facility, require the construction and evaluation of *multivariate* densities. Beginning with Diebold et al. (1999), the literature has proposed several approaches for testing whether a sequence of multivariate density estimates coincides with the corresponding true densities (e. g., Clements and Smith, 2000, 2002; Corradi and Swanson, 2006; Bai and Chen, 2008; González-Rivera and Yoldas, 2012; Ko and Park, 2013; Ziegel and Gneiting, 2014). In our view, however, this strand of literature has to date neglected two important issues (Ziegel and Gneiting (2014) being an exception). First, the existing tests depend on the ordering of variables in a multivariate model, an issue that is mentioned in many papers, but that has, until now, not been dealt with. This is highly problematic because it requires the presentation of many different test results (which might lead to inconclusive results) and even makes room for "cheating" if a researcher decides to report only those results which correspond to one particular ("preferred") ordering.[1] Second, all empirical applications and many of the theoretical results focus on the bivariate case. However, systems of higher dimensionality are required to render many applications, especially in finance, useful. We address both issues in this paper.

Following Diebold et al. (1999), the most commonly used approach for testing the calibration of multivariate density forecasts is based on the Rosenblatt (1952) probability integral transform (PIT). It relies on the factorization of the multivariate density into conditional distributions because these, in turn, can be used to form independent PITs which, for well-specified models, follow a uniform distribution.[2] Suitable transformations of these conditional PITs then lead to a reduction of the multivariate testing problem to a univariate one. These univariate tests can be implemented using any goodness-of-fit test (e. g., of the Kolmogorov type or of the smooth type). How well a testing approach works depends crucially on the chosen transformation.

Diebold et al. (1999) suggest stacking all conditional PITs, which yields a sample of $dn$ independent PITs (where $d$ refers to the dimension of the density and $n$ is the size of the sample used to evaluate the density forecasts). Clements and Smith (2000) propose to use the sequence of products of the conditional PITs because this preserves the temporal connection between PITs from one period and increases power against misspecification of the correlation between the model variables. For the bi-variate case, Clements and Smith (2002) suggest using the ratio of the conditional PITs because this improves on the "product approach" whenever the correlation exceeds that implied by the null hypothesis. In reality, however, the true correlation structure might be unknown, i. e., it might be impossible to determine whether the "product approach" or the "ratio approach" should be preferred; Ko and Park (2013) propose a location-adjusted version of the "product approach" that has superior power in both cases.

To test whether the transformations follow the distribution implied by the null hypothesis, a number of papers have suggested goodness-of-fit tests tailored to the context of evaluating density forecasts based on time series models. In particular, these papers deal with the issues of parameter uncertainty and dynamic misspecification. We want to highlight two of them: using

[1] Essentially, this is one form of "data snooping" as discussed in White (2000).

[2] Henceforth, we use the term 'conditional distributions' in a way that includes the one marginal distribution that is needed for the factorization of the joint density. In addition, we will refer to the PITs of the conditional distributions as *conditional PITs*.

the Khmaladze martingale transformation, Bai and Chen (2008) construct different distribution-free test statistics based on the conditional PITs which eliminates the effects of parameter estimation.[3] Corradi and Swanson (2006) use a bootstrap to construct a goodness-of-fit test that allows for dynamic misspecification under the null hypothesis.

In this paper, we contribute to the literature on density forecast evaluation in the following way. We propose new transformations of the conditional PITs which can be combined with *any* goodness-of-fit test for univariate distributions. The new transformations have a number of properties that renders them preferable to existing approaches. First, they are *order invariant*, a concept we define below. Second, they are applicable to densities of arbitrary dimension in contrast to the existing approaches that generally focus on the bivariate case. Third, they have better power against a wide range of alternatives. Furthermore, for the case of multivariate normal distributions, we propose adjusted transformations, based on the idea of randomization proposed by Durbin (1961), which can be used when parameters have to be estimated and evaluation is done in-sample. This makes our approach a suitable tool also for specification testing, even though in this paper we focus on the evaluation of out-of-sample forecasts. Our preferred transformations (called $Z_t^{2*}$ and $Z_t^{2\dagger}$ below) are constructed as the sum of squares of normal transformations of conditional PITs corresponding to all possible orderings of the variables and are distributed as a mixture of chi-squared random variables.

Our main results are as follows. First, the distortions in rejection rates caused by a tendentious application of tests which are not order invariant ("cheating") can be very substantial. Second, the order-invariant tests that we propose have better power properties relative to existing tests against a wide range of deviations from the null model. Third, this remains the case when parameter uncertainty is taken into account. Finally, we show that the new tests are helpful for testing the appropriateness of density forecasts based on sophisticated multivariate models for financial returns and for evaluating multivariate macroeconomic forecasts. In particular, we show that the potential for "cheating" is immense in these applications and that our robust tests are required to draw unambiguous conclusions.

The remainder of this paper is organized as follows. In Section 2, we introduce the notation, generalize existing tests, and derive new tests to evaluate multivariate densities. In Section 3, we assess the properties of a wide range of tests by means of Monte Carlos simulations. In Section 4, we demonstrate the usefulness of the newly proposed tests using two empirical applications. Section 5 concludes. The Appendix contains all proofs, descriptions of univariate goodness-of-fit tests, and additional simulation results.

## 2 Theory

Consider a vector valued time series $\{Y_t\}_{t=1}^n = \{[Y_{1,t}, \dots, Y_{d,t}]\}_{t=1}^n$ with true conditional density $f_{Y_t}(Y_t|\Omega_{t-1})$, where $\Omega_{t-1}$ denotes the information set available at time $t-1$. Suppose that we have a density forecast $\hat{f}_{Y_t}(Y_t|\Omega_{t-1})$ with corresponding cumulative density function (CDF)

[3] Note that by looking at transformations of test statistics for individual conditional PITs, their approach is somewhat different from the more commonly used approach that looks at test statistics for transformations of conditional PITs.

$\hat{F}_{Y_t}(Y_t|\Omega_{t-1})$. Furthermore, let $\hat{F}_{Y_i}(Y_{i,t}|\Omega_{t-1})$ denote the forecast for the $i^{th}$ marginal distribution function and denote by $\hat{F}_{Y_i|Y_{i-1},\dots,Y_1}(Y_{i,t}|Y_{i-1,t},\dots,Y_{1,t},\Omega_{t-1})$ the predictive conditional distribution of $Y_{i,t}$ given $Y_{i-1,t},\dots,Y_{1,t}$.

We are interested in testing the null hypothesis that the forecast density coincides with the true density, or formally:[4]

$$H_0 : \hat{f}_{Y_t}(Y_t|\Omega_{t-1}) = f_{Y_t}(Y_t|\Omega_{t-1}) \tag{2.1}$$

One important condition for $H_0$ to be true is that the density forecasts have to be *properly calibrated*. The latter term refers to the consistency between the realized values of $Y_t$ and $\hat{f}_{Y_t}$ (Gneiting et al., 2007). For the univariate case it can be shown that $H_0$ implies that the probability integral transform (PIT) of $Y_t$ with respect to $\hat{f}_{Y_t}$, given by $\hat{F}_{Y_t}(Y_t)$, is uniformly distributed between 0 and 1. The latter fact can be used to test for proper density calibration in the univariate case (e. g., Dawid, 1984; Diebold et al., 1998). The uniformity can be checked either by graphical methods, such as QQ-plots and histograms, or by goodness-of-fit tests, such as the Kolmogorov-Smirnov test, the Anderson-Darling test, or Neyman's smooth test.

Unfortunately, matters are more complicated in the multivariate case because the distribution of the multivariate PITs of $Y_t$, i.e., the distribution of the random variable $F(Y_t)$, is in general unknown for $d > 1$; see, e. g., Genest and Rivest (2001). In essence, the task then is to reduce the multivariate problem to a univariate one by using suitable transformations. One way to approach this problem, proposed in Ziegel and Gneiting (2014), is to work with the Kendall distribution function for $\hat{F}_{Y_t}(Y_t|\Omega_{t-1})$, given by

$$\mathcal{K}_{\hat{F}}(w) = Pr\left(\hat{F}_{Y_t}(Y_t|\Omega_{t-1}) \leq w\right) \quad \text{for} \quad w \in [0,1],$$

and the corresponding copula probability integral transform. It remains a problem, however, that $\mathcal{K}_{\hat{F}}$ is available in closed form only for special cases and, in general, needs to be approximated by simulation.

The other (more commonly used) way to approach this problem is based on the result by Rosenblatt (1952) that relies on the factorization of the joint densities into the product of conditional densities

$$\hat{f}_{Y_t}(Y_t) = \hat{f}_{Y_d|Y_{d-1},\dots,Y_1}(Y_{d,t}) \times \ldots \times \hat{f}_{Y_2|Y_1}(Y_{2,t}) \times \hat{f}_{Y_1}(Y_{1,t}). \tag{2.2}$$

Then the sequences of *conditional PITs* for the elements of $Y_t$

$$\begin{aligned} U_t^1 &= \hat{F}_{Y_1}(Y_{1,t}), \\ U_t^{2|1} &= \hat{F}_{Y_2|Y_1}(Y_{2,t}), \\ &\vdots \\ U_t^{d|d-1,\dots,1} &= \hat{F}_{Y_d|Y_{d-1},\dots,Y_1}(Y_{d,t}) \end{aligned} \tag{2.3}$$

[4] To simplify notation, we henceforth suppress the dependence on $\Omega_{t-1}$ and the conditioning variables in the function arguments.

are independent of each other and distributed $\mathcal{U}(0,1)$. In other words, transforming each element of $Y_t$ by the corresponding predictive conditional cumulative density yields independent sequences of uniformly distributed random variables.

Diebold et al. (1999) then achieve the reduction of dimension by stacking all conditional PITs, noting that this produces a sequence of random variables which are distributed $\mathcal{U}(0,1)$. More formally, if we let

$$S_t = [U_t^{d|d-1,\dots,1}, \dots, U_t^1]', \tag{2.4}$$

then $S = [S_1', S_2', \dots, S_n']'$ constitutes a vector of variables which are uniformly distributed under $H_0$.

Instead of stacking the conditional PITs, other approaches advocate transforming the vector-valued random variable $Y_t$ into a scalar random variable and then computing PITs for this transformed random variable. This is also the approach that we use when developing our new tests. To formalize the idea, consider the general transform function $g_t(\cdot) : \mathbb{R}^d \to \mathbb{R}$ and define the transformed series $W_t = g_t(Y_t)$ with distribution function $F_{W_t}$ estimated by $\hat{F}_{W_t}$. The PIT of $W_t$ is given by

$$U_t^W = \hat{F}_{W_t}(W_t). \tag{2.5}$$

Testing $H_0$ then is equivalent to testing whether $U_t^W \sim \mathcal{U}(0,1)$. Well-known tests can be used to implement this. Below, we rely on Neyman's smooth test (Neyman, 1937), the Kolmogorov-Smirnov test and a test suggested by Knüppel (2015).[5] The tests are reviewed in Appendix B and in Section 2.4.

For one-step-ahead density forecast—this is what we have referred to so far—the PITs are also independently distributed across time under $H_0$, i. e., $U_t \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$. Mitchell and Wallis (2011) call density forecasts which satisfy both features *completely calibrated*. We revisit the issue of multi-step forecasts, which cause the PITs to be autocorrelated, in Section 2.4.

Following the seminal contribution of Diebold et al. (1999), different transformations $g_t(\cdot)$ have been considered in the literature to test for the proper calibration of densities. Clements and Smith (2000) propose using the product of the conditional PITs, Clements and Smith (2002) (for the bi-variate case) look at their ratio, while Ko and Park (2013) advocate using the product of the demeaned conditional PITs. We reconsider these approaches in Section 2.2 where we generalize them to the case of densities with arbitrary dimensions, before suggesting, as we argue, preferable transformations.

## 2.1 Ordering of the Variables

So far, we have implicitly assumed that there exists a natural ordering of variables from 1 to $d$. This, of course, is not really true and, as already mentioned in most papers on the topic (Diebold et al., 1999; Clements and Smith, 2002; Hong and Li, 2005; Ishida, 2005), sorting the elements in $Y_t$ in a different way–and, thus, factorizing the multivariate density in a different way–will generally lead to different results. Specifically, the Rosenblatt transform in (2.3) clearly

[5] The Anderson-Darling test, which is usually reported to have good properties for univariate testing problems, showed surprisingly poor power in preliminary simulations and was therefore not considered.

depends on the ordering of the variables in $Y_t$. There are $d!$ different orderings of the variables, leading to different conditional PITs. Consequently, the outcome of a hypothesis test based on the transformed variable will depend on the selected ordering. This is an undesirable property for a test since a researcher who is interested in supporting or discrediting a certain model may perform the hypothesis test for all distinct orderings and only report the outcome with the largest or smallest p-value. Note that while it is certainly true that for low-dimensional cases results for all possible permutations can be presented and discussed, this becomes quickly impossible for larger $d$. In addition, even when multiple test statistics are presented, it is unclear how an overall decision should be formed based on those.

We use the following notation for different permutations of the variables. Let $\pi_k$ for $k = 1, \ldots, d!$ be the set of all possible permutations of the data. Furthermore, let $\pi_k(i)$ denote the index (or "position") of variable $i$ in the $k^{th}$ permutation. Then, the conditional PITs under permutation $\pi_k$ are given by

$$\begin{aligned} U_t^{\pi_k(1)} &= \hat{F}_{Y_{\pi_k(1)}}(Y_{\pi_k(1),t}) \\ U_t^{\pi_k(2)|\pi_k(1)} &= \hat{F}_{Y_{\pi_k(2)}|Y_{\pi_k(1)}}(Y_{\pi_k(2),t}) \\ &\vdots \\ U_t^{\pi_k(d)|\pi_k(d-1),\ldots,\pi_k(1)} &= \hat{F}_{Y_{\pi_k(d)}|Y_{\pi_k(d-1),t},\ldots,Y_{\pi_k(1),t}}(Y_{\pi_k(d),t}). \end{aligned} \tag{2.6}$$

The following definition formalizes the concept that the exact permutation of the data is not relevant for the test outcome.

**Definition 1.** *Let $T(\pi_k)$ be a test statistic based on $\{Y_t\}_{t=1}^n$ under permutation $\pi_k$. We call a test statistic $T(\pi_k)$* ***order invariant*** *if $T(\pi_k) = T(\pi_j)$, $\forall\, k \neq j$.*

In the next section, we show that existing tests are order invariant only under very restrictive conditions and we derive new tests that are always order invariant.

### 2.2 Tests Based on the Rosenblatt Transformation

In this section, we generalize existing tests (Clements and Smith, 2000; Ko and Park, 2013) for the case of an arbitrarily dimensioned density and, derive new tests which we consider to be preferable because they are order invariant and, as shown below, have better power properties in a wide range of situations.

Clements and Smith (2000) propose to evaluate density forecasts based on the product of the conditional PITs corresponding to one particular permutation of the variables[6]. In this case, the transformation function $g_t(\cdot)$ is given by

$$P_{t,d} = g(Y_t) = \prod_{i=1}^{d} U_t^{i|1:i-1}, \tag{2.7}$$

[6] Clements and Smith (2002) consider their ratio to be an alternative. We do not discuss this transformation since it is not obvious how to extend the ratio to higher dimensions.

where $U_t^{i|1:i-1}$ denotes the conditional probability integral transform of variable $Y_{i,t}$ given the variables $Y_{1,t}$ to $Y_{i-1,t}$, and is defined as $U_t^1$ for $i = 1$. The authors derive the distribution of $P_{t,d}$ for $d = 2, 3$. In the following proposition we generalize these results for arbitrary $d$.

**Proposition 1.** *Let $P_{t,d}$ be given by the expression in (2.7). Under $H_0$ it has the following probability density function (PDF) and CDF:*

$$f_{P_d}(P_{t,d}) = \frac{(-1)^{d-1}}{(d-1)!} \log^{d-1}(P_{t,d}) \tag{2.8}$$

$$F_{P_d}(P_{t,d}) = P_{t,d} \sum_{i=0}^{d-1} f_{P_{d-i}}(P_{t,d}) \tag{2.9}$$

Note that for $d = 2, 3$ the density derived in Clements and Smith (2000) is recovered.

Ko and Park (2013) explain why tests based on $P_{t,d}$ have good power only against correlations lower than the hypothesized value. They suggest a location-adjusted version which does not suffer from this asymmetry and which is given by

$$P^*_{t,d} = g(Y_t) = \prod_{i=1}^{d} (U_t^{i|1:i-1} - 0.5). \tag{2.10}$$

Ko and Park only consider the case $d = 2$. We generalize their results to any value of $d$ in the following proposition.[7]

**Proposition 2.** *Let $P^*_{t,d}$ be given by the expression in (2.10). Under $H_0$ it has the following PDF and CDF:*

$$f_{P^*_d}(P^*_{t,d}) = \frac{2^{d-1}}{(d-1)!} \log^{d-1} \left| \frac{1}{2^d P^*_{t,d}} \right|$$

$$F_{P^*_d}(P^*_{t,d}) = P^*_{t,d} 2^{d-1} \sum_{i=1}^{d} \frac{1}{(d-i)!} \log^{d-i} \left| \frac{1}{2^d P^*_{t,d}} \right| + \frac{1}{2}$$

Although the distributions of $P_{t,d}$ and $P^*_{t,d}$ depend on the dimension of the application, we suppress the index $d$ in what follows. Below, we also refer to these approaches by $P$ and $P^*$.

Tests based on the transformations suggested by Diebold et al. (1999), Clements and Smith (2000), and Ko and Park (2013) are not, in general, insensitive to the choice of the permutation. In the following proposition, we show under which conditions these three transformations are order invariant.

**Proposition 3.** *Test statistics $T(\pi_k)$ based on $\{P_t\}_{t=1}^n$, $\{P_t^*\}_{t=1}^n$ and on the stacked transformation $\{S_t\}_{t=1}^n$ are order invariant if and only if under $H_0$ the variables $Y_{1,t}, \ldots, Y_{d,t}$ are independent, i. e., when $\hat{f}_{Y_t}(Y_t) = \hat{f}_{Y_1}(Y_{1,t}) \times \ldots \times \hat{f}_{Y_d}(Y_{d,t})$.*

[7] Note that the density given in the appendix of Ko and Park (2013) needs to be multiplied by a factor of 2.

We continue by introducing a transformation that leads to order-invariant test statistics under less restrictive conditions and forms the basis for additional transformations which always lead to order invariant tests. Consider the transformation

$$Z_{t,d}^2 = \sum_{i=1}^{d} \left( \Phi^{-1} \left( U_t^{i|1:i-1} \right) \right)^2 , \tag{2.11}$$

where $\Phi$ denotes the CDF of the standard normal distribution. $H_0$ implies that $Z_{t,d}^2 \sim \chi_d^2$, where $\chi_d^2$ denotes the chi-squared distribution with $d$ degrees of freedom. Denoting by $F_{\chi_d^2}$ the CDF of this distribution, $U_t^{Z^2} = F_{\chi_d^2}(Z_{t,d}^2)$ is distributed $\mathcal{U}(0,1)$ under $H_0$.[8] As the following proposition shows, tests based on $Z_t^2$ are order invariant under normality.

