Received: 11 October 2018 | Revised: 29 November 2019

DOI: 10.1002/jae.2755

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Order-invariant tests for proper calibration of multivariate density forecasts

Jonas Dovern[1,2] | **Hans Manner**[3]

[1]Institute for Economics, Friedrich-Alexander University Erlangen-Nürnberg, Nuremberg, Germany

[2]CESifo, Munich, Germany

[3]Institute of Economics, University of Graz, Graz, Austria

Correspondence
Hans Manner, Institute of Economics, University of Graz, Universitätsstr. 15/F4, 8010 Graz, Austria.
Email: hans.manner@uni-graz.at

Summary

Established tests for proper calibration of multivariate density forecasts based on Rosenblatt probability integral transforms can be manipulated by changing the order of variables in the forecasting model. We derive order-invariant tests. The new tests are applicable to densities of arbitrary dimensions and can deal with parameter estimation uncertainty and dynamic misspecification. Monte Carlo simulations show that they often have superior power relative to established approaches. We use the tests to evaluate generalized autoregressive conditional heteroskedasticity-based multivariate density forecasts for a vector of stock market returns and macroeconomic forecasts from a Bayesian vector autoregression with time-varying parameters.

1 | INTRODUCTION

The use of density forecasts has recently become common in many scientific fields (Gneiting & Katzfuss, 2014) and, in particular, in many areas of economics. Density forecasts are increasingly used, for instance, in the fields of energy economics (Huurman, Ravazzolo, & Zhou, 2012), demand management (Taylor, 2012), finance (Hallam & Olmo, 2014), and macroeconomics (Clark, 2011; Wolters, 2015). Many tasks, such as the computation of value-at-risk measures for portfolios containing multiple assets or the planning of production for a firm that serves many markets from one central production facility, require the construction and evaluation of *multivariate* density forecasts. Beginning with Smith (1985) and Diebold, Hahn, and Tay (1999), the literature has proposed several approaches for testing whether a sequence of multivariate density forecasts coincides with the corresponding true densities (e.g., Bai & Chen, 2008; Clements and Smith, 2000, 2002; Corradi & Swanson, 2006a; Ko & Park, 2013).

This strand of the literature has neglected two important issues. First, established tests depend on the order of variables in a multivariate model.[1] This offers room for data mining if a researcher decides to report only those results that correspond to one particular ("preferred") order. Second, most empirical applications and many of the theoretical results focus on the bivariate case. However, many applications, especially in finance, require models of higher dimensionality to be 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). Examples include Clements and Smith, (2000, 2002) and Ko and Park, (2013, 2019). This approach relies on a factorization of the forecast distribution into condi-

[1]Note that there is a complementary literature on methods for comparing the relative accuracy of (multivariate) density forecasts (see, e.g.,Gneiting & Katzfuss, 2014). This literature usually compares scores of different density forecast models that do not suffer from dependence on the order of variables.

tional distributions because these, in turn, can be used to form independent PITs which, for well-specified models, follow a uniform distribution. Suitable transformations of these conditional PITs then lead to a reduction of the multivariate testing problem to a univariate one.

We contribute to the literature by proposing new variants of such transformations of the conditional PITs. The new transformations have a number of advantages. First, they are *order invariant*—a concept we define below—meaning that test results do not depend on the order of variables in the forecasting model. We show that the distortions in rejection rates caused by a tendentious application of the established tests, which are not order invariant, can be very substantial. Second, the new tests are applicable to densities of arbitrary dimension. Third, they have better power (relative to established tests) against a wide range of alternatives. In two applications, we show that the new tests are helpful for testing the appropriateness of density forecasts based on sophisticated multivariate models for vectors of financial returns and to evaluate macroeconomic density forecasts. In particular, we show that the potential for data mining is immense when using the established tests in practice and that our order-invariant tests are required to draw unambiguous conclusions.

Our tests can also be used when dynamic misspecification and parameter uncertainty have to be taken into account. These two aspects are of utmost importance in practical applications that involve parametric forecasting models. Corradi and Swanson (2006b) present a comprehensive overview of both aspects.