**Proposition 4.** *Test statistics $T(\pi_k)$ based on $\{Z_t^2\}_{t=1}^n$ are order invariant if under $H_0$ $Y_t \sim \mathcal{N}(\mu, \Sigma)$, i. e., when $Y_t$ follows a multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$.*

**Remark:** The proof in Appendix A shows that under the null hypothesis of normality it holds that $Z_t^2 = (Y_t - \mu)'\Sigma^{-1}(Y_t - \mu)$, which is the transformation proposed by Ishida (2005). Of course, $Z_t^2$ can also be used to test non-Gaussian densities. In this case, however, the corresponding test statistics are not generally order invariant, except for the obvious case of independence.

Ideally, we would like to have a transformation that is order invariant in general. Such a transformation can be constructed as follows. Consider all possible permutations $\pi_k$ for $k = 1, \ldots, d!$ of the variables and the corresponding sequences of conditional PITs defined by (2.6). This yields a total of $d \times d!$ terms. However, only $d \times 2^{d-1}$ of those terms are distinct. To see why, note that there are $d$ variables that can each be ordered first to last. When a particular variable is ordered second, there are $\binom{d-1}{1} = d-1$ possible conditioning variables. When this variable is ordered third, there are $\binom{d-1}{2}$ distinct sets of conditioning variables (each containing two of the other variables), and so on. Finally, when the variable is ordered last, the number of conditioning sets is $\binom{d-1}{d-1} = 1$. Therefore, the overall number of distinct PITs is $d \times \sum_{k=0}^{d-1} \binom{d-1}{k} = d \times 2^{d-1}$.

The transformation that we propose is similar in structure to $Z_t^2$ defined by (2.11) but considers the sum over all distinct conditional PITs derived from all possible permutations of the variables. To formalize, let $\gamma_i^k$ for $k = 1, \ldots, 2^{d-1}$ be the set of all sets of conditioning variables (including the empty set) corresponding to all distinct conditional PITs for $Y_{i,t}$. Then the suggested transformation has the form

$$Z_t^{2*} = \sum_{i=1}^{d} \sum_{k=1}^{2^{d-1}} \left( \Phi^{-1} \left( U_t^{i|\gamma_i^k} \right) \right)^2 . \tag{2.12}$$

Since all distinct conditional PITs enter into this transformation, order invariance is ensured for any test statistic based on $Z_t^{2*}$. However, due to the fact that the terms in the sum are not independent in general, $Z_t^{2*}$ does not follow a $\chi^2$ distribution under $H_0$. The following proposition gives its distribution under normality.

[8] Again, we'll denote the transformation as $Z_t^2$ and refer to the approach by $Z^2$.

**Proposition 5.** *Let $Y_t \sim \mathcal{N}(\mu, \Sigma)$. Then $Z_t^{2*}$ is distributed as $\sum_{i=1}^{d} \lambda_i Z_i^2$, for independent $\mathcal{N}(0,1)$ variables $Z_1, \ldots, Z_d$ and $\lambda_1, \ldots, \lambda_d$ the non-zero eigenvalues of the rank $d$ matrix $R_{Z^*}$, which is the correlation matrix of all distinct terms $\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right) \forall\, i, k$ entering $Z_t^{2*}$. A typical entry of $R_{Z^*}$ is given by*

$$Corr\left(\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right), \Phi^{-1}\left(U_t^{j|\gamma_j^l}\right)\right) = (\Sigma_{i,i} - \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,i})^{-1/2}(\Sigma_{j,j} - \Sigma_{j,\gamma_j^l}\Sigma^{-1}_{\gamma_j^l,\gamma_j^l}\Sigma_{\gamma_j^l,j})^{-1/2}\times$$
$$(\Sigma_{i,j} - \Sigma_{j,\gamma_j^l}\Sigma^{-1}_{\gamma_j^l,\gamma_j^l}\Sigma_{\gamma_j^l,i} - \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,j} + \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,\gamma_j^l}\Sigma^{-1}_{\gamma_j^l,\gamma_j^l}\Sigma_{\gamma_j^l,j}),$$

*where the $\Sigma_{r,c}$ $(r, c \in \{i, \gamma_i^k\})$ are scalars, vectors, and matrices containing those elements of $\Sigma$ that are defined by the row(s) corresponding to the variable(s) defined by $r$ and the column(s) corresponding to the variable(s) defined by $c$.*

**Remark:** When $Y_t$ does not follow a multivariate normal distribution, the above result does not hold. In this case, the terms $\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right)$ do not jointly follow a multivariate normal distribution (even though they are marginally normally distributed) because their dependence structure is unknown in general. Such cases include non-Gaussian parametric distributions but also situations in which the analytic form of the distribution is unknown and the PITs are based on non-parametric estimations of conditional and marginal densities.

In those cases, the distribution of $Z_t^{2*}$ can be obtained straightforwardly by Monte Carlo simulation, as long as it is possible to generate random draws from the hypothesized model, which is, e. g., the case for models estimated with Bayesian methods. The Monte Carlo simulation can be used to obtain a nonparametric approximation to the distribution of $Z_t^{2*}$ along the following lines:

1. For each period in the evaluation sample ($t$), generate $B'$ distinct conditional forecasts, $\hat{y}_t^{(b)}$, based on the model under $H_0$. These forecasts should reflect the same kinds of uncertainty that are also taken into account during the construction of the conditional predictive densities $\hat{f}_{y_t}(y_t)$.

2. Given $\hat{f}_{y_t}(y_t)$, compute $\Phi^{-1}\left(U_{t,(b)}^{i|\gamma_i^k}\right)$, $\forall i, k$, for all $B'$ simulated conditional forecasts. In other words, compute all distinct "inverse" conditional and marginal PITs for each of the simulated conditional forecasts obtained under $H_0$.

3. Based on the set of $\Phi^{-1}\left(U_{t,(b)}^{i|\gamma_i^k}\right)$, compute $Z_{t,(b)}^{2*}$, i. e., construct a set of $B'$ transformed statistics under $H_0$.

4. Compute $U_t^{Z^{2*}} = Pr\left(Z_t^{2*} < Z_{t,(b)}^{2*}\right)$ by simply counting how often the transformed statistic based on the actual realizations is smaller than the transformed statistics based on conditional forecasts that are generated under $H_0$.

5. Apply preferred test to the sequence of $U_t^{Z^{2*}}$ to test the null hypothesis of a $\mathcal{U}(0,1)$ distribution.

When $d$ is very large, the number of terms entering $Z_t^{2*}$ can become very large.[9] In this case, it appears sensible to use a transformation in which the number of terms grows only linearly with $d$. A transformation that is always order invariant can be obtained by considering only those conditional PITs corresponding to the distribution of $Y_{i,t}$ for $i = 1, \dots, d$ conditional on all other variable. Denoting those conditional PITs by $U_t^{i|-i}$, where $i| - i$ denotes variable $i$ conditional on the set $\{1, \dots, d\} \backslash i$, the transformation is given by

$$Z_t^{2\dagger} = \sum_{i=1}^{d} \left( \Phi^{-1} \left( U_t^{i|-i} \right) \right)^2 . \tag{2.13}$$

Under normality of $Y_t$, the distribution $Z_t^{2\dagger}$ is given by the following corollary to Proposition 5.

**Corollary 1.** *Let $Y_t \sim \mathcal{N}(\mu, \Sigma)$. Then $Z_t^{2\dagger}$ is distributed as $\sum_{i=1}^{d} \lambda_i Z_i^2$, for independent $\mathcal{N}(0,1)$ variables $Z_1, \dots, Z_d$ and $\lambda_1, \dots, \lambda_d$ the eigenvalues of the matrix $R_{Z^\dagger}$, which is the correlation matrix of all terms $\Phi^{-1}\left(U_t^{i|-i}\right)$ for $i = 1, \dots, d$ entering $Z_t^{2\dagger}$. A typical entry of $R_{Z^\dagger}$ is given by*

$$\begin{aligned} Corr\left(\Phi^{-1}\left(U_t^{i|-i}\right), \Phi^{-1}\left(U_t^{j|-j}\right)\right) &= (\Sigma_{i,i} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i})^{-1/2}(\Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j})^{-1/2} \times \\ &(\Sigma_{i,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,i} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,j} + \Sigma_{i,-i}\Sigma_{\gamma_i^k,-i}^{-1}\Sigma_{-i,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}), \end{aligned}$$

*where the index $-i$ denotes all rows/columns of $\Sigma$ except for the $i^{th}$ one.*

This transformation still uses information from all variables but fewer terms enter into the statistic. In the non-Gaussian case, the same approach as for $Z_t^{2*}$ can be used to approximate the distribution of $Z_t^{2\dagger}$. Our simulations show that neither $Z_t^{2*}$ nor $Z_t^{2\dagger}$ is superior in general. Below, we refer to tests based on the latter two transformations by $Z^{2*}$ and $Z^{2\dagger}$.

### 2.3 Estimated Parameters

So far, we have neglected the issue of parameter uncertainty because in this paper we are mainly concerned with the out-of-sample evaluation of predictive densities. Treating the density forecasts as primitives as suggested by Berkowitz (2001), we can, therefore, abstract from parameter uncertainty. We simply test whether the density forecasts are properly calibrated without any reference to the density generating model. Therefore, we can assume for all transformations presented in Section 2.2 that the parameters under the null model are known. In contrast, when testing the in-sample goodness-of-fit of the hypothesized model, the parameters need to be estimated from the same sample used to evaluate the model fit and one implicitly tests a composite hypothesis. This changes the distribution of the corresponding test statistics. Ignoring this issue will, in general, lead to undersized tests. Bai (2003) and Bai and Chen (2008) overcome this problem by relying on the Khmaladze transformation, whereas Andrews (1997) solves this problem by using a parametric bootstrap. The latter solution could easily be applied to our tests for testing general model specifications.

[9] With $d = 10$, for instance, the number of terms equals 5,120.

However, here we suggest a simple method which can be used to adjust the transformations studied above to take into account the estimation of parameters when $Y_t$ follows a multivariate normal distribution. The idea is to apply an appropriate randomization to the transformations, which offsets the effect of using estimated instead of true parameters. This idea dates back to Durbin (1961) and has been studied in Wagle (1968), González-Barrios et al. (2010), and Szkutnik (2012).

To see the general idea, let $\{Y_{i,t}\}_{t=1}^n$ be $N(\mu_i, \sigma_i^2)$ distributed and let

$$\hat{Z}_{i,t} = \frac{Y_{i,t} - \hat{\mu}_i}{\hat{\sigma}_i}, \tag{2.14}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the usual sample mean and standard deviation. The distribution of $\hat{Z}_{i,t}$ is not standard normal due to the standardization with the sample mean and standard deviation but has the distribution given in David and Johnson (1948). Now let $m$ be a random variable drawn from a $\mathcal{N}(0, 1/n)$, and let $s^2$ be a random variable drawn from a $\chi^2_{n-1}$ divided by $n-1$. Then Durbin (1961) shows that the sequence $\{\hat{Z}'_{i,t} = s\hat{Z}_{i,t} + m\}_{t=1}^n$ is i.i.d. $\mathcal{N}(0, 1)$ distributed.

Let us start with the distribution of $Z_t^2$ when $\Sigma = \mathrm{Cov}(Y_t)$ is replaced by the sample covariance matrix $\widehat{\Sigma}$ and $\mu = E(Y_t)$ is replaced by the sample mean. First, note that in the present situation $Z_t^2$ can be written as

$$Z_t^2 = \sum_{i=1}^{d} \left( \frac{Y_{i,t|1:i-1} - \mu_{i|1:i-1}}{\sigma_{i|1:i-1}} \right)^2, \tag{2.15}$$

where $Y_{i|1:i-1}$ denotes $Y_{i,t}$ conditional on $Y_{1,t}, \ldots, Y_{i-1,t}$, and $\mu_{i|1:i-1}$ and $\sigma_{i|1:i-1}$ are the corresponding conditional mean and standard deviation. In practice, however, these are replaced by the estimators $\hat{\mu}_{i|1:i-1}$ and $\hat{\sigma}_{i|1:i-1}$ giving the feasible form

$$\hat{Z}_t^2 = \sum_{i=1}^{d} \left( \frac{Y_{i,t|1:i-1} - \hat{\mu}_{i|1:i-1}}{\hat{\sigma}_{i|1:i-1}} \right)^2, \tag{2.16}$$

which does not follow a $\chi^2_d$ distribution. The following proposition shows how Durbin's randomization can be used to recover that distribution.

**Proposition 6.** *Let $Y_t \sim \mathcal{N}(\mu, \Sigma)$ with $\mu$ and $\Sigma$ unknown. Let $m_i \sim \mathcal{N}(0, 1/n)$ and $s_i^2 \sim \chi^2_{n-1}/(n-1)$ for $i = 1, \ldots, d$ independent of each other. Let $\hat{Z}_{i,t|1:i-1} = \frac{Y_{i,t|1:i-1} - \hat{\mu}_{i|1:i-1}}{\hat{\sigma}_{i|1:i-1}}$. Then $\tilde{Z}_t^2 = \sum_{i=1}^d \left( s_i \hat{Z}_{i,t|1:i-1} + m_i \right)^2$ follows a $\chi^2_d$ distribution.*

The transformations $P_t$ and $P_t^*$, as well as the stacked transformation $S_t$ by Diebold et al. (1999) can be adjusted similarly in this situation. Let

$$\tilde{U}_t^{i|1:i-1} = \Phi(s_i \hat{Z}_{i,t|1:i-1} + m_i).$$

Then the transformations $\tilde{S}_t$, $\tilde{P}_t$, and $\tilde{P}_t^*$ are defined as in (2.4), (2.7), and (2.10) with $U_t^{i|1:i-1}$ replaced by $\tilde{U}_t^{i|1:i-1}$.

Next we turn to the distributions of $Z_t^{2*}$ and $Z_t^{2\dagger}$ when parameters are estimated.

**Proposition 7.** *Let $Y_t \sim \mathcal{N}(\mu, \Sigma)$ with $\mu$ and $\Sigma$ unknown. Let $R_{Z^*}$ be the correlation matrix defined in Proposition 5. Let $m = (m_1, \ldots, m_{d \cdot d^{n-1}}) \sim MV\mathcal{N}(0, \frac{1}{n} R_{Z^*})$ and $s^2 = (s_1^2, \ldots, s_{d \cdot d^{n-1}}^2) = diag(S)$, with $S \sim \mathcal{W}(R_{Z^*}, n-1)/(n-1)$ a random matrix from a Wishart distribution. Then $\tilde{Z}_t^{2*} = \sum_{i=1}^{d} \sum_{k=1}^{2^{d-1}} \left( s_{i|\gamma_i^k} \hat{Z}_{it|\gamma_i^k} + m_{i|\gamma_i^k} \right)^2$ is distributed as $\sum_{i=1}^{d} \lambda_i Z_i^2$, for independent $\mathcal{N}(0,1)$ variables $Z_1, \ldots, Z_d$ and $\lambda_1, \ldots, \lambda_d$ the non-zero eigenvalues of the matrix $R_{Z^*}$.*

The transformation $\tilde{Z}_t^{2\dagger}$ is defined analogously and its distribution follows straightforwadly from Proposition 7 and Corollary 1. Note, however, that the result in Proposition 7 is not directly applicable since the distribution of $\tilde{Z}_t^{2*}$ depends on the matrix $R_{Z^*}$ through the random draws $m$ and $s^2$, as well as the eigenvalues $\lambda_1, \ldots, \lambda_d$, which is unknown if $\Sigma$ is unknown. However, a feasible transformation can be computed based on the estimated covariance matrix $\hat{\Sigma}$, which does not affect the transformation asymptotically.

**Corollary 2.** *The result of Proposition 7 continues to hold as $n \to \infty$ when $R_{Z^*}$ is replaced by $\hat{R}_{Z^*}$, computed as in Proposition 5 based on a consistent estimator $\hat{\Sigma}$.*

Our simulations below show that using the estimated covariance matrix to apply Proposition 7 also works well in finite samples with sample sizes as small as 50 observations. Note that the approach we describe above can in principle be adapted to non-Gaussian models. However, this is a non-trivial task and we leave it for future research.

### 2.4 Autocorrelation

Above, we assume that the PITs are i.i.d. which implies that they are independent across time. This, however, is only true in general if the model that is used to generate the predictive densities is not dynamically misspecified and/or if we restrict ourselves to one-step-ahead forecasts. Whenever one of the two conditions is violated, the sequences of PITs are subject to some form of autocorrelation.

We argue that dynamic misspecification is no big concern in practice because forecasting models, in general, can be appropriately specified without major costs. To check whether this is indeed the case, one can either pre-test for autocorrelations in the sequence of PITs, $\{U_t^W\}_{t=1}^n$, or use tests that are designed for testing the joint hypothesis of properly calibrated densities and autocorrelation-free PITs. The latter approach is, for instance, put forward by Berkowitz (2001), who develops a joint likelihood ratio test for $H_0 : \Phi^{-1}(U_t) \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, or Hong and Li (2005), who develop a nonparametric specification test for the same hypothesis in the context of continuous-time models.

In contrast, PITs based on $h$-step-ahead density forecasts will generally follow a moving average process of order $h-1$. Such multi-step forecasts are frequently required by decision makers and therefore need to be produced by forecasters; see Section 4.2 below. Thus, in this case the autocorrelation is “a feature and not a bug” and one needs to deal with it by using tests that take this form of autocorrelation into account. One test that allows us to account for autocorrelation in a straightforward way was recently suggested by Knüppel (2015). This test is based on a set of raw moments of some suitable transformation of the PITs. To be specific, Knüppel (2015) suggests transforming the PITs into standard uniformly distributed variables (with mean 0, unit variance, skewness of 0, and kurtosis of 1.8) using $\tilde{U}_t^W = \sqrt{12}\left(U_t^W - \frac{1}{2}\right)$. Below, we follow this suggestion. Denoting $\hat{m}_r = 1/n \sum_{t=1}^n \left(\tilde{U}_t^W\right)^r$ and $m_r = E\left[\left(\tilde{U}_t^W\right)^r\right]$, a number of $N$ differences between empirical and expected raw moments is collected in one vector

$$\hat{\mathbf{D}}_{r_1 r_2 \ldots r_N} = \begin{bmatrix} \hat{m}_{r_1} - m_{r_1} \\ \hat{m}_{r_2} - m_{r_2} \\ \vdots \\ \hat{m}_{r_N} - m_{r_N} \end{bmatrix}. \tag{2.17}$$

Again, we follow Knüppel's suggestion and use the first four moments to form $\hat{\mathbf{D}}_{1234}$. Knüppel (2015) shows that under $H_0$ (and subject to some mild conditions) $\sqrt{n}\,\hat{\mathbf{D}}_{r_1 r_2 \ldots r_N}$ converges to a multivariate normal distribution with mean 0 and a covariance matrix that is given by the long-run covariance, $\boldsymbol{\Omega}_{r_1 r_2 \ldots r_N}$, of the vector

$$\mathbf{d}_t = \begin{bmatrix} \left(\tilde{U}_t^W\right)^{r_1} - m_{r_1} \\ \left(\tilde{U}_t^W\right)^{r_2} - m_{r_2} \\ \vdots \\ \left(\tilde{U}_t^W\right)^{r_N} - m_{r_N} \end{bmatrix}.$$

The test proposed in Knüppel (2015) is based on the following statistic which follows a $\chi^2$ distribution under $H_0$:

$$\alpha_{r_1 r_2 \ldots r_N} = n \hat{\mathbf{D}}'_{r_1 r_2 \ldots r_N} \hat{\boldsymbol{\Omega}}^{-1}_{r_1 r_2 \ldots r_N} \hat{\mathbf{D}}_{r_1 r_2 \ldots r_N} \sim \chi^2_N, \tag{2.18}$$

where $\hat{\boldsymbol{\Omega}}_{r_1 r_2 \ldots r_N}$ is a consistent estimator of $\boldsymbol{\Omega}_{r_1 r_2 \ldots r_N}$. Autocorrelation in the PITs can be taken into account straightforwardly by using a suitable HAC estimator for the long-run covariance $\boldsymbol{\Omega}_{r_1 r_2 \ldots r_N}$ such as the one proposed by Newey and West (1987). The truncation lag can be either chosen based on a rule of thumb or be determined based on the specific context, e. g., in the case of h-step-ahead forecasts.

## 3 Monte Carlo Studies

We use Monte Carlo simulations to analyze i) how severe the size and power distortions caused by "cheating" with the order-dependent approaches can be, ii) the size and power of the tests based on the transformations discussed in the previous section, and iii) how well the randomization approach works in the presence of estimated parameters. We assume that the data generating process (DGP) under the null hypothesis is a multivariate normal distribution given by

$$y_t \sim \mathcal{N}(0, \Sigma), \tag{3.1}$$

with the $d \times d$ covariance matrix $\Sigma$ being such that all $d$ elements of $y_t$ have unit variances ($\sigma_i^2 = 1$ for $i = 1, \ldots, d$) and the correlation between any two elements of $y_t$ is equal to 0.5 ($\rho_{ij} = 0.5$ for all $i \neq j$). We consider dimensions of $d = \{2, 3, \ldots, 6\}$ and sample sizes of $n = \{50, 100, 200\}$. Throughout the paper, we use 10,000 iterations for our Monte Carlo simulations.

We consider five different alternative DGPs which imply different (combinations of) deviations from $H_0$:

- **Alternative 1 ($H_1$):** The data are generated from a multivariate normal distribution with $\sigma_i^2 = 1.1$ and $\rho_{ij} = 0.5$.
- **Alternative 2 ($H_2$):** The data are generated from a multivariate normal distribution with $\sigma_i^2 = 1.0$ and $\rho_{ij} = 0.4$.
- **Alternative 3 ($H_3$):** The data are generated from a multivariate normal distribution with $\sigma_i^2 = 1.1$ and $\rho_{ij} = 0.4$.
- **Alternative 4 ($H_4$):** The data are generated from a multivariate $t$ distribution with 8 degrees of freedom with $\sigma_i^2 = 1.0$ and $\rho_{ij} = 0.5$.
- **Alternative 5 ($H_5$):** The data are generated from a multivariate $t$ distribution with 8 degrees of freedom with $\sigma_i^2 = 1.1$ and $\rho_{ij} = 0.4$.

To test whether the PITs of the various transformed variables, $U_t^W$, are uniformly distributed we use Neyman's smooth (NS) test (Neyman, 1937) for our baseline results and present robustness checks based on the Kolmogorov-Smirnov (KS) test as well as the test proposed by Knüppel (2015) which we abbreviate by "K".

### 3.1 Potential for "Cheating"

In this section, we present results that address the issue of whether considering different permutations of the data can have a serious impact on the outcomes of the tests that are not order invariant.[10] The question that we ask is: what rejection rates do we obtain if we always choose the permutation for which we obtain either the highest or the lowest test statistic? The idea behind this exercise is the following: a researcher who wants to discredit (support) the hypothesis that a particular model produces good density forecasts could, in principle, search all permutations and select the one which yields the highest (lowest) test statistic; hence the term "cheating". We present results for $H_0$ and $H_5$ based on $n = 100$; results are similar for other alternatives and available upon request. We only consider Neyman's smooth test here, but results for other tests are very similar.

Figure 1 shows how severe the issue of "cheating" is under the null hypothesis. The solid line indicates the nominal size of 5 % which, as we show below, is also obtained in practice for all tests considered if the latter are applied properly (meaning that the ordering of variables is chosen randomly). The other lines refer to the rejection frequencies that we obtain for the

[10] Note that tests based on $Z^2$ are order invariant under the Gaussian setup that we use here.

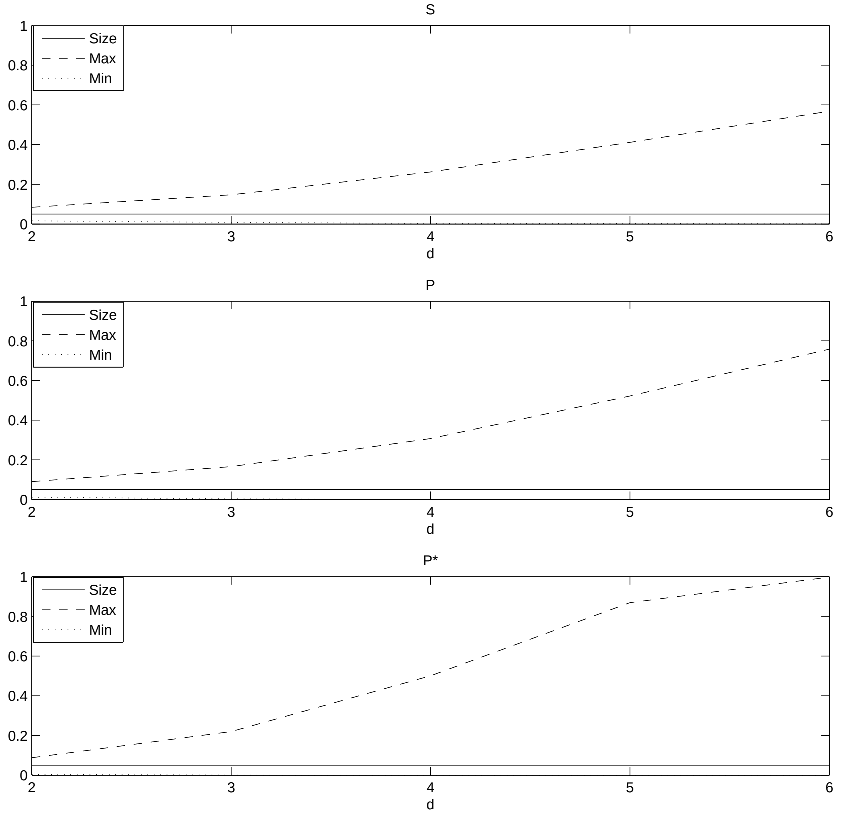

Figure 1: Potential for cheating under $H_0$.

tests based on $S$, $P$, and $P^*$, respectively, when we always choose that permutation for which we obtain the highest (lowest) test statistic. At the lower end of obtainable rejection rates, it is clearly possible to virtually never reject the null hypothesis for any dimension. On the other hand, the (true) null hypothesis can be rejected much too frequently if one chooses those permutations yielding high test statistics. For $d = 2$ the "room for cheating" is rather limited, with obtainable rejection rates being around 10 %. Once the dimension (and consequently the number of possible permutations) increases, obtainable rejection rates increase quickly. They lie above 50% for $d = 6$ for all transformations considered and reach virtually 100 % for the test based on $P^*$.

Now we turn to the effects of "cheating" on the rejection rates under the alternative. Figure 2 shows three lines for each of the tests considered. The solid lines indicate the power that is obtained for different $d$ when the tests are applied properly. The upper (lower) lines show the rejection rates that one obtains when always selecting the highest (lowest) test statistic across all possible permutations of the $d$ variables. The range of obtainable rejection rates is considerable in all cases. The potential for "cheating" is lowest for tests based on $S$. In this case, using the smallest test statistic leads to a rejection frequency of about 60%, in contrast to a little over 80% for the properly used test. For the other approaches the interval increases with $d$ and ranges from virtually 0 to almost 1 for $d = 6$. This means that even though the data are generated

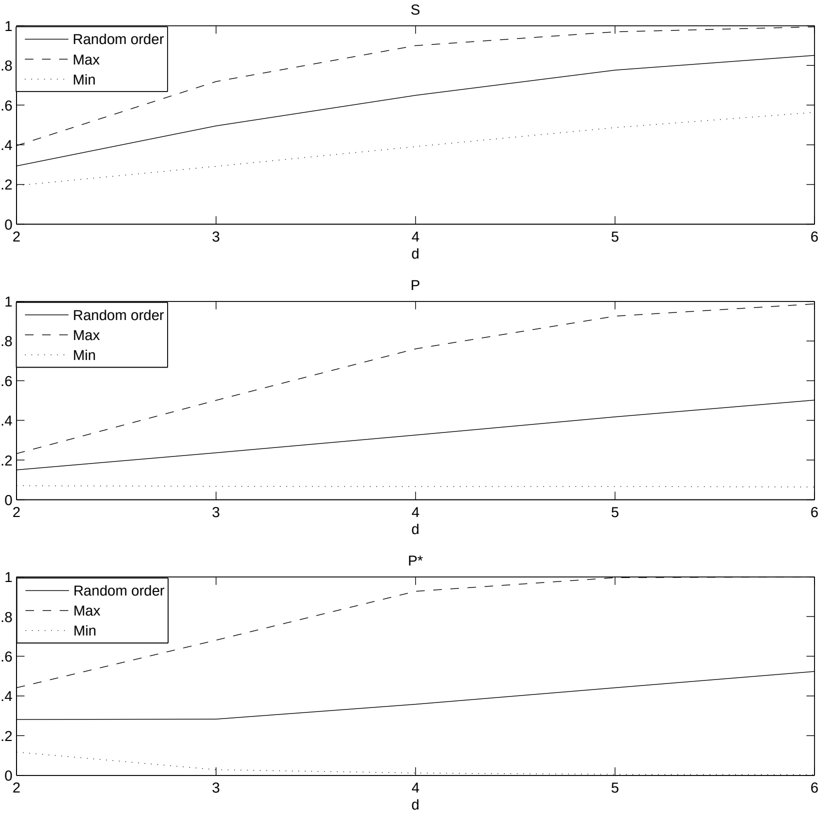

Figure 2: Potential for cheating under $H_5$.

from a different DGP, a researcher would be able to purposely select permutations in such a way that $H_0$ is almost never rejected.

### 3.2 Size and Power

We start by discussing the results under the assumption of known parameters. The Monte Carlo results concerning the size and power for the different transformations can be found in Table 1. These results are based on Neyman's smooth test. Corresponding results based on alternative tests are shown in Appendix C and generally support our conclusions. Focusing on the upper panel of the table, we see that none of the approaches suffers from substantial size-distortions. In all cases, the obtained actual sizes are very close to the nominal size of 5 %.

In terms of power, the second panel of the table reveals that tests based on our three new transformations and on $S$ perform best when deviations of the variances ($H_1$) have to be detected. Their powers are very close to each other for all considered sample sizes, with tests based on $Z^2$ and $Z^{2^*}$ performing marginally better than that based on $S$ which, in turn, has slightly higher power than the test based on $Z^{2^\dagger}$. The third panel, referring to $H_2$, shows that the three new approaches consistently outperform the tests based on the previously suggested transformations, showing that they are better suited for detecting deviations from $H_0$ in terms of the correlation structure of the multivariate density. The improvements of our new tests over tests based on

$P$ and $P^*$ are substantial, and while the test based on $Z^2$ outperforms that based on $S$ only marginally, the other two new approaches have substantially higher power against $H_2$ than the latter. Combining misspecification of variances and correlations in $H_3$ leads to the results shown in the fourth panel of the table. Here, the new approaches consistently outperform the tests based on $S$, $P$, or $P^*$. Relative to the test based on $S$ the outperformance is substantial only for small samples ($n = 50$), as $S$ and the new tests quickly approach a power of 1 for moderate to large sample sizes.

Turning to the power properties of the different tests in terms of detecting misspecification of the kurtosis, the results relating to $H_4$ show that the new approaches outperform all existing tests by a wide margin. Especially for the small sample size the results are stunning: in general, the power of the new approaches exceeds that of even the best-performing old approach threefold. Adding wrongly calibrated variances and correlations to the misspecification of the distribution in $H_5$ leads to a decrease of this outperformance. This is because there is little room for the power of the new approaches to improve while, at the same time, tests based on the old transformations gain a lot of power through these additional deviations from $H_0$. However, our new approaches still clearly outperform the existing approaches.

Now, we turn to the case in which the model parameters have to be estimated from the available data sample. We analyze how strongly the performance in terms of size and power properties declines relative to the case of known parameter values. Table 2 shows the results based on Neyman's smooth tests and, again, robustness checks using alternative tests for uniformity can be found in Appendix C. In general, the results indicate that tests based on all transformations are substantially undersized if one does not take into account that parameters are estimated from the available sample. Using the transformations based on the randomization approach described in Section 2.3, in general, yields correctly sized tests. The tests based on $P$ and $P^*$ are an exception; the former are still undersized for all considered settings while the latter are still undersized for the case of bivariate densities. With regard to the performance under $H_4$ we conclude that having to deal with estimated parameters results in a considerable loss of power; at the same time, the ranking of the competing tests remains unaffected.

## 4 Applications

In this section we provide two applications of our tests. Both applications use models for more than two variables and both show that using tests that are not order invariant potentially gives the researcher the opportunity to manipulate the reported results in many situations. In Section 4.1, we consider the problem of forecasting the distribution of weekly stock index returns for five countries. The models are a DCC-GARCH and more flexible model specifications based on a time-varying t-copula with fat-tailed and potentially asymmetric GARCH models as the margins. In Section 4.2, we analyze the ability of the Bayesian vector autoregressive model by Primiceri (2005) to forecast the multivariate density of macroeconomic variables for the US.

Table 1: Size and power - known parameters (Neyman's smooth test)

| Size | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.047 | 0.051 | 0.047 | 0.051 | 0.050 | 0.052 | 0.051 | 0.050 | 0.052 | 0.045 | 0.054 | 0.055 | 0.053 | 0.051 | 0.050 | 0.051 | 0.052 | 0.053 |
| $d = 3$ | 0.049 | 0.047 | 0.048 | 0.047 | 0.053 | 0.052 | 0.050 | 0.047 | 0.050 | 0.047 | 0.049 | 0.047 | 0.052 | 0.047 | 0.047 | 0.053 | 0.052 | 0.055 |
| $d = 4$ | 0.049 | 0.050 | 0.048 | 0.050 | 0.048 | 0.051 | 0.053 | 0.052 | 0.050 | 0.045 | 0.049 | 0.047 | 0.051 | 0.050 | 0.051 | 0.045 | 0.051 | 0.049 |
| $d = 5$ | 0.051 | 0.049 | 0.051 | 0.049 | 0.052 | 0.054 | 0.051 | 0.047 | 0.048 | 0.049 | 0.048 | 0.047 | 0.050 | 0.050 | 0.049 | 0.049 | 0.053 | 0.052 |
| $d = 6$ | 0.047 | 0.051 | 0.049 | 0.048 | 0.047 | 0.048 | 0.049 | 0.054 | 0.049 | 0.047 | 0.049 | 0.048 | 0.052 | 0.049 | 0.049 | 0.052 | 0.053 | 0.051 |
| Power against $H_1$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.197 | 0.137 | 0.139 | 0.198 | 0.199 | 0.164 | 0.328 | 0.210 | 0.211 | 0.336 | 0.336 | 0.273 | 0.556 | 0.338 | 0.358 | 0.596 | 0.583 | 0.484 |
| $d = 3$ | 0.253 | 0.150 | 0.162 | 0.273 | 0.272 | 0.227 | 0.435 | 0.223 | 0.240 | 0.464 | 0.472 | 0.394 | 0.739 | 0.385 | 0.411 | 0.777 | 0.766 | 0.673 |
| $d = 4$ | 0.323 | 0.164 | 0.184 | 0.338 | 0.335 | 0.289 | 0.557 | 0.247 | 0.280 | 0.603 | 0.597 | 0.516 | 0.859 | 0.435 | 0.487 | 0.893 | 0.882 | 0.822 |
| $d = 5$ | 0.385 | 0.175 | 0.205 | 0.418 | 0.407 | 0.360 | 0.654 | 0.279 | 0.315 | 0.698 | 0.693 | 0.635 | 0.925 | 0.472 | 0.539 | 0.948 | 0.940 | 0.909 |
| $d = 6$ | 0.449 | 0.193 | 0.219 | 0.482 | 0.475 | 0.434 | 0.741 | 0.301 | 0.361 | 0.783 | 0.763 | 0.721 | 0.961 | 0.527 | 0.600 | 0.977 | 0.971 | 0.954 |
| Power against $H_2$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.066 | 0.046 | 0.100 | 0.067 | 0.072 | 0.106 | 0.077 | 0.052 | 0.144 | 0.081 | 0.080 | 0.145 | 0.100 | 0.063 | 0.244 | 0.106 | 0.105 | 0.235 |
| $d = 3$ | 0.090 | 0.052 | 0.083 | 0.098 | 0.111 | 0.168 | 0.136 | 0.063 | 0.103 | 0.146 | 0.166 | 0.274 | 0.219 | 0.086 | 0.136 | 0.247 | 0.283 | 0.492 |
| $d = 4$ | 0.135 | 0.060 | 0.106 | 0.149 | 0.175 | 0.238 | 0.217 | 0.076 | 0.139 | 0.241 | 0.290 | 0.406 | 0.377 | 0.114 | 0.199 | 0.429 | 0.513 | 0.691 |
| $d = 5$ | 0.174 | 0.065 | 0.121 | 0.195 | 0.247 | 0.308 | 0.306 | 0.095 | 0.154 | 0.350 | 0.436 | 0.538 | 0.546 | 0.145 | 0.259 | 0.612 | 0.730 | 0.836 |
| $d = 6$ | 0.225 | 0.075 | 0.138 | 0.252 | 0.324 | 0.373 | 0.409 | 0.115 | 0.200 | 0.462 | 0.570 | 0.643 | 0.706 | 0.187 | 0.327 | 0.762 | 0.856 | 0.915 |
| Power against $H_3$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.305 | 0.135 | 0.256 | 0.326 | 0.325 | 0.378 | 0.510 | 0.203 | 0.423 | 0.549 | 0.549 | 0.627 | 0.810 | 0.341 | 0.696 | 0.844 | 0.848 | 0.904 |
| $d = 3$ | 0.504 | 0.180 | 0.292 | 0.550 | 0.559 | 0.619 | 0.788 | 0.290 | 0.473 | 0.834 | 0.849 | 0.894 | 0.980 | 0.493 | 0.757 | 0.987 | 0.991 | 0.996 |
| $d = 4$ | 0.664 | 0.227 | 0.350 | 0.724 | 0.762 | 0.791 | 0.929 | 0.387 | 0.587 | 0.953 | 0.966 | 0.976 | 0.999 | 0.660 | 0.881 | 0.999 | 1.000 | 1.000 |
| $d = 5$ | 0.801 | 0.290 | 0.436 | 0.850 | 0.875 | 0.890 | 0.980 | 0.489 | 0.690 | 0.988 | 0.993 | 0.995 | 1.000 | 0.792 | 0.940 | 1.000 | 1.000 | 1.000 |
| $d = 6$ | 0.892 | 0.346 | 0.508 | 0.923 | 0.941 | 0.946 | 0.997 | 0.589 | 0.787 | 0.998 | 0.999 | 0.999 | 1.000 | 0.881 | 0.976 | 1.000 | 1.000 | 1.000 |
| Power against $H_4$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.107 | 0.077 | 0.080 | 0.183 | 0.188 | 0.156 | 0.160 | 0.105 | 0.123 | 0.302 | 0.302 | 0.241 | 0.299 | 0.172 | 0.218 | 0.544 | 0.545 | 0.439 |
| $d = 3$ | 0.142 | 0.085 | 0.095 | 0.322 | 0.314 | 0.257 | 0.241 | 0.122 | 0.171 | 0.545 | 0.538 | 0.431 | 0.437 | 0.207 | 0.313 | 0.843 | 0.837 | 0.729 |
| $d = 4$ | 0.177 | 0.091 | 0.125 | 0.481 | 0.472 | 0.391 | 0.311 | 0.143 | 0.214 | 0.763 | 0.750 | 0.652 | 0.563 | 0.247 | 0.413 | 0.970 | 0.970 | 0.925 |
| $d = 5$ | 0.231 | 0.109 | 0.149 | 0.620 | 0.622 | 0.551 | 0.399 | 0.165 | 0.269 | 0.889 | 0.883 | 0.822 | 0.677 | 0.291 | 0.519 | 0.996 | 0.995 | 0.987 |
| $d = 6$ | 0.264 | 0.114 | 0.173 | 0.747 | 0.736 | 0.670 | 0.456 | 0.186 | 0.327 | 0.961 | 0.955 | 0.924 | 0.752 | 0.344 | 0.619 | 1.000 | 1.000 | 0.998 |
| Power against $H_5$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.186 | 0.108 | 0.181 | 0.299 | 0.300 | 0.320 | 0.293 | 0.155 | 0.275 | 0.486 | 0.478 | 0.514 | 0.504 | 0.248 | 0.478 | 0.756 | 0.762 | 0.795 |
| $d = 3$ | 0.298 | 0.155 | 0.185 | 0.507 | 0.525 | 0.524 | 0.493 | 0.233 | 0.276 | 0.762 | 0.773 | 0.778 | 0.772 | 0.402 | 0.465 | 0.958 | 0.963 | 0.968 |
| $d = 4$ | 0.406 | 0.199 | 0.232 | 0.666 | 0.689 | 0.688 | 0.656 | 0.323 | 0.364 | 0.904 | 0.913 | 0.911 | 0.905 | 0.563 | 0.597 | 0.994 | 0.997 | 0.996 |
| $d = 5$ | 0.521 | 0.260 | 0.283 | 0.793 | 0.802 | 0.791 | 0.781 | 0.430 | 0.446 | 0.965 | 0.971 | 0.968 | 0.965 | 0.691 | 0.711 | 1.000 | 1.000 | 1.000 |
| $d = 6$ | 0.588 | 0.304 | 0.326 | 0.859 | 0.872 | 0.865 | 0.847 | 0.505 | 0.514 | 0.988 | 0.991 | 0.989 | 0.987 | 0.784 | 0.799 | 1.000 | 1.000 | 1.000 |

**Notes:** Rejection frequencies of Neyman's smooth test based on the transformations introduced in Section 2.2 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. The alternative hypotheses are defined in Section 3. All Monte Carlo simulations are based on 10,000 iterations.

## 4.1 Predicting the Distribution of Stock Market Returns

As a first application, we consider the problem of forecasting the joint distribution of five international stock market indices. Our data consist of weekly returns of the MSCI indices for the US, Japan, UK, Australia, and Germany and were obtained from Datastream. The sample spans the period from January 1971 until October 2013 for a total of 2,232 weekly returns. We consider eight different time periods of four years for which we evaluate density forecasts. These (out-of-sample) evaluation periods are (1) 1981-1984, (2) 1985-1988, (3) 1989-1992, (4) 1993-1996, (5) 1997-2000, (6) 2001-2004, (7) 2005-2008, and (8) 2009-2013. For each period, the previous ten years are considered as in-sample data to estimate the models of interest. The models are re-estimated for each week using a recursive scheme.

Three competing models of increasing complexity are considered, namely (i) a Gaussian DCC-GARCH model (Engle, 2002), (ii) a time-varying Student t-copula with t-GARCH mar-

Table 2: Size and power - estimated parameters (Neyman's smooth test)

| Size (original test) | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.030 | 0.027 | 0.012 | 0.022 | 0.023 | 0.021 | 0.026 | 0.026 | 0.014 | 0.023 | 0.024 | 0.020 | 0.026 | 0.022 | 0.012 | 0.020 | 0.021 | 0.022 |
| $d = 3$ | 0.028 | 0.029 | 0.033 | 0.022 | 0.023 | 0.021 | 0.030 | 0.025 | 0.032 | 0.022 | 0.023 | 0.023 | 0.030 | 0.024 | 0.033 | 0.024 | 0.024 | 0.023 |
| $d = 4$ | 0.030 | 0.030 | 0.038 | 0.024 | 0.024 | 0.020 | 0.026 | 0.023 | 0.039 | 0.022 | 0.022 | 0.022 | 0.025 | 0.028 | 0.035 | 0.023 | 0.022 | 0.022 |
| $d = 5$ | 0.034 | 0.031 | 0.038 | 0.023 | 0.020 | 0.021 | 0.029 | 0.024 | 0.035 | 0.021 | 0.021 | 0.023 | 0.026 | 0.026 | 0.036 | 0.024 | 0.024 | 0.025 |
| $d = 6$ | 0.033 | 0.032 | 0.035 | 0.027 | 0.026 | 0.024 | 0.027 | 0.027 | 0.038 | 0.027 | 0.025 | 0.026 | 0.029 | 0.027 | 0.035 | 0.021 | 0.021 | 0.023 |
| Size (adjusted test) | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.054 | 0.041 | 0.026 | 0.053 | 0.051 | 0.054 | 0.054 | 0.036 | 0.024 | 0.052 | 0.053 | 0.049 | 0.054 | 0.037 | 0.025 | 0.047 | 0.050 | 0.047 |
| $d = 3$ | 0.057 | 0.036 | 0.047 | 0.055 | 0.051 | 0.055 | 0.053 | 0.035 | 0.052 | 0.049 | 0.052 | 0.051 | 0.050 | 0.034 | 0.050 | 0.051 | 0.053 | 0.047 |
| $d = 4$ | 0.061 | 0.039 | 0.049 | 0.055 | 0.053 | 0.050 | 0.055 | 0.037 | 0.051 | 0.048 | 0.049 | 0.053 | 0.052 | 0.032 | 0.049 | 0.051 | 0.052 | 0.047 |
| $d = 5$ | 0.064 | 0.038 | 0.049 | 0.057 | 0.051 | 0.050 | 0.058 | 0.032 | 0.054 | 0.048 | 0.048 | 0.051 | 0.054 | 0.032 | 0.050 | 0.051 | 0.048 | 0.053 |
| $d = 6$ | 0.063 | 0.040 | 0.050 | 0.052 | 0.049 | 0.049 | 0.057 | 0.039 | 0.048 | 0.054 | 0.052 | 0.051 | 0.050 | 0.029 | 0.052 | 0.052 | 0.049 | 0.049 |
| Power against $H_4$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.100 | 0.062 | 0.058 | 0.165 | 0.158 | 0.138 | 0.153 | 0.084 | 0.091 | 0.270 | 0.270 | 0.221 | 0.278 | 0.153 | 0.180 | 0.499 | 0.495 | 0.401 |
| $d = 3$ | 0.112 | 0.059 | 0.085 | 0.238 | 0.234 | 0.186 | 0.210 | 0.095 | 0.147 | 0.467 | 0.465 | 0.362 | 0.389 | 0.178 | 0.289 | 0.775 | 0.771 | 0.648 |
| $d = 4$ | 0.126 | 0.062 | 0.098 | 0.327 | 0.308 | 0.224 | 0.246 | 0.101 | 0.184 | 0.638 | 0.625 | 0.500 | 0.496 | 0.209 | 0.374 | 0.922 | 0.918 | 0.839 |
| $d = 5$ | 0.154 | 0.069 | 0.100 | 0.409 | 0.390 | 0.265 | 0.280 | 0.109 | 0.208 | 0.771 | 0.755 | 0.615 | 0.578 | 0.240 | 0.451 | 0.979 | 0.977 | 0.936 |
| $d = 6$ | 0.155 | 0.077 | 0.098 | 0.461 | 0.439 | 0.283 | 0.315 | 0.119 | 0.236 | 0.854 | 0.841 | 0.712 | 0.646 | 0.270 | 0.532 | 0.996 | 0.995 | 0.980 |

**Notes:** Rejection frequencies of Neyman's smooth test based on the transformations introduced in Section 2.2 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. The alternative hypotheses are defined in Section 3. All Monte Carlo simulations are based on 10,000 iterations.

gins,[11] and (iii) a time-varying t-copula with skewed-t-GJR-GARCH margins. Formally, for the DCC-GARCH model the marginal models for $i = 1, \ldots, d$ are given by

$$\begin{aligned} Y_{i,t} &= \mu_i + \varepsilon_{i,t} \\ \varepsilon_{i,t} &= \sqrt{h_{i,t}} z_{i,t} \\ h_{i,t} &= \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1} \end{aligned}$$

with $z_{i,t} \sim \mathcal{N}(0,1)$, $\omega_i, \alpha_i, \beta_i \geq 0$ and $\alpha_i + \beta_i < 1$. The correlation matrix $R_t$ of the innovations $z_t = [z_{1,t}, \ldots, z_{d,t}]$ is given by

$$R_t = \operatorname{diag}(Q_t)^{-1/2} Q_t \operatorname{diag}(Q_t)^{-1/2}, \tag{4.1}$$

where

$$Q_t = (1 - \alpha_c - \beta_c)\bar{Q} + \alpha_c z_{t-1}' z_{t-1} + \beta_c Q_{t-1}, \tag{4.2}$$

with $\alpha_c, \beta_c \geq 0$, $\alpha_c + \beta_c \leq 1$, and $\bar{Q} = E(z_t' z_t)$, which in practice is estimated with the sample covariance matrix of $z_t$.

For the second model, the marginal models are the same as above, with the difference that the innovations $z_{i,t}$ follow a t-distribution with $\nu_i$ degrees of freedom. The dependence between the t-distributed GARCH innovations $z_t$ is given by a t-copula with degrees of freedom $\nu_c$ and correlation matrix $R_t$. Let $U_{i,t} = T_{\nu_i}\left(\sqrt{\frac{\nu_i}{\nu_i - 2}} \frac{Y_{it} - \mu_i}{\sqrt{h_{it}}}\right)$, where $T_{\nu_i}$ denotes the CDF of a univariate

[11] The time-varying correlation matrix of the copula is driven by DCC-type dynamics as described in the text, see also Manner and Reznikova (2012).

t-distribution with $\nu_i$ degrees of freedom. Then the t-copula is given by

$$C(U_{1,t}, \ldots U_{d,t}; R_t, \nu_c) = T^d_{\nu_c}(T^{-1}_{\nu_c}(U_{1,t}), \ldots, T^{-1}_{\nu_c}(U_{d,t}); R_t),$$

where $T^d_{\nu_c}$ stands for the CDF of the $d$-dimensional t-distribution with $\nu_c$ degrees of freedom. For details and properties of the t-copula see, e. g., Joe (2014). The evolution of the correlation matrix is given by (4.1) and (4.2), but with $z_{i,t}$ replaced by $T^{-1}_{\nu_c}(U_{i,t})\sqrt{\frac{\nu_c-2}{\nu_c}}$. Note that this model is slightly more flexible than a DCC-GARCH model based on a multivariate t-distribution since the copula approach allows all marginal series to have distinct degrees-of-freedom which are also different from the degrees of freedom of the copula. The estimation of the copula-based model is naturally done in two steps, ensuring numerical stability at the price of a small loss in statistical efficiency; see Joe (2005) on two-step estimation of copula models.

The third model is made even more flexible by assuming that the GARCH innovations $z_{i,t}$ follow the skewed-t distribution of Hansen (1994) and by relying on the GJR-GARCH model of Glosten et al. (1993), for which the conditional variance follows

$$h_{i,t} = \omega_i + \alpha_i \varepsilon^2_{i,t-1} + \beta_i h_{i,t-1} + \gamma_i \varepsilon^2_{i,t-1} I(\varepsilon_{i,t-1} < 0).$$

The dependence is again given by the DCC-t-copula model.

For each model and each time period, we compute the Rosenblatt PITs and further transform the data with the methods studied in Section 2.2. Recall that for non-Gaussian models the distribution of $Z^{2*}_t$ and $Z^{2\dagger}_t$ is not known. Therefore, we compute the PITs by Monte Carlo simulation as explained in Section 2.2. The null hypothesis of a correctly predicted density is then tested with Neyman's smooth test (Neyman, 1937) and the test by Knüppel (2015), the latter test being robust to autocorrelation in the transformed series.[12] In Tables 3 and 4, we report the p-values based on the different transformations introduced in Section 2.2. For those tests which are not order invariant we consider all $5! = 120$ permutations of the data. We report the p-value of a random permutation of the variables (based on the arbitrary order in which we downloaded the data: US, JP, UK, AU, GE) and, in brackets, the smallest and largest p-values across all permutations.

Overall, the results are mixed and depend on the time period under study. However, a few things clearly stand out. First of all, the Gaussian DCC model is rejected by all tests for all time periods except the 1997-2000 period. Second, model specifications (ii) and (iii) perform much better, but are still rejected for some periods. Notably, most tests reject these models for the aforementioned 1997-2000 period. This suggests that during that period returns had much lighter tails than in previous years. A shorter in-sample period may be appropriate to reflect such non-stationarities. Second, the more flexible specification (iii) does not yield more appropriate forecasts for all periods, confirming the known fact that model complexity may yield superior in-sample fit, but not necessarily a better forecasting performance. Third, the potential for "cheating" using the tests based on $S$, $P$, and $P^*$, by Diebold et al. (1999), Clements and

[12] Here, we use an automatic selection of the truncation lag following Andrews (1991).

Table 3: Density forecast evaluation for stock market returns (Neyman's smooth test)

| **Gaussian DCC** | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
|---|---|---|---|---|---|---|
| 1981-1984 | 0.022 [0.007, 0.189] | 0.001 [0.000, 0.019] | 0.013 [0.005, 0.948] | 0.013 | 0.012 | 0.005 |
| 1985-1988 | 0.000 [0.000, 0.000] | 0.000 [0.000, 0.001] | 0.647 [0.004, 0.914] | 0.000 | 0.001 | 0.001 |
| 1989-1992 | 0.000 [0.000, 0.001] | 0.142 [0.001, 0.457] | 0.016 [0.000, 0.075] | 0.000 | 0.000 | 0.000 |
| 1993-1996 | 0.000 [0.000, 0.000] | 0.004 [0.000, 0.021] | 0.019 [0.000, 0.019] | 0.000 | 0.000 | 0.000 |
| 1997-2000 | 0.169 [0.009, 0.642] | 0.015 [0.000, 0.382] | 0.727 [0.059, 0.981] | 0.010 | 0.066 | 0.089 |
| 2001-2004 | 0.000 [0.000, 0.000] | 0.002 [0.001, 0.120] | 0.014 [0.000, 0.051] | 0.000 | 0.000 | 0.000 |
| 2005-2008 | 0.000 [0.000, 0.167] | 0.000 [0.000, 0.004] | 0.001 [0.000, 0.609] | 0.000 | 0.000 | 0.000 |
| 2009-2013 | 0.000 [0.000, 0.000] | 0.009 [0.000, 0.066] | 0.015 [0.000, 0.506] | 0.000 | 0.000 | 0.000 |
| **t-GARCH-tDCC-Cop** | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| 1981-1984 | 0.286 [0.054, 0.645] | 0.001 [0.000, 0.032] | 0.347 [0.227, 0.997] | 0.538 [0.330, 0.538] | 0.375 | 0.199 |
| 1985-1988 | 0.000 [0.000, 0.000] | 0.001 [0.000, 0.012] | 0.538 [0.057, 0.991] | 0.113 [0.029, 0.206] | 0.057 | 0.350 |
| 1989-1992 | 0.409 [0.016, 0.425] | 0.787 [0.045, 0.911] | 0.970 [0.009, 0.970] | 0.162 [0.102, 0.290] | 0.146 | 0.125 |
| 1993-1996 | 0.007 [0.001, 0.052] | 0.044 [0.001, 0.114] | 0.685 [0.010, 0.750] | 0.013 [0.011, 0.032] | 0.061 | 0.075 |
| 1997-2000 | 0.056 [0.005, 0.089] | 0.070 [0.001, 0.720] | 0.626 [0.002, 0.952] | 0.011 [0.009, 0.017] | 0.006 | 0.094 |
| 2001-2004 | 0.000 [0.000, 0.001] | 0.024 [0.001, 0.377] | 0.145 [0.001, 0.335] | 0.000 [0.000, 0.000] | 0.000 | 0.000 |
| 2005-2008 | 0.000 [0.000, 0.446] | 0.000 [0.000, 0.032] | 0.015 [0.002, 0.950] | 0.055 [0.047, 0.166] | 0.499 | 0.079 |
| 2009-2013 | 0.001 [0.000, 0.007] | 0.151 [0.001, 0.431] | 0.013 [0.002, 0.972] | 0.040 [0.024, 0.051] | 0.028 | 0.023 |
| **st-GJR-tDCC-Cop** | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| 1981-1984 | 0.307 [0.040, 0.587] | 0.001 [0.000, 0.022] | 0.192 [0.123, 0.998] | 0.756 [0.614, 0.774] | 0.694 | 0.200 |
| 1985-1988 | 0.000 [0.000, 0.000] | 0.002 [0.000, 0.045] | 0.640 [0.048, 0.995] | 0.005 [0.002, 0.011] | 0.004 | 0.005 |
| 1989-1992 | 0.049 [0.000, 0.050] | 0.617 [0.033, 0.924] | 0.603 [0.002, 0.967] | 0.241 [0.157, 0.357] | 0.202 | 0.221 |
| 1993-1996 | 0.004 [0.000, 0.028] | 0.042 [0.002, 0.076] | 0.586 [0.007, 0.656] | 0.014 [0.011, 0.038] | 0.050 | 0.047 |
| 1997-2000 | 0.033 [0.001, 0.208] | 0.218 [0.001, 0.926] | 0.223 [0.002, 0.979] | 0.008 [0.007, 0.014] | 0.022 | 0.159 |
| 2001-2004 | 0.000 [0.000, 0.000] | 0.050 [0.002, 0.348] | 0.027 [0.000, 0.042] | 0.000 [0.000, 0.000] | 0.000 | 0.000 |
| 2005-2008 | 0.000 [0.000, 0.210] | 0.006 [0.000, 0.463] | 0.003 [0.000, 0.714] | 0.010 [0.009, 0.024] | 0.020 | 0.008 |
| 2009-2013 | 0.277 [0.007, 0.729] | 0.704 [0.076, 0.790] | 0.108 [0.008, 0.967] | 0.728 [0.684, 0.754] | 0.328 | 0.202 |

**Notes:** The table shows p-values corresponding to the different transformations introduced in Section 2.2 and using Neyman's smooth test (Neyman, 1937). The data are weekly MSCI stock index returns for the US, Japan, UK, Australia and Germany. Forecasts are evaluated for the stated periods and the previous 10 years of data are used as the in-sample period. For transformations which are not order invariant, the numbers in brackets show the lowest and highest obtained p-values across all permutations of the variables; for these transformations, the first p-value is for an arbitrarily selected permutation. The models, introduced in Section 4.1, are a Gaussian DCC-GARCH model, t-GARCH margins and a t-copula with DCC correlation matrix, and a GJR-GARCH model with skewed-t innovations and a t-copula with DCC correlation matrix.

Smith (2000), and Ko and Park (2013), respectively, is immense. For the majority of periods one can find a permutation that rejects or does not reject the density forecasts of any of the models we study. Note, however, that in line with our results from Section 3.1, the range of obtainable p-values is a little smaller for tests based on $S$ than for the ones based on $P$ and $P^*$. Turning to the results for $Z^2$ which are not order invariant for the non-Gaussian models, one can see that the range of the p-values is very limited and that there is almost no room for cheating based on this transformation.

In summary, we recommend evaluating the density forecast solely based on $Z^{2*}$ and $Z^{2\dagger}$, and possibly based on $Z^2$. The results based on the other tests are not reliable as different permutations can lead to substantially different conclusions regarding the performance of the models. Furthermore, our Monte Carlo simulations show that the new tests are superior in terms of power. Thus looking only at the last two columns of Tables 3 and 4, we see that at the 1% significance level the copula-based and fat-tailed model specifications (ii) and (iii) are only rejected for one or two time periods. Taking into account that for each test we perform eight different hypothesis tests (for the different periods), we are actually in a situation of multiple hypothesis testing. Using a Bonferroni correction, a test at the 5% significance level should thus reject when the p-value is smaller than $0.05/8 = 0.0063$. Thus, overall, the models are only rejected for the 1997-2000 period, for which the Gaussian DCC model (i) is appropriate.

Table 4: Density forecast evaluation for stock market returns (Knüppel test)

| **Gaussian DCC** | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
|---|---|---|---|---|---|---|
| 1981-1984 | 0.025 [0.008, 0.227] | 0.009 [0.005, 0.062] | 0.022 [0.009, 0.954] | 0.012 | 0.013 | 0.010 |
| 1985-1988 | 0.000 [0.000, 0.000] | 0.001 [0.000, 0.005] | 0.633 [0.003, 0.903] | 0.005 | 0.019 | 0.010 |
| 1989-1992 | 0.001 [0.000, 0.002] | 0.173 [0.002, 0.489] | 0.026 [0.000, 0.067] | 0.000 | 0.000 | 0.000 |
| 1993-1996 | 0.000 [0.000, 0.000] | 0.006 [0.000, 0.031] | 0.019 [0.000, 0.023] | 0.000 | 0.000 | 0.000 |
| 1997-2000 | 0.174 [0.011, 0.713] | 0.015 [0.001, 0.429] | 0.703 [0.048, 0.977] | 0.037 | 0.180 | 0.166 |
| 2001-2004 | 0.000 [0.000, 0.001] | 0.007 [0.005, 0.172] | 0.052 [0.000, 0.105] | 0.000 | 0.000 | 0.001 |
| 2005-2008 | 0.000 [0.000, 0.265] | 0.000 [0.000, 0.024] | 0.024 [0.001, 0.671] | 0.000 | 0.000 | 0.001 |
| 2009-2013 | 0.000 [0.000, 0.000] | 0.009 [0.000, 0.120] | 0.015 [0.000, 0.627] | 0.000 | 0.000 | 0.000 |
| **t-GARCH-tDCC-Cop** | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| 1981-1984 | 0.279 [0.053, 0.641] | 0.009 [0.002, 0.079] | 0.325 [0.127, 0.997] | 0.416 [0.261, 0.423] | 0.181 | 0.066 |
| 1985-1988 | 0.000 [0.000, 0.000] | 0.003 [0.000, 0.034] | 0.522 [0.037, 0.990] | 0.196 [0.057, 0.283] | 0.111 | 0.385 |
| 1989-1992 | 0.369 [0.009, 0.389] | 0.864 [0.078, 0.939] | 0.968 [0.003, 0.968] | 0.128 [0.070, 0.269] | 0.123 | 0.115 |
| 1993-1996 | 0.002 [0.000, 0.028] | 0.055 [0.002, 0.113] | 0.660 [0.001, 0.714] | 0.002 [0.001, 0.006] | 0.016 | 0.038 |
| 1997-2000 | 0.049 [0.003, 0.066] | 0.087 [0.005, 0.747] | 0.588 [0.002, 0.960] | 0.003 [0.002, 0.005] | 0.002 | 0.033 |
| 2001-2004 | 0.000 [0.000, 0.001] | 0.037 [0.006, 0.437] | 0.193 [0.000, 0.427] | 0.010 [0.010, 0.016] | 0.008 | 0.002 |
| 2005-2008 | 0.000 [0.000, 0.503] | 0.001 [0.000, 0.071] | 0.086 [0.001, 0.964] | 0.136 [0.112, 0.294] | 0.660 | 0.189 |
| 2009-2013 | 0.001 [0.000, 0.007] | 0.171 [0.002, 0.466] | 0.005 [0.000, 0.968] | 0.080 [0.063, 0.111] | 0.050 | 0.086 |
| **st-GJR-tDCC-Cop** | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| 1981-1984 | 0.305 [0.043, 0.583] | 0.013 [0.001, 0.063] | 0.190 [0.134, 0.998] | 0.650 [0.489, 0.678] | 0.575 | 0.086 |
| 1985-1988 | 0.000 [0.000, 0.000] | 0.006 [0.000, 0.061] | 0.634 [0.047, 0.993] | 0.047 [0.022, 0.066] | 0.025 | 0.030 |
| 1989-1992 | 0.048 [0.000, 0.048] | 0.660 [0.068, 0.938] | 0.626 [0.001, 0.950] | 0.206 [0.133, 0.318] | 0.176 | 0.187 |
| 1993-1996 | 0.001 [0.000, 0.013] | 0.035 [0.001, 0.052] | 0.476 [0.000, 0.578] | 0.003 [0.002, 0.011] | 0.019 | 0.028 |
| 1997-2000 | 0.034 [0.000, 0.215] | 0.249 [0.009, 0.933] | 0.221 [0.002, 0.975] | 0.009 [0.009, 0.014] | 0.023 | 0.078 |
| 2001-2004 | 0.000 [0.000, 0.000] | 0.070 [0.005, 0.355] | 0.035 [0.000, 0.057] | 0.001 [0.001, 0.001] | 0.001 | 0.000 |
| 2005-2008 | 0.000 [0.000, 0.270] | 0.023 [0.002, 0.531] | 0.004 [0.001, 0.684] | 0.058 [0.055, 0.088] | 0.060 | 0.043 |
| 2009-2013 | 0.286 [0.005, 0.706] | 0.656 [0.090, 0.783] | 0.104 [0.004, 0.968] | 0.744 [0.704, 0.772] | 0.335 | 0.219 |

**Notes:** The table shows p-values corresponding to the different transformations introduced in Section 2.2 and using the test by Knüppel (2015). The data are weekly MSCI stock index returns for the US, Japan, UK, Australia and Germany. Forecasts are evaluated for the stated periods and the previous 10 years of data are used as the in-sample period. For transformations which are not order invariant, the numbers in brackets show the lowest and highest obtained p-values across all permutations of the variables; for these transformations, the first p-value is for an arbitrarily selected permutation. The models, introduced in Section 4.1, are a Gaussian DCC-GARCH model, t-GARCH margins and a t-copula with DCC correlation matrix, and a GJR-GARCH model with skewed-t innovations and a t-copula with DCC correlation matrix.

### 4.2 Evaluating Macroeconomic Density Forecasts

As a second application, we demonstrate how the new tests developed in this paper can be applied in the area of macroeconomic forecasting. To demonstrate that our new tests, in contrast to available approaches, are not prone to cheating and that they can also be used to evaluate higher-order forecasts and when densities are estimated nonparametrically, we evaluate macroeconomic density forecasts for the US economy which we generate using the model by Primiceri (2005).

The model is a Bayesian vector autoregressive (VAR) model with time-varying parameters, which is designed to track changes in macroeconomic volatility and structural changes that alter the economic transmission channels. As in Primiceri (2005), we model the unemployment rate ($u_t$), the log-difference of the chain weighted GDP price index ($\Delta p_t$), and the yield of three-month Treasury bills ($i_t$). The data are downloaded from FRED and cover the sample from 1953$q$1 until 2015$q$2. Collecting all variables in one vector, $y_t = [u_t, \Delta p_t, i_t]'$, the main equation of the model can be written as

$$y_t = c_t + B_{1,t} y_{t-1} + \cdots + B_{k,t} y_{t-k} + u_t \qquad t = 1, \ldots, n. \tag{4.3}$$

Here, $c_t$ is a vector of time-varying intercepts, the $B_{i,t}$ for $i = 1, \ldots, k$, are matrices with time-varying coefficients, and $u_t$ is a vector of unobservable shocks with a time-varying covariance matrix $\Sigma_t$. We follow the exact specification of Primiceri (2005) in allowing for rather flexible processes that govern the variation of the model's parameters over time. In essence, all time-varying parameters (including those of the covariance matrix) are specified as random walk processes and the covariance matrix of the vector of innovations to these processes is assumed to have a block diagonal structure. Details can be found in Section 2 of Primiceri (2005).

The model can be estimated using Bayesian methods. We follow the specification of priors as in Primiceri (2005). We use Gibbs sampling to evaluate numerically the posterior distribution of the parameters and unobserved states of the model. Note that we use the corrected algorithm (Del Negro and Primiceri, 2015) that implies a different ordering of the Markov Chain Monte Carlo steps.[13]

We use a recursive scheme to generate density forecasts, $\hat{f}_{y_{t+h}}(y_{t+h}|\mathcal{F}_t)$, with forecast horizons $h = 1, \ldots, 4$. The period between $1982q4 + h$ and $2014q2 + h$ is used as our evaluation sample. Thus, we start by estimating the model using data until $1982q4$ and constructing density forecasts for $1983q1$, $1983q2$, $1983q3$, and $1983q4$. Subsequently, we recursively add one observation to our estimation sample and shift the forecast period one quarter forward. This yields a sequence of 127 density forecasts for each forecast horizon.

The form of $\hat{f}_{y_{t+h}}(y_{t+h}|\mathcal{F}_t)$ is unknown in general. For $h = 1$ the conditional forecasts follow a multivariate normal distribution conditional on the parameters of the model but not unconditionally. For $h > 1$, a second source for deviations from a Gaussian distribution is given by the fact that the conditional forecasts are non-linear functions of the model parameters. Therefore, we estimate the predictive densities nonparametrically. All results are based on samples of $B = 5,000$ draws from the posterior distribution of the model parameters that we obtain by keeping every $10^{th}$ draw from a sample of 50,000 draws, after a burn-in phase of 5,000 draws. For each of these draws, we simulate corresponding draws from the implied predictive density, $\hat{y}_{t+h}^{(b)}$, which reflect estimation uncertainty and shocks that occur during the forecast period (see Krüger et al., 2016). We use a nonparametric kernel estimator with a (second order) Gaussian kernel (with fixed bandwidths) to estimate the different conditional and marginal distributions that are needed to compute the conditional PITs under all possible permutations.[14] Since the data-driven determination of optimal bandwidths is computationally demanding, we do so only for every twelfth period and keep the bandwidths fixed for all intermediate periods. When we re-optimize the bandwidths, we rely on least-squares cross-validation (Li et al., 2013).[15]

As stated above, the distributions of the random variables $Z_t^{2*}$ and $Z_t^{2\dagger}$ are not known unless the data is assumed to follow a multivariate normal distribution. In our case, however, we only have simulated samples from unknown predictive densities. In order to compute the PITs for $Z_t^{2*}$ and for $Z_t^{2\dagger}$, we simulate their distribution by repeatedly computing $Z_t^{2*}$ and $Z_t^{2\dagger}$ under $H_0$, i. e., under the assumption that the realized values of $y_t$ over the evaluation sample are indeed generated by the model that we use to form our predictive densities; see Section 2.2 for details on the algorithm.

[13] We thank Fabian Krüger for providing his 'bvarsv' package for R which we used to estimate the model.

[14] For this application, $d = 3$ and we need to estimate $\hat{F}_{u_{t+h}|\Delta p_{t+h}, i_{t+h}}(u_{t+h}|\Delta p_{t+h}, i_{t+h}; \mathcal{F}_t)$, in short notation $\hat{F}_{u|\Delta p,i}(u_{t+h})$, as well as $\hat{F}_{\Delta p|u,i}(\Delta p_{t+h})$, …, $\hat{F}_{u|\Delta p}(u_{t+h})$, $\hat{F}_{u|i}(u_{t+h})$, …,$\hat{F}_{i|\Delta p}(i_{t+h})$, $\hat{F}_u(u_{t+h})$, $\hat{F}_{\Delta p}(\Delta p_{t+h})$, and $\hat{F}_i(i_{t+h})$. These are a total of twelve (conditional) distributions that we have to estimate for each forecast period and horizon.

[15] All nonparametric estimations are executed using the 'np' package for R (Hayfield and Racine, 2008).

For comparison, we also check whether results differ if we use an approximation and assume that all conditional forecasts follow a multivariate normal distribution (as, e. g., suggested by Adolfson et al., 2007). In this case, the mean and the covariance matrix completely determine the predictive density. We estimate both quantities as $\bar{y}_{t+h} = (1/B)\sum_{b=1}^{B} \hat{y}_{t+h}^{(b)}$ and $\Sigma_{t+h} = (1/B)\sum_{b=1}^{B}\left(\hat{y}_{t+h}^{(b)} - \bar{y}_{t+h}\right)\left(\hat{y}_{t+h}^{(b)} - \bar{y}_{t+h}\right)'$.

The test proposed by Knüppel (2015) explicitly allows us to account for autocorrelated PITs. Since for $h > 1$ the PITs will be subject to autocorrelation, we report results using this test along with results based on the Neyman smooth (NS) test which is shown to have good properties in our Monte Carlo simulations. Test results for $h = 1$ and $h = 4$ are summarized in Table 5.[16] The upper panel lists results based on the nonparametric approach while the lower panel lists those based on the approximative assumption of normally distributed predictive densities. We show p-values for both tests, for the different transformations, and for all possible permutations of the variables. For those transformations that are order invariant we show only one p-value.

We first focus on the nonparametric predictive densities. By and large, the tests based on our preferred transformations, $Z^{2^*}$ and $Z^{2^\dagger}$, indicate for all forecast horizons that the conditional predictive densities are well calibrated. At a 5 % significant level, only the NS test based on $Z^{2^\dagger}$ marginally rejects $H_0$ for $h = 4$ (p-value of 0.045). The evidence based on those transformations that are not order invariant is mixed. The variation in p-values across permutations is large in almost all cases, indicating that the "cheating" issue can be very relevant in practice even for low-dimensional models when using transformations which are not order invariant.[17] Assuming a significance level of 5 %, decisions made based on the NS test (K test) are dependent on the choice of the permutation in 2 (1) cases for $h = 1$ and in 4 (0) cases for $h = 4$. In general, however, both tests do not reject the null hypothesis for the majority of permutations for $h = 1$. In contrast, the NS test rejects the null of properly calibrated density forecasts in most cases for $h = 4$ while the K test yields large p-values also in these cases.

The results corresponding to the approximative approach provide strong evidence against the null hypothesis of well-calibrated predictive densities. All order-invariant transformations yield p-values very close to 0 for both $h = 1$ and $h = 4$. For $h = 1$, the tests based on the other transformations mostly reject the null hypothesis for the majority of permutations (the product transformation $P$ being an exception). For $h = 4$, the NS test tends to reject $H_0$ while the K test does not lead to any rejections of the null hypothesis. The latter dissent can be explained by two factors: first, our Monte Carlo simulations showed that the NS test has more power, in general, than the K test. Second, accounting for autocorrelation induces a tendency to reject $H_0$ less frequently.

In general, we conclude that (i) "cheating" can be a very relevant issue in practice, (ii) the VAR model with time-varying parameters proposed by Primiceri (2005) seems to generate well-calibrated multivariate density forecasts, and (iii) the latter result holds true for properly estimated predictive densities but not when using a Gaussian approximation.

[16] Results for $h = 2$ and $h = 3$ are very similar and available on request.

[17] The average (across forecast horizons and transformations) standard deviation of p-values (across permutations) is 0.09 for the NS test and 0.12 for the K test.

Table 5: Tests for proper calibration of macroeconomic forecasts

| **Nonparametric densities** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NST | | | | | | KT | | | | | |
| $h = 1$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $u_t - \Delta p_t - i_t$ | 0.374 | 0.667 | 0.022 | 0.006 | 0.230 | 0.107 | 0.374 | 0.675 | 0.023 | 0.071 | 0.382 | 0.292 |
| $u_t - i_t - \Delta p_t$ | 0.552 | 0.216 | 0.769 | 0.158 | | | 0.555 | 0.256 | 0.732 | 0.254 | | |
| $\Delta p_t - u_t - i_t$ | 0.402 | 0.644 | 0.005 | 0.004 | | | 0.403 | 0.652 | 0.007 | 0.066 | | |
| $\Delta p_t - i_t - u_t$ | 0.385 | 0.184 | 0.055 | 0.083 | | | 0.381 | 0.092 | 0.049 | 0.156 | | |
| $i_t - u_t - \Delta p_t$ | 0.366 | 0.314 | 0.366 | 0.112 | | | 0.374 | 0.254 | 0.362 | 0.185 | | |
| $i_t - \Delta p_t - u_t$ | 0.484 | 0.556 | 0.271 | 0.164 | | | 0.504 | 0.531 | 0.301 | 0.263 | | |
| | NST | | | | | | KT | | | | | |
| $h = 4$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $u_t - \Delta p_t - i_t$ | 0.052 | 0.042 | 0.208 | 0.034 | 0.063 | 0.045 | 0.440 | 0.301 | 0.638 | 0.361 | 0.487 | 0.457 |
| $u_t - i_t - \Delta p_t$ | 0.051 | 0.038 | 0.194 | 0.088 | | | 0.384 | 0.219 | 0.413 | 0.449 | | |
| $\Delta p_t - u_t - i_t$ | 0.010 | 0.067 | 0.024 | 0.022 | | | 0.245 | 0.442 | 0.331 | 0.323 | | |
| $\Delta p_t - i_t - u_t$ | 0.020 | 0.007 | 0.000 | 0.007 | | | 0.185 | 0.107 | 0.121 | 0.286 | | |
| $i_t - u_t - \Delta p_t$ | 0.122 | 0.004 | 0.755 | 0.039 | | | 0.525 | 0.087 | 0.839 | 0.355 | | |
| $i_t - \Delta p_t - u_t$ | 0.032 | 0.008 | 0.240 | 0.129 | | | 0.416 | 0.187 | 0.488 | 0.444 | | |
| **Normal approximation** | | | | | | | | | | | | |
| | NST | | | | | | KT | | | | | |
| $h = 1$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $u_t - \Delta p_t - i_t$ | 0.032 | 0.110 | 0.058 | 0.001 | 0.001 | 0.000 | 0.031 | 0.147 | 0.045 | 0.012 | 0.012 | 0.006 |
| $u_t - i_t - \Delta p_t$ | 0.027 | 0.116 | 0.154 | | | | 0.033 | 0.151 | 0.163 | | | |
| $\Delta p_t - u_t - i_t$ | 0.032 | 0.125 | 0.021 | | | | 0.032 | 0.162 | 0.016 | | | |
| $\Delta p_t - i_t - u_t$ | 0.007 | 0.150 | 0.005 | | | | 0.009 | 0.138 | 0.003 | | | |
| $i_t - u_t - \Delta p_t$ | 0.005 | 0.166 | 0.009 | | | | 0.007 | 0.164 | 0.014 | | | |
| $i_t - \Delta p_t - u_t$ | 0.009 | 0.149 | 0.008 | | | | 0.013 | 0.128 | 0.018 | | | |
| | NST | | | | | | KT | | | | | |
| $h = 4$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $u_t - \Delta p_t - i_t$ | 0.020 | 0.004 | 0.595 | 0.000 | 0.000 | 0.002 | 0.196 | 0.067 | 0.712 | 0.016 | 0.015 | 0.049 |
| $u_t - i_t - \Delta p_t$ | 0.057 | 0.142 | 0.267 | | | | 0.303 | 0.412 | 0.530 | | | |
| $\Delta p_t - u_t - i_t$ | 0.028 | 0.008 | 0.605 | | | | 0.241 | 0.074 | 0.727 | | | |
| $\Delta p_t - i_t - u_t$ | 0.086 | 0.144 | 0.238 | | | | 0.532 | 0.521 | 0.671 | | | |
| $i_t - u_t - \Delta p_t$ | 0.122 | 0.007 | 0.097 | | | | 0.659 | 0.136 | 0.439 | | | |
| $i_t - \Delta p_t - u_t$ | 0.099 | 0.004 | 0.305 | | | | 0.629 | 0.115 | 0.639 | | | |

**Notes:** The table shows the p-values corresponding to the various combinations of tests on uniformity and transformations of the multivariate PITs for all possible permutations of the data. For those transformations that yield order-invariant test statistics, we only report one p-value. NST refers to Neyman's smooth test (Neyman, 1937). KT refers to the test proposed by Knüppel (2015).

## 5 Conclusion

In this paper we show how order-invariant tests can be derived for testing the proper calibration of multivariate densities of arbitrary dimension. We demonstrate that distortions in rejection rates can be very large when "cheating" based on existing tests which are not order invariant. Furthermore, we show that the new tests have very good power properties for a wide range of deviations from the null hypothesis; this holds true, in particular, when the data exhibit fat tails that are not taken into account by the null model. We want to stress again that our approach, which essentially relies on transforming the multivariate problem to a univariate one, is compatible with any existing method for testing univariate distributions and we recommend using the powerful Neyman smooth test in general, but the test by Knüppel (2015) whenever one is concerned with autocorrelation. The new tests remain superior when using modified versions for the case that parameters of the Gaussian DGP have to be estimated from the data (for instance, when testing the in-sample fit of a density model), although a loss of power results in this case.

In the previous section, we have presented two empirical applications to demonstrate the usefulness of our approach. We believe there is a wide range of other applications in various fields. First, the proposed methods can be applied whenever properly calibrated density forecasts are crucial to form well-informed decisions (about production, investment, pricing, etc.) and could foster the use of multivariate density forecasts in situations in which decisions are based on loss functions that take more than one variable as arguments. Second, the proposed methods can be used to improve the specification of multivariate models taking higher moments into account; obvious applications of this kind are common in financial econometrics, e. g., for estimating the Value-at-Risk of a portfolio, but it can be expected that the modeling of the dependence structure of higher moments of multivariate data becomes more common also for demand management or in macroeconomics (e. g., Tay, 2015).

Our study leaves room for future research along several dimensions. First, especially for financial applications, it would be interesting to extend those results of our paper which are limited to the case of multivariate Gaussian processes under the null hypothesis to more general settings. In particular, the application of the randomization device to test composite hypothesis in the context of goodness-of-fit tests needs to be extended to more general distributions. Second, we believe it may be possible to develop tests with even better power for very high-dimensional densities; this could be achieved by selecting the terms entering the $Z^{*\dagger}$ transformation in a data-driven way or by assigning weights to the conditional PITs entering the transformations. This would be relevant, for instance, when modeling the joint behavior of future returns of large portfolios and when working with densities derived from large-scale macroeconomic VAR models (Crespo Cuaresma et al., 2014; Dovern et al., 2015). Finally, our approach may be compared with tests based on the Kendall distribution function as studied in Ziegel and Gneiting (2014) or Genest and Rivest (2001).

## References

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., and Thorsrud, L. A. (2014). Nowcasting GDP in real time: A density combination approach. *Journal of Business & Economic Statistics*, 32(1):48–68.

Adolfson, M., Lindé, J., and Villani, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Reviews*, 26(2-4):289–328.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Andrews, D. W. K. (1997). A conditional kolmogorov test. *Econometrica*, 65(5):1097–1128.

Bai, J. (2003). Testing parametric conditional distributions of dynamic models. *The Review of Economics and Statistics*, 85(3):531–549.

Bai, J. and Chen, Z. (2008). Testing multivariate distributions in GARCH models. *Journal of Econometrics*, 143(1):19–36.

Bera, A. K. and Ghosh, A. (2002). Neyman's smooth test and its applications in econometrics. In Ullah, A., Wan, A. T. K., and Chaturvedi, A., editors, *Handbook of Applied Econometrics and Statistical Inference*, pages 177–230. Marcel Dekker, New York.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–74.

Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29(3):327–341.

Clements, M. P. and Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting*, 19:144–165.

Clements, M. P. and Smith, J. (2002). Evaluating multivariate forecast densities: A comparison of two approaches. *International Journal of Forecasting*, 18(3):397–407.

Corradi, V. and Swanson, N. R. (2006). Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics*, 133(2):779–806.

Crespo Cuaresma, J., Feldkircher, M., and Huber, F. (2014). Forecasting with Bayesian global vector autoregressive models: A comparison of priors. Working Papers 189, Oesterreichische Nationalbank (Austrian Central Bank).

David, F. N. and Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35:182–190.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *J. Roy. Statist. Soc. Ser. A*, 147(2):278–292.

De Gooijer, J. G. (2007). Power of the Neyman smooth test for evaluating multivariate forecast densities. *Journal of Applied Statistics*, 34(4):371–381.

Del Negro, M. and Primiceri, G. E. (2015). Time varying structural vector autoregressions and monetary policy: a corrigendum. *Review of Economic Studies*, forthcoming.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):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. *The Review of Economics and Statistics*, 81(4):661–673.

Dovern, J., Feldkircher, M., and Huber, F. (2015). Does joint modelling of the world economy pay off? Evaluating global forecasts from a Bayesian GVAR. Working Papers 200, Oesterreichische Nationalbank (Austrian Central Bank).

Durbin, J. (1961). Some methods of constructing exact tests. *Biometrika*, 48:41–55.

Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics*, 20:339–350.

Genest, C. and Rivest, L.-P. (2001). On the multivariate probability integral transformation. *Statistics & Probability Letters*, 53(4):391–399.

Ghosh, A. and Bera, A. K. (2015). Density forecast evaluation for dependent financial data: Theory and applications. mimeo.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69(2):243–268.

González-Barrios, J. M., O'Reilly, F., and Rueda, R. (2010). Durbin's random substitution and conditional Monte Carlo. *Metrika*, 72:369–383.

González-Rivera, G. and Yoldas, E. (2012). Autocontour-based evaluation of multivariate predictive densities. *International Journal of Forecasting*, 28(2):328–342.

Hallam, M. and Olmo, J. (2014). Semiparametric density forecasts of daily financial returns from intraday data. *Journal of Financial Econometrics*, 12(2):408–432.

Hansen, E. B. (1994). Autoregressive density estimation. *International Economic Review*, 35:705–730.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Herbst, E. and Schorfheide, F. (2012). Evaluating DSGE model forecasts of comovements. *Journal of Econometrics*, 171(2):152–166.

Hong, Y. and Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies*, 18(1):37–84.

Huurman, C., Ravazzolo, F., and Zhou, C. (2012). The power of weather. *Computational Statistics & Data Analysis*, 56(11):3793–3807.

Ishida, I. (2005). Scanning multivariate conditional densities with probability integral transforms. CARF F-Series CARF-F-045, Center for Advanced Research in Finance, Faculty of Economics, The University of Tokyo.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Kitsul, Y. and Wright, J. H. (2013). The economics of options-implied inflation probability density functions. *Journal of Financial Economics*, 110(3):696–711.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33(2):270–281.

Ko, S. I. M. and Park, S. Y. (2013). Multivariate density forecast evaluation: A modified approach. *International Journal of Forecasting*, 29(3):431–441.

Krüger, F., Clark, T. E., and Ravazzolo, F. (2016). Using entropic tilting to combine BVAR forecasts with external nowcasts. *Journal of Business & Economic Statistics*, forthcoming.

Li, Q., Lin, J., and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, 31(1):57–65.

Manner, H. and Reznikova, O. (2012). A survey on time-varying copulas: Specification, simulations and application. *Econometric Reviews*, 31(6):654–687.

Mitchell, J. and Wallis, K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26(6):1023–1040.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–08.

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

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72(3):821–852.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472.

Shackleton, M. B., Taylor, S. J., and Yu, P. (2010). A multi-horizon comparison of density forecasts for the s&p 500 using index returns and option prices. *Journal of Banking & Finance*, 34(11):2678–2693.

Szkutnik (2012). On the durbin-wagle randomization device and some of its applications. *Journal of Multivariate Analysis*, 109:103–108.

Tay, A. (2015). A brief survey of density forecasting in macroeconomics. Macroeconomic Review October 2015, Monetary Authority of Singapore.

Taylor, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3):534–549.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wagle, B. (1968). Multivariate beta distribution and a test for multivariate normality. *Journal of the Royal Statistical Society B*, 30:511–516.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Wolters, M. H. (2015). Evaluating point and density forecasts of DSGE models. *Journal of Applied Econometrics*, 30(1):74–96.

Ziegel, J. F. and Gneiting, T. (2014). Copula calibration. *Electronic Journal of Statistics*, 8(2):26192638.

## Appendix A Proofs

***Proof of Proposition 1.*** The proof is done by induction. To simplify notations we drop the time subscript and replace the conditional PITs by a sequence of independent $\mathcal{U}(0,1)$ random variables $U_1, U_2, \ldots$.

Step 1 $(d = 2)$: For $d = 2$ the density is given by

$$f_{P_2}(P_2) = \frac{(-1)^1}{1!}\log(P_2) = -\log^2(P_2),$$

which is equal to the density derived in Clements and Smith (2000). Note that we could also start at $d = 1$, for which the density is equal to 1, corresponding to the uniform distribution.
Step 2 $(d \to d+1)$: Consider the change of variables

$$P_{d+1} = P_d U_{d+1}$$

The determinant of the Jacobian for the inverse transformation is

$$J = \det \frac{\partial(P_d, U_{d+1})}{\partial(P_{d+1}, U_{d+1})} = \begin{vmatrix} \frac{1}{U_{d+1}} & -\frac{P_{d+1}}{U^2_{d+1}} \\ 0 & 1 \end{vmatrix} = \frac{1}{U_{d+1}}.$$

The joint density of $P_{d+1}$ and $U_{d+1}$ is

$$f_{P_{d+1},U_{d+1}}(P_{d+1}, U_{d+1}) = f_{P_d}\left(\frac{P_{d+1}}{U_{d+1}}\right) \cdot \frac{1}{U_{d+1}} = \frac{(-1)^{d-1}}{(d-1)!}\log^{d-1}\left(\frac{P_{d+1}}{U_{d+1}}\right) \cdot \frac{1}{U_{d+1}},$$

with $0 < P_{d+1} < U_{d+1} < 1$. Therefore, the marginal PDF of $P_{d+1}$ is

$$f_{P_{d+1}}(P_{d+1}) = \int_{P_{d+1}}^{1} f_{P_d}\left(\frac{P_{d+1}}{U_{d+1}}\right) \cdot \frac{1}{U_{d+1}} d \cdot U_{d+1} = \frac{(-1)^d}{d!}\log^d(P_{d+1}) = f_{d+1}(P_{d+1}).$$

To show that the CDF is correct first note that

$$f'_{P_d}(P_d) = \frac{(-1)^{d-1}}{(d-2)!}\log^{d-2}(P_d) \cdot \frac{1}{P_d} = -1 \cdot f_{P_{d-1}}(P_d) \cdot \frac{1}{P_d}.$$

It follows that

$$F'(P_d) = \sum_{i=0}^{d-1} f_{P_{d-i}}(P_d) - \sum_{i=1}^{d-1} f_{P_{d-i}}(P_d) = f_{P_d}(P_d).$$

□

***Proof of Proposition 2.*** Again, the proof is done by induction and again for simplicity we consider a sequence of independent $\mathcal{U}(0,1)$ random variables $U_1, U_2, \ldots$.

Step 1 $(d = 2)$: Consider the change of variables

$$P^*_2 = (U_1 - 0.5)(U_2 - 0.5) = U^*_1 U^*_2.$$

The determinant of the Jacobian is

$$J = \frac{1}{U^*_2},$$

so the joint density of $P_2^*$ and $U_2^*$ is given by

$$f_{P_2^*,U_2^*} = \left|\frac{1}{U_2^*}\right|.$$

Integrating out $U_2^*$ gives

$$f_{P_2^*}(P_2^*) = \int_{-1/2}^{1/2} \left|\frac{1}{U_2^*}\right| 2 \cdot U_2^* = 2 \cdot \int_{|2P_2^*|}^{1/2} \frac{1}{U_2^*} = 2\log(U_2^*)\Big|_{|2P_2^*|}^{1/2} = 2\log\left|\frac{1}{4P_2^*}\right|,$$

where the second equality follows from the symmetry around 0 and the fact that $|2P_2^*| < |U_2^*| < 1/2$.

Step 2 ($d \to d+1$): Consider the following change of variables

$$P_{d+1}^* = P_d^*(U_{d+1} - 0.5) = P_d^* U_{d+1}^*.$$

The determinant of the Jacobian is

$$J = \frac{1}{U_{d+1}^*},$$

and therefore the joint density of $P_{d+1}^*$ and $U_{d+1}^*$ is

$$f_{P_{d+1}^*,U_{d+1}^*} = f_{P_d^*}\left(\frac{P_{d+1}^*}{U_{d+1}^*}\right)\left|\frac{1}{U_{d+1}^*}\right|.$$

The PDF of $P_{d+1}^*$ then is

$$\begin{aligned} f_{P_{d+1}^*}(P_{d+1}^*) &= \int_{-1/2}^{1/2} f_{P_d^*}\left(\frac{P_{d+1}^*}{U_{d+1}^*}\right)\left|\frac{1}{U_{d+1}^*}\right| d \cdot U_{d+1}^* \\ &= 2 \cdot \int_{|2^d P_{d+1}^*|}^{1/2} \frac{2^{d-1}}{(d-1)!} \log^{d-1}\left(\frac{U_{d+1}^*}{2^d\left|P_{d+1}^*\right|}\right) \frac{1}{U_{d+1}} d \cdot U_{d+1} \\ &= 2 \cdot \frac{2^{d-1}}{(d-1)!}\frac{1}{d} \log^d\left(\frac{U_{d+1}^*}{2^d\left|P_{d+1}^*\right|}\right)\Bigg|_{|2^d P_2^*|}^{1/2} = \frac{2^d}{(d)!}\log^d\left|\frac{1}{2^{d+1}P_{d+1}^*}\right|. \end{aligned}$$

Again the symmetry around 0 and the fact that $|2^d P_{d+1}^*| < |U_{d+1}^*| < 1/2$ was used.

Now consider the CDF. Note that

$$\frac{f'_{P^*d}(P_d^*)}{2^{d-1}} = \frac{1}{(d-2)!}\log\left|\frac{1}{2^d P_d^*}\right|(-1)\frac{1}{P_d}.$$

Then using the product rule

$$\begin{aligned} F'_{P_d^*}(P_d^*) &= 2^{d-1}\sum_{i=1}^{d}\frac{1}{(d-i)!}\log^{d-i}\left|\frac{1}{2^d P_d^*}\right| P_d^* - P_d^* \sum_{i=2}^{d}\frac{1}{(d-i)!}\log^{d-i}\left|\frac{1}{2^d P_d^*}\right|\frac{1}{P^*d} \\ &= \frac{2^{d-1}}{(d-1)!}\log^{d-1}\left|\frac{1}{2^d P_d^*}\right| = f_{P_d^*}(P_d^*). \end{aligned}$$

The addition of 1/2 (see Proposition 2) ensures that the CDF lies between 0 and 1. □

***Proof of Proposition 3.*** Under independence, we have $U_t^{i|1:i-1} = U_t^i$, i. e., the conditional CDF is equal to the marginal CDF. In this case, the product transformation reduces to $P_{t,d} = \prod_{i=1}^d U_t^i$. This is clearly robust to permutations. The same argument can be made for the location-adjusted version $P^*_{t,d}$. The stacked transformation then becomes $S_t = [U_t^1, \dots, U_t^d]'$, which again is obviously order invariant.

Now consider the following two permutations: $\pi_1 = (1, 2, 3, \dots, d)$ and $\pi_2 = (2, 1, 3, \dots, d)$. For these permutations, the product transformations only differ in their first two components. So w.l.g., we only check that independence is needed for $U_t^1 \cdot U_t^{2|1} = U_t^2 \cdot U_t^{1|2}$ to hold. The latter equality is equivalent to $\frac{U_t^1}{U_t^2} = \frac{U_t^{1|2}}{U_t^{2|1}}$ for all $t$, which does not hold in general, unless we have independence.

For these two permutations order invariance in $S_t$ is given only if $[U_t^1 \; U_t^{2|1}]'$ is equal to $[U_t^2 \; U_t^{1|2}]'$ for all $t$, which again only holds under independence. □

***Proof of Proposition 4.*** W.l.g. let $\mu = 0$, which can be achieved by demeaning the original data. Rewrite $Y_t$ as

$$\begin{aligned} Y_{1,t} &= Z_{1,t} \\ Y_{2,t} &= \beta_{2,1} Y_{1,t} + Z_{2t} \\ &\vdots \\ Y_{d,t} &= \beta_{d,1} Y_{1,t} + \beta_{d,2} Y_{2,t} + \ldots + \beta_{d,d-1} Y_{d-1,t} + Z_{d,t}, \end{aligned}$$

with $Z_{i,t}$ normally distributed. Writing this more compactly we obtain

$$BY_t = Z_t,$$

where $Z_t = (Z_{1,t}, \dots, Z_{d,t})'$, with

$$\mathbb{E}(Z_t Z_t') = D = \operatorname{diag}\begin{pmatrix} \sigma_1^2 \\ \sigma_{2|1}^2 \\ \vdots \\ \sigma_{d|1:d-1}^2 \end{pmatrix}$$

and

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots \\ -\beta_{2,1} & 1 & 0 & \dots \\ -\beta_{3,1} & -\beta_{3,2} & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the matrix of population regression coefficients, whose precise form in terms of the covariance matrix is directly available using standard results on conditional normal random variables. It holds that $U_t^1 = \Phi(Z_{1,t}/\sigma_1), U_t^{2|1} = \Phi(Z_{2,t}/\sigma_{2|1}), \dots, U_t^{d|1:d-1} = \Phi(Z_{d,t}/\sigma_{d|1:d-1})$. Furthermore, note that

$$\operatorname{Cov}(Y_t) = \mathbb{E}(Y_t Y_t') = \Sigma = B^{-1} D B^{-1'}.$$

Consequently,

$$Z_t^2 = (Z_t' D^{-1/2})(D^{-1/2} Z_t) = Y_t' B' D^{-1} B Y_t = Y_t' \Sigma^{-1} Y_t.$$

The last term is clearly invariant to the ordering of the variables. □

***Proof of Proposition 5.*** Consider the generic term $\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right) \sim \mathcal{N}(0,1)$, where $\gamma_i^k$ stands for a set of indices representing the conditioning variables. Under normality, these terms are also jointly normally distributed. Then the fact that $Z_t^{*2}$ has a mixture of independent $\chi_1^2$ random variables follows directly from Lemma 17.1 in van der Vaart (1998). The weights of the mixture are given by the eigenvalues of the covariance matrix of the terms $\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right)$ for all $i = 1,\dots,d$ and $k = 1,\dots,2^{d-1}$. This matrix is actually a correlation matrix due to the unit variance of the inverse normal transformation. To compute this correlation matrix, we start with the covariance between $Y^t_{i|\gamma_i^k}$ and $Y^t_{j|\gamma_j^l}$. Then, dropping the time index, $Y_i$ conditional on the vector $Y_{\gamma_i^k}$ is

$$Y_i|Y_{\gamma_i^k} = Y_i - \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}Y_{\gamma_i^k},$$

which has variance equal to $\Sigma_{ii} - \Sigma_{i,\gamma_i^k}R^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,i}$. Consequently,

$$\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right) = \frac{Y_i - \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}Y_{\gamma_i^k}}{(\Sigma_{ii} - \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,i})^{1/2}}$$

and analogously for $\Phi^{-1}\left(U_t^{j|\gamma_j^l}\right)$. Then the computation of the covariance/correlation is straightforward. The reduced rank of $R_{Z^*}$ follows from the fact that all conditional variables $Y^t_{i|\gamma_i^k}$ are a linear combination of the original $d$ variables. □

***Proof of Proposition 6.*** Consider

$$\hat{Z}_{i,t|1:i-1} = \frac{Y_{i,t|1:i-1} - \hat{\mu}_{i|1:i-1}}{\hat{\sigma}_{i|1:i-1}} = \frac{(Y_{i,t|1:i-1} - \mu_{i|1:i-1}) - (\hat{\mu}_{i|1:i-1} - \mu_{i|1:i-1})}{\sigma_{i|1:i-1}}\left(\frac{\sigma_{i|1:i-1}}{\hat{\sigma}_{i|1:i-1}}\right)$$

$$\Leftrightarrow \hat{Z}_{i,t|1:i-1}\left(\frac{\hat{\sigma}_{i|1:i-1}}{\sigma_{i|1:i-1}}\right) + \left(\frac{\hat{\mu}_{i|1:i-1} - \mu_{i|1:i-1}}{\sigma_{i|1:i-1}}\right) = \frac{Y_{i,t|1:i-1} - \mu_{i|1:i-1}}{\sigma_{i|1:i-1}} = Z_{i,t|1:i-1}.$$

Now note that $Z_{i,t|1:i-1} \sim \mathcal{N}(0,1)$, $\left(\frac{\hat{\sigma}_{i|1:i-1}}{\sigma_{i|1:i-1}}\right) \sim \sqrt{\chi^2_{n-1}/(n-1)}$, and $\left(\frac{\hat{\mu}_{i|1:i-1} - \mu_{i|1:i-1}}{\sigma_{i|1:i-1}}\right) \sim \mathcal{N}(0,1/n)$. Due to the independence of $Y_{i,t|1:i-1}$ for $i = 1,\dots,d$, the estimated conditional means and variances are also independent across $i$. The result from the proposition then follows from Durbin (1961) and Szkutnik (2012).

□

***Proof of Proposition 7.*** First note that the joint distribution of the estimated vector of conditional means consisting of the $d \cdot d^{n-1}$ terms $\hat{\mu}_i|\gamma_i^k$, $\forall i,k$, denoted by $\hat{\mu}_c$, is given by

$$\hat{\mu}_c = [\hat{\mu}_1,\dots,\hat{\mu}_{d\cdot d^{n-1}}]' \sim MV\mathcal{N}(\mu_c, \Sigma_{Z^*}),$$

with a typical entry of $\Sigma_{Z^*}$ being given by

$$\text{Cov}\left(Y_t^{i|\gamma_i^k}, Y_t^{j|\gamma_j^l}\right) =$$
$$(\Sigma_{i,j} - \Sigma_{j,\gamma_j^l}\Sigma^{-1}_{\gamma_j^l,\gamma_j^l}\Sigma_{\gamma_j^l,i} - \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,j} + \Sigma_{i,\gamma_i^k}\Sigma^{-1}_{\gamma_i^k,\gamma_i^k}\Sigma_{\gamma_i^k,\gamma_j^l}\Sigma^{-1}_{\gamma_j^l,\gamma_j^l}\Sigma_{\gamma_j^l,j}),$$

see Proposition 5 and its proof for details. Then for $D = \text{diag}(\sigma_1^2, \dots, \sigma_{d \cdot d^{n-1}}^2)$, the diagonal matrix containing the diagonal elements of $\Sigma_{Z^*}$, we have

$$D^{-1/2}(\hat{\mu}_c - \mu_c) \sim MV\mathcal{N}(0, \frac{1}{n} R_{Z^*}),$$

which justifies the randomization vector $m$. Furthermore,

$$\hat{\Sigma}_{Z^*} \sim \mathcal{W}(\Sigma_{Z^*}, n-1)/(n-1)$$

and thus

$$D^{-1/2}\hat{\Sigma}_{Z^*} \sim \mathcal{W}(R_{Z^*}, n-1)/(n-1),$$

justifying the randomization vector $s^2$.

Next, consider the distinct terms $\tilde{Z}_{i,t|\gamma_i^k} = s_{i|\gamma_i^k} \hat{Z}_{i,t|\gamma_i^k} + m_{i|\gamma_i^k}$ entering the transformation. By the proof of Proposition 6, these terms are each $\mathcal{N}(0,1)$ distributed, but they are not independent in general. However, the randomization by a single draw from the distribution $s_{i|\gamma_i^k}$ and $m_{i|\gamma_i^k}$ does not alter the correlation, i. e., $Corr\left(\tilde{Z}_{i,t|\gamma_i^k}, \tilde{Z}_{jt|\gamma_j^{k'}}\right) = Corr\left(Z_{i,t|\gamma_i^k}, Z_{jt|\gamma_j^{k'}}\right)$. By Proposition 5 the distribution of $\tilde{Z}_t^{2*}$ follows. □

***Proof of Corollary 2.*** For the feasible test based on $\hat{R}_{Z^*}$, due to the fact that $\hat{\Sigma} \to \Sigma$ as $n \to \infty$ it follows that $\hat{R}_{Z^*} \to R_{Z^*}$ as $n \to \infty$. Similarly, the eigenvalues of $\hat{R}_{Z^*}$ converge to the eigenvalues of $R_{Z^*}$. The result follows by the continuous mapping theorem. □

## Appendix B Tests for Uniformity

In this paper, we consider three tests for whether the transformation $\hat{F}_{W_t}(W_t)$ is uniformly distributed: Neyman's smooth test, the Kolmogorov-Smirnov test, and a test proposed in Knüppel (2015) that allows us to account for autocorrelation in a straightforward manner.

### B.1 Neyman's Smooth Test

Bera and Ghosh (2002) and De Gooijer (2007) advocates testing uniformity with Neyman's smooth test (Neyman, 1937). The test statistic based on the first four normalized Legendre polynomials is given by

$$\Psi_4^2 = \sum_{i=1}^{4} u_i^2, \tag{B.1}$$

with $u_1^2 = 3n\hat{\mu}_1^2$, $u_2^2 = \frac{45n(\hat{\mu}_2-1/3)^2}{4}$, $u_3^2 = \frac{7n(5\hat{\mu}_3-3\hat{\mu}_1)^2}{4}$, $u_4^2 = \frac{9n(35(\hat{\mu}_4-1/5)-30(\hat{\mu}_2-1/3))^2}{64}$, and $\hat{\mu}_i = \left(\sum_{t=1}^{n}(2U_t-1)^i\right)/n$. Under $H_0$ the statistic $\Psi_4^2 \sim \chi_4^2$.

### B.2 Kolmogorov-Smirnov Test

This test is based on the maximum distance between the empirically observed distribution function and the theoretical distribution function under $H_0$. In our case, the latter is known to be $\mathcal{U}(0,1)$. The empirical distribution function is given by

$$U_n(x) = P_n\left(\hat{F}_{W_t}(W_t) < x\right) = \frac{1}{n}\sum_{t=1}^{n} I\left(\hat{F}_{W_t}(W_t) < x\right) \tag{B.2}$$

and measures how many sample points are below x. Now the Kolmogorov-Smirnov test statistic is given by the supremum of the set of distances between those two functions

$$D = \sup_x |U_n(x) - F_u(x)|, \tag{B.3}$$

with $F_u$ denoting the CDF of the $\mathcal{U}(0,1)$ distribution. Under $H_0$, the (scaled) test statistic asymptotically converges to the Kolmogorov distribution which is based on a Brownian bridge ($B(t)$). Thus, for large $n$:

$$\sqrt{n}D \sim \sup_{t\in[0,1]} |B(t)| \tag{B.4}$$

Since for continuous distribution functions (such as in our case) the distribution of $D$ under $H_0$ is independent of this function, values for finite $n$ are available in tabulated form.

### B.3 Knüppel's Test Based on Raw-Moments

The test by Knüppel (2015) is summarized in Section 2.4.

## Appendix C Additional Results from Monte Carlo Simulations

Table C.1: Size and power - known parameters (Kolmogorov-Smirnov test)

| Size | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.051 | 0.048 | 0.047 | 0.050 | 0.049 | 0.050 | 0.047 | 0.052 | 0.051 | 0.049 | 0.049 | 0.050 | 0.053 | 0.053 | 0.049 | 0.049 | 0.050 | 0.049 |
| $d = 3$ | 0.054 | 0.051 | 0.049 | 0.053 | 0.055 | 0.052 | 0.048 | 0.047 | 0.049 | 0.047 | 0.048 | 0.049 | 0.050 | 0.049 | 0.048 | 0.045 | 0.048 | 0.048 |
| $d = 4$ | 0.051 | 0.046 | 0.052 | 0.046 | 0.046 | 0.046 | 0.048 | 0.052 | 0.056 | 0.051 | 0.052 | 0.053 | 0.052 | 0.049 | 0.050 | 0.048 | 0.050 | 0.052 |
| $d = 5$ | 0.052 | 0.049 | 0.049 | 0.047 | 0.052 | 0.049 | 0.049 | 0.052 | 0.053 | 0.050 | 0.048 | 0.052 | 0.048 | 0.048 | 0.053 | 0.050 | 0.053 | 0.050 |
| $d = 6$ | 0.054 | 0.050 | 0.051 | 0.049 | 0.051 | 0.053 | 0.051 | 0.045 | 0.054 | 0.052 | 0.051 | 0.052 | 0.048 | 0.050 | 0.050 | 0.049 | 0.050 | 0.052 |
| Power against $H_1$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.079 | 0.078 | 0.074 | 0.189 | 0.190 | 0.173 | 0.105 | 0.095 | 0.096 | 0.317 | 0.318 | 0.284 | 0.151 | 0.134 | 0.121 | 0.553 | 0.552 | 0.498 |
| $d = 3$ | 0.090 | 0.088 | 0.079 | 0.262 | 0.261 | 0.223 | 0.126 | 0.121 | 0.097 | 0.471 | 0.469 | 0.410 | 0.224 | 0.199 | 0.144 | 0.762 | 0.761 | 0.685 |
| $d = 4$ | 0.099 | 0.103 | 0.087 | 0.337 | 0.339 | 0.300 | 0.158 | 0.159 | 0.116 | 0.597 | 0.591 | 0.526 | 0.290 | 0.250 | 0.181 | 0.880 | 0.879 | 0.819 |
| $d = 5$ | 0.114 | 0.114 | 0.092 | 0.421 | 0.415 | 0.362 | 0.180 | 0.185 | 0.129 | 0.710 | 0.701 | 0.639 | 0.372 | 0.333 | 0.208 | 0.944 | 0.941 | 0.907 |
| $d = 6$ | 0.124 | 0.129 | 0.099 | 0.491 | 0.477 | 0.432 | 0.222 | 0.217 | 0.143 | 0.788 | 0.781 | 0.739 | 0.455 | 0.395 | 0.242 | 0.977 | 0.973 | 0.955 |
| Power against $H_2$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.041 | 0.054 | 0.095 | 0.063 | 0.064 | 0.080 | 0.049 | 0.063 | 0.134 | 0.077 | 0.078 | 0.105 | 0.053 | 0.073 | 0.211 | 0.101 | 0.102 | 0.169 |
| $d = 3$ | 0.048 | 0.065 | 0.057 | 0.104 | 0.109 | 0.153 | 0.053 | 0.074 | 0.066 | 0.147 | 0.157 | 0.247 | 0.067 | 0.109 | 0.073 | 0.248 | 0.269 | 0.445 |
| $d = 4$ | 0.055 | 0.076 | 0.068 | 0.157 | 0.172 | 0.238 | 0.062 | 0.093 | 0.077 | 0.249 | 0.285 | 0.398 | 0.093 | 0.148 | 0.089 | 0.438 | 0.498 | 0.677 |
| $d = 5$ | 0.053 | 0.084 | 0.067 | 0.203 | 0.236 | 0.300 | 0.074 | 0.118 | 0.081 | 0.364 | 0.430 | 0.535 | 0.127 | 0.193 | 0.107 | 0.620 | 0.704 | 0.824 |
| $d = 6$ | 0.064 | 0.097 | 0.076 | 0.270 | 0.316 | 0.372 | 0.095 | 0.145 | 0.088 | 0.482 | 0.560 | 0.644 | 0.171 | 0.257 | 0.135 | 0.772 | 0.852 | 0.911 |
| Power against $H_3$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.085 | 0.091 | 0.134 | 0.283 | 0.283 | 0.306 | 0.135 | 0.143 | 0.223 | 0.498 | 0.499 | 0.543 | 0.249 | 0.222 | 0.409 | 0.794 | 0.793 | 0.836 |
| $d = 3$ | 0.109 | 0.142 | 0.113 | 0.483 | 0.496 | 0.545 | 0.212 | 0.233 | 0.156 | 0.790 | 0.805 | 0.838 | 0.484 | 0.412 | 0.289 | 0.977 | 0.981 | 0.988 |
| $d = 4$ | 0.169 | 0.204 | 0.137 | 0.692 | 0.712 | 0.736 | 0.350 | 0.348 | 0.218 | 0.931 | 0.942 | 0.955 | 0.720 | 0.609 | 0.407 | 0.998 | 0.999 | 0.999 |
| $d = 5$ | 0.227 | 0.260 | 0.167 | 0.815 | 0.840 | 0.857 | 0.487 | 0.458 | 0.264 | 0.984 | 0.987 | 0.991 | 0.888 | 0.744 | 0.518 | 1.000 | 1.000 | 1.000 |
| $d = 6$ | 0.292 | 0.326 | 0.198 | 0.905 | 0.924 | 0.931 | 0.635 | 0.564 | 0.332 | 0.997 | 0.997 | 0.998 | 0.964 | 0.846 | 0.629 | 1.000 | 1.000 | 1.000 |
| Power against $H_4$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.055 | 0.054 | 0.052 | 0.164 | 0.165 | 0.152 | 0.084 | 0.070 | 0.074 | 0.281 | 0.281 | 0.243 | 0.129 | 0.104 | 0.104 | 0.497 | 0.499 | 0.442 |
| $d = 3$ | 0.069 | 0.065 | 0.059 | 0.241 | 0.242 | 0.205 | 0.104 | 0.082 | 0.080 | 0.426 | 0.423 | 0.358 | 0.185 | 0.128 | 0.129 | 0.710 | 0.709 | 0.624 |
| $d = 4$ | 0.088 | 0.074 | 0.066 | 0.333 | 0.330 | 0.286 | 0.131 | 0.097 | 0.092 | 0.566 | 0.560 | 0.484 | 0.249 | 0.163 | 0.165 | 0.859 | 0.856 | 0.778 |
| $d = 5$ | 0.106 | 0.075 | 0.074 | 0.406 | 0.402 | 0.352 | 0.158 | 0.119 | 0.110 | 0.688 | 0.678 | 0.603 | 0.316 | 0.208 | 0.204 | 0.940 | 0.936 | 0.896 |
| $d = 6$ | 0.118 | 0.091 | 0.083 | 0.503 | 0.491 | 0.445 | 0.196 | 0.136 | 0.139 | 0.784 | 0.769 | 0.715 | 0.388 | 0.237 | 0.252 | 0.979 | 0.975 | 0.959 |
| Power against $H_5$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.045 | 0.056 | 0.092 | 0.092 | 0.093 | 0.119 | 0.052 | 0.064 | 0.129 | 0.133 | 0.133 | 0.182 | 0.052 | 0.075 | 0.206 | 0.252 | 0.252 | 0.352 |
| $d = 3$ | 0.049 | 0.070 | 0.059 | 0.157 | 0.167 | 0.209 | 0.058 | 0.091 | 0.062 | 0.301 | 0.319 | 0.398 | 0.084 | 0.133 | 0.073 | 0.594 | 0.622 | 0.716 |
| $d = 4$ | 0.056 | 0.093 | 0.059 | 0.253 | 0.273 | 0.319 | 0.075 | 0.127 | 0.072 | 0.486 | 0.524 | 0.585 | 0.128 | 0.195 | 0.084 | 0.831 | 0.865 | 0.896 |
| $d = 5$ | 0.071 | 0.102 | 0.073 | 0.359 | 0.394 | 0.418 | 0.104 | 0.165 | 0.082 | 0.653 | 0.702 | 0.736 | 0.195 | 0.284 | 0.104 | 0.945 | 0.963 | 0.970 |
| $d = 6$ | 0.080 | 0.121 | 0.074 | 0.440 | 0.482 | 0.504 | 0.135 | 0.189 | 0.090 | 0.780 | 0.817 | 0.833 | 0.294 | 0.365 | 0.129 | 0.984 | 0.990 | 0.993 |

**Notes:** Rejection frequencies of Kolmogorov-Smirnov test based on the transformations introduced in Section 2.2 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. The alternative hypotheses are defined in Section 3. All Monte Carlo simulations are based on 10,000 iterations.

Table C.2: Size and power - known parameters (Knüppel test)

| Size | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.056 | 0.056 | 0.055 | 0.056 | 0.056 | 0.057 | 0.053 | 0.052 | 0.051 | 0.051 | 0.053 | 0.051 | 0.048 | 0.049 | 0.051 | 0.048 | 0.048 | 0.048 |
| $d = 3$ | 0.047 | 0.053 | 0.055 | 0.056 | 0.054 | 0.055 | 0.051 | 0.053 | 0.052 | 0.054 | 0.055 | 0.053 | 0.052 | 0.052 | 0.052 | 0.055 | 0.053 | 0.055 |
| $d = 4$ | 0.050 | 0.057 | 0.057 | 0.060 | 0.059 | 0.059 | 0.047 | 0.049 | 0.053 | 0.054 | 0.054 | 0.053 | 0.048 | 0.052 | 0.056 | 0.052 | 0.053 | 0.056 |
| $d = 5$ | 0.051 | 0.059 | 0.060 | 0.057 | 0.055 | 0.057 | 0.047 | 0.051 | 0.055 | 0.053 | 0.052 | 0.053 | 0.049 | 0.052 | 0.049 | 0.049 | 0.047 | 0.047 |
| $d = 6$ | 0.053 | 0.060 | 0.056 | 0.057 | 0.059 | 0.057 | 0.051 | 0.054 | 0.046 | 0.049 | 0.051 | 0.049 | 0.055 | 0.056 | 0.052 | 0.048 | 0.052 | 0.050 |
| Power against $H_1$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.098 | 0.065 | 0.063 | 0.117 | 0.117 | 0.108 | 0.209 | 0.109 | 0.111 | 0.261 | 0.261 | 0.213 | 0.436 | 0.225 | 0.236 | 0.510 | 0.511 | 0.425 |
| $d = 3$ | 0.146 | 0.072 | 0.071 | 0.174 | 0.175 | 0.148 | 0.313 | 0.123 | 0.136 | 0.375 | 0.378 | 0.315 | 0.641 | 0.260 | 0.287 | 0.718 | 0.719 | 0.616 |
| $d = 4$ | 0.199 | 0.076 | 0.082 | 0.232 | 0.228 | 0.200 | 0.431 | 0.150 | 0.168 | 0.503 | 0.497 | 0.428 | 0.795 | 0.309 | 0.356 | 0.858 | 0.851 | 0.785 |
| $d = 5$ | 0.258 | 0.085 | 0.097 | 0.306 | 0.298 | 0.258 | 0.541 | 0.159 | 0.190 | 0.628 | 0.617 | 0.556 | 0.894 | 0.363 | 0.423 | 0.933 | 0.926 | 0.882 |
| $d = 6$ | 0.316 | 0.095 | 0.104 | 0.365 | 0.355 | 0.312 | 0.641 | 0.194 | 0.221 | 0.713 | 0.703 | 0.647 | 0.945 | 0.417 | 0.493 | 0.969 | 0.964 | 0.946 |
| Power against $H_2$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.041 | 0.062 | 0.072 | 0.054 | 0.054 | 0.059 | 0.053 | 0.066 | 0.106 | 0.063 | 0.064 | 0.097 | 0.071 | 0.079 | 0.196 | 0.086 | 0.086 | 0.174 |
| $d = 3$ | 0.057 | 0.070 | 0.050 | 0.070 | 0.076 | 0.105 | 0.086 | 0.080 | 0.056 | 0.103 | 0.113 | 0.197 | 0.161 | 0.098 | 0.090 | 0.203 | 0.231 | 0.412 |
| $d = 4$ | 0.082 | 0.073 | 0.062 | 0.101 | 0.114 | 0.153 | 0.143 | 0.090 | 0.075 | 0.177 | 0.211 | 0.314 | 0.297 | 0.131 | 0.133 | 0.371 | 0.444 | 0.626 |
| $d = 5$ | 0.107 | 0.080 | 0.062 | 0.135 | 0.163 | 0.211 | 0.224 | 0.106 | 0.097 | 0.281 | 0.344 | 0.441 | 0.471 | 0.153 | 0.178 | 0.557 | 0.664 | 0.792 |
| $d = 6$ | 0.140 | 0.085 | 0.060 | 0.174 | 0.214 | 0.257 | 0.305 | 0.110 | 0.116 | 0.375 | 0.470 | 0.552 | 0.629 | 0.193 | 0.224 | 0.716 | 0.822 | 0.890 |
| Power against $H_3$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.146 | 0.070 | 0.119 | 0.203 | 0.203 | 0.230 | 0.344 | 0.117 | 0.269 | 0.439 | 0.438 | 0.505 | 0.700 | 0.245 | 0.565 | 0.788 | 0.787 | 0.855 |
| $d = 3$ | 0.304 | 0.098 | 0.125 | 0.393 | 0.406 | 0.447 | 0.654 | 0.195 | 0.289 | 0.752 | 0.771 | 0.821 | 0.953 | 0.400 | 0.624 | 0.979 | 0.983 | 0.992 |
| $d = 4$ | 0.479 | 0.139 | 0.171 | 0.581 | 0.612 | 0.651 | 0.868 | 0.287 | 0.396 | 0.925 | 0.941 | 0.956 | 0.997 | 0.577 | 0.777 | 0.999 | 1.000 | 1.000 |
| $d = 5$ | 0.662 | 0.183 | 0.216 | 0.749 | 0.782 | 0.804 | 0.955 | 0.378 | 0.509 | 0.981 | 0.985 | 0.989 | 1.000 | 0.731 | 0.879 | 1.000 | 1.000 | 1.000 |
| $d = 6$ | 0.784 | 0.227 | 0.278 | 0.857 | 0.884 | 0.897 | 0.988 | 0.488 | 0.616 | 0.996 | 0.998 | 0.998 | 1.000 | 0.831 | 0.945 | 1.000 | 1.000 | 1.000 |
| Power against $H_4$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.116 | 0.083 | 0.092 | 0.119 | 0.120 | 0.094 | 0.173 | 0.114 | 0.128 | 0.228 | 0.229 | 0.181 | 0.304 | 0.174 | 0.227 | 0.477 | 0.478 | 0.379 |
| $d = 3$ | 0.148 | 0.085 | 0.103 | 0.184 | 0.184 | 0.148 | 0.251 | 0.119 | 0.158 | 0.418 | 0.419 | 0.324 | 0.447 | 0.198 | 0.305 | 0.783 | 0.782 | 0.657 |
| $d = 4$ | 0.194 | 0.087 | 0.122 | 0.285 | 0.283 | 0.220 | 0.311 | 0.133 | 0.198 | 0.617 | 0.611 | 0.504 | 0.576 | 0.235 | 0.396 | 0.944 | 0.940 | 0.877 |
| $d = 5$ | 0.233 | 0.093 | 0.134 | 0.404 | 0.397 | 0.328 | 0.393 | 0.150 | 0.244 | 0.792 | 0.781 | 0.690 | 0.674 | 0.276 | 0.487 | 0.989 | 0.987 | 0.967 |
| $d = 6$ | 0.280 | 0.101 | 0.149 | 0.510 | 0.494 | 0.433 | 0.468 | 0.163 | 0.296 | 0.892 | 0.875 | 0.821 | 0.752 | 0.311 | 0.580 | 0.998 | 0.998 | 0.995 |
| Power against $H_5$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.081 | 0.060 | 0.080 | 0.127 | 0.127 | 0.144 | 0.163 | 0.094 | 0.156 | 0.299 | 0.300 | 0.330 | 0.357 | 0.177 | 0.334 | 0.613 | 0.612 | 0.667 |
| $d = 3$ | 0.156 | 0.083 | 0.066 | 0.258 | 0.267 | 0.274 | 0.323 | 0.147 | 0.119 | 0.561 | 0.583 | 0.599 | 0.644 | 0.300 | 0.272 | 0.899 | 0.911 | 0.923 |
| $d = 4$ | 0.243 | 0.108 | 0.077 | 0.379 | 0.401 | 0.401 | 0.499 | 0.209 | 0.157 | 0.758 | 0.786 | 0.783 | 0.838 | 0.455 | 0.377 | 0.982 | 0.987 | 0.987 |
| $d = 5$ | 0.335 | 0.127 | 0.086 | 0.498 | 0.528 | 0.531 | 0.650 | 0.284 | 0.209 | 0.880 | 0.896 | 0.892 | 0.927 | 0.594 | 0.483 | 0.996 | 0.998 | 0.997 |
| $d = 6$ | 0.432 | 0.162 | 0.110 | 0.610 | 0.636 | 0.626 | 0.746 | 0.355 | 0.252 | 0.942 | 0.953 | 0.949 | 0.974 | 0.702 | 0.592 | 0.999 | 1.000 | 1.000 |

**Notes:** Rejection frequencies of Knüppel (2015) test based on the transformations introduced in Section 2.2 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. The alternative hypotheses are defined in Section 3. All Monte Carlo simulations are based on 10,000 iterations.

## Table C.3: Size and power - estimated parameters (Kolmogorov-Smirnov test)

| Size (original test) | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.046 | 0.040 | 0.005 | 0.007 | 0.007 | 0.006 | 0.046 | 0.041 | 0.005 | 0.006 | 0.007 | 0.008 | 0.042 | 0.038 | 0.005 | 0.006 | 0.006 | 0.006 |
| $d = 3$ | 0.049 | 0.042 | 0.049 | 0.005 | 0.005 | 0.006 | 0.043 | 0.035 | 0.047 | 0.003 | 0.003 | 0.005 | 0.045 | 0.036 | 0.040 | 0.004 | 0.004 | 0.004 |
| $d = 4$ | 0.051 | 0.043 | 0.047 | 0.004 | 0.005 | 0.005 | 0.046 | 0.038 | 0.047 | 0.003 | 0.003 | 0.004 | 0.044 | 0.035 | 0.046 | 0.003 | 0.003 | 0.004 |
| $d = 5$ | 0.055 | 0.046 | 0.049 | 0.003 | 0.003 | 0.004 | 0.048 | 0.043 | 0.049 | 0.003 | 0.002 | 0.003 | 0.043 | 0.037 | 0.046 | 0.002 | 0.002 | 0.003 |
| $d = 6$ | 0.053 | 0.043 | 0.047 | 0.004 | 0.003 | 0.005 | 0.048 | 0.036 | 0.048 | 0.003 | 0.002 | 0.004 | 0.047 | 0.037 | 0.047 | 0.003 | 0.002 | 0.002 |
| Size (adjusted test) | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.050 | 0.043 | 0.006 | 0.052 | 0.057 | 0.061 | 0.052 | 0.047 | 0.006 | 0.051 | 0.051 | 0.063 | 0.053 | 0.048 | 0.005 | 0.049 | 0.052 | 0.063 |
| $d = 3$ | 0.055 | 0.045 | 0.049 | 0.053 | 0.056 | 0.060 | 0.052 | 0.046 | 0.052 | 0.053 | 0.058 | 0.059 | 0.048 | 0.044 | 0.047 | 0.048 | 0.053 | 0.056 |
| $d = 4$ | 0.057 | 0.051 | 0.051 | 0.051 | 0.059 | 0.062 | 0.051 | 0.046 | 0.054 | 0.051 | 0.055 | 0.056 | 0.052 | 0.046 | 0.054 | 0.049 | 0.051 | 0.054 |
| $d = 5$ | 0.058 | 0.053 | 0.049 | 0.062 | 0.063 | 0.061 | 0.056 | 0.048 | 0.047 | 0.050 | 0.055 | 0.055 | 0.054 | 0.046 | 0.053 | 0.050 | 0.054 | 0.054 |
| $d = 6$ | 0.066 | 0.054 | 0.050 | 0.059 | 0.060 | 0.057 | 0.054 | 0.048 | 0.050 | 0.055 | 0.058 | 0.057 | 0.056 | 0.049 | 0.051 | 0.052 | 0.051 | 0.053 |
| Power against $H_4$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.060 | 0.054 | 0.025 | 0.149 | 0.144 | 0.134 | 0.085 | 0.063 | 0.035 | 0.253 | 0.257 | 0.237 | 0.133 | 0.097 | 0.059 | 0.457 | 0.455 | 0.414 |
| $d = 3$ | 0.071 | 0.054 | 0.059 | 0.175 | 0.168 | 0.147 | 0.100 | 0.077 | 0.077 | 0.355 | 0.352 | 0.282 | 0.178 | 0.116 | 0.133 | 0.642 | 0.641 | 0.530 |
| $d = 4$ | 0.081 | 0.058 | 0.062 | 0.194 | 0.190 | 0.144 | 0.120 | 0.083 | 0.094 | 0.438 | 0.433 | 0.339 | 0.224 | 0.145 | 0.158 | 0.772 | 0.764 | 0.660 |
| $d = 5$ | 0.080 | 0.064 | 0.063 | 0.215 | 0.200 | 0.140 | 0.139 | 0.095 | 0.098 | 0.526 | 0.503 | 0.382 | 0.270 | 0.162 | 0.188 | 0.870 | 0.864 | 0.767 |
| $d = 6$ | 0.090 | 0.067 | 0.064 | 0.226 | 0.210 | 0.127 | 0.156 | 0.102 | 0.100 | 0.590 | 0.560 | 0.427 | 0.315 | 0.193 | 0.215 | 0.930 | 0.921 | 0.843 |

**Notes:** Rejection frequencies of Kolmogorov-Smirnov test based on the transformations introduced in Section 2.2 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. The alternative hypotheses are defined in Section 3. All Monte Carlo simulations are based on 10,000 iterations.

## Table C.4: Size and power - estimated parameters (Knüppel test)

| Size (original test) | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.027 | 0.022 | 0.016 | 0.033 | 0.033 | 0.029 | 0.025 | 0.023 | 0.015 | 0.031 | 0.031 | 0.024 | 0.026 | 0.021 | 0.012 | 0.025 | 0.026 | 0.024 |
| $d = 3$ | 0.026 | 0.023 | 0.038 | 0.044 | 0.043 | 0.036 | 0.029 | 0.023 | 0.036 | 0.033 | 0.033 | 0.030 | 0.029 | 0.024 | 0.034 | 0.029 | 0.030 | 0.026 |
| $d = 4$ | 0.028 | 0.024 | 0.042 | 0.054 | 0.055 | 0.044 | 0.026 | 0.022 | 0.039 | 0.034 | 0.033 | 0.030 | 0.026 | 0.027 | 0.037 | 0.031 | 0.029 | 0.027 |
| $d = 5$ | 0.033 | 0.027 | 0.043 | 0.061 | 0.058 | 0.045 | 0.029 | 0.023 | 0.037 | 0.037 | 0.037 | 0.035 | 0.026 | 0.024 | 0.037 | 0.032 | 0.033 | 0.033 |
| $d = 6$ | 0.032 | 0.028 | 0.041 | 0.075 | 0.074 | 0.058 | 0.028 | 0.026 | 0.040 | 0.047 | 0.044 | 0.042 | 0.028 | 0.024 | 0.038 | 0.030 | 0.031 | 0.031 |
| Size (adjusted test) | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.056 | 0.035 | 0.038 | 0.063 | 0.062 | 0.051 | 0.050 | 0.033 | 0.033 | 0.051 | 0.052 | 0.050 | 0.053 | 0.035 | 0.028 | 0.049 | 0.055 | 0.049 |
| $d = 3$ | 0.056 | 0.033 | 0.062 | 0.065 | 0.067 | 0.062 | 0.055 | 0.034 | 0.057 | 0.058 | 0.057 | 0.056 | 0.051 | 0.033 | 0.058 | 0.054 | 0.057 | 0.053 |
| $d = 4$ | 0.055 | 0.032 | 0.067 | 0.064 | 0.068 | 0.060 | 0.054 | 0.032 | 0.061 | 0.060 | 0.056 | 0.057 | 0.054 | 0.036 | 0.053 | 0.051 | 0.051 | 0.053 |
| $d = 5$ | 0.061 | 0.030 | 0.070 | 0.074 | 0.077 | 0.068 | 0.055 | 0.031 | 0.064 | 0.059 | 0.062 | 0.060 | 0.052 | 0.032 | 0.055 | 0.054 | 0.058 | 0.058 |
| $d = 6$ | 0.059 | 0.032 | 0.068 | 0.080 | 0.086 | 0.071 | 0.056 | 0.033 | 0.061 | 0.068 | 0.067 | 0.063 | 0.055 | 0.035 | 0.056 | 0.058 | 0.061 | 0.060 |
| Power against $H_4$ | $n = 50$ | | | | | | $n = 100$ | | | | | | $n = 200$ | | | | | |
| | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ | $S$ | $P$ | $P^*$ | $Z^2$ | $Z^{2*}$ | $Z^{2\dagger}$ |
| $d = 2$ | 0.093 | 0.054 | 0.060 | 0.099 | 0.093 | 0.087 | 0.152 | 0.086 | 0.093 | 0.206 | 0.201 | 0.169 | 0.283 | 0.151 | 0.173 | 0.436 | 0.434 | 0.349 |
| $d = 3$ | 0.114 | 0.049 | 0.090 | 0.139 | 0.129 | 0.097 | 0.200 | 0.081 | 0.132 | 0.337 | 0.333 | 0.254 | 0.393 | 0.159 | 0.280 | 0.696 | 0.696 | 0.561 |
| $d = 4$ | 0.131 | 0.054 | 0.090 | 0.176 | 0.168 | 0.119 | 0.246 | 0.091 | 0.165 | 0.479 | 0.468 | 0.353 | 0.496 | 0.189 | 0.341 | 0.878 | 0.877 | 0.764 |
| $d = 5$ | 0.141 | 0.051 | 0.090 | 0.214 | 0.202 | 0.130 | 0.295 | 0.100 | 0.189 | 0.619 | 0.600 | 0.450 | 0.575 | 0.217 | 0.422 | 0.958 | 0.954 | 0.884 |
| $d = 6$ | 0.147 | 0.049 | 0.092 | 0.246 | 0.232 | 0.126 | 0.323 | 0.102 | 0.202 | 0.716 | 0.701 | 0.536 | 0.644 | 0.246 | 0.481 | 0.987 | 0.985 | 0.953 |

**Notes:** Rejection frequencies of Knüppel (2015) test based on the transformations introduced in Section 2.2 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. The alternative hypotheses are defined in Section 3. All Monte Carlo simulations are based on 10,000 iterations.