Dynamic misspecification refers to the fact that a forecaster potentially uses only a subset of the relevant information to form a conditional density forecast. In most fields of economics and finance such misspecification is very likely. Dynamic misspecification causes the PITs to be serially correlated. A number of papers propose tests that are robust against dynamic misspecification; that is, they preserve this misspecification under the null hypothesis. Pioneering work in this context has been done by Corradi and Swanson (2006a), who show that a block bootstrap can be used to adjust Kolmogorov-type tests under such conditions. In contrast, Rossi and Sekhposyan (2013) relax this assumption and propose a test for correct specification of density forecasts that is robust, in addition, to structural breaks. Other papers (Berkowitz, 2001; Hong, Li, & Zhao, 2007; Ko & Park, 2019; Lin & Wu, 2017) jointly test for uniformity and the i.i.d. property of the PITs, thereby testing the null hypothesis of completely calibrated densities (Mitchell & Wallis, 2011).

Parameter estimation uncertainty arises whenever a parametric forecast model is used to construct density forecasts whose parameters are estimated based on finite samples. Whether estimation uncertainty has to be dealt with when evaluating a sequence of predictive densities depends on the exact formulation of the null hypothesis one is interested in. One common approach is to test whether the forecast distribution belongs to a given parametric density family with parameters evaluated at their pseudo-true values. The alternative view, proposed in Rossi and Sekhposyan (2019), is to test for the ability of a model to produce correct forecast distributions evaluated at the estimated parameter values. Bai (2003) (for the univariate case) and Bai and Chen (2008) (for multivariate densities) combine the Kolmogorov test with Khmaladze's martingale transformation to obtain a test which is distribution free in the presence of estimated parameters. Andrews (1997) solves this problem by using a parametric bootstrap. More recently, Chen (2011) adapts a number of tests from the parameter-free context to parameter-dependent density forecast evaluation, building on insights from Newey (1985) and Tauchen (1985) in the in-sample case and from West (1996) and West and McCracken (1998) in the out-of-sample case. This is the approach that we use in our paper.

Thus there is a wide range of views regarding how density forecasts should be tested (Table 1). The methods that we propose below are compatible with any combinations of those views.

TABLE 1 Classification of testing problems

Treatment of dynamic misspecification:	Known parameters/ Forecasts as primitives	Estimated parameters
Tests that ignore dynamic misspecification	Diebold et al. (1998, 1999), Clements and Smith (2000, 2002), Ko and Park (2013)	Andrews (1997), Bai (2003), Chen (2011)
Tests robust to dynamic misspecification	Knüppel (2015)	Corradi and Swanson (2006a)
Tests that test for dynamic misspecification	Berkowitz (2001), Rossi and Sekhposyan (2019)	Hong and Li (2005), Hong et al. (2007), Ko and Park (2019), Lin and Wu (2017), González-Rivera and Sun (2015)

Note. This table contains a nonexhaustive collection of papers taking different views on how parameter estimation and dynamic misspecification should be treated when testing the calibration of (predictive) densities.

The remainder of this paper is organized as follows. In Section 2 we describe the testing problem, generalize established tests, and derive new tests to evaluate multivariate densities. In Section 3 we assess the finite-sample properties of different tests by means of Monte Carlo simulations. In Section 4 we demonstrate the usefulness of the newly proposed tests in an application to forecast the distribution of a vector of stock returns. Section 5 concludes.

2 | THEORY

2.1 | Setup and test hypothesis

Let $Y_t = [Y_{1,t} \dots Y_{d,t}]'$ be a vector-valued continuous random variable with true (but unknown) conditional distribution function (CDF) $G_{Y_t}(y|\mathfrak{F}_{t-1})$, where $\mathfrak{F}_{t-1}$ denotes the relevant information set available at time $t-1$. Furthermore, we consider the predictive CDF $F_{Y_t}(y|\Omega_{t-1}, \theta_0)$ with corresponding conditional probability density function (PDF) $F_{Y_t}(y|\Omega_{t-1}, \theta_0)$, where $\Omega_{t-1} \subseteq \mathfrak{F}_{t-1}$ is the information set available to the researcher and θ_0 denotes a parameter vector with compact and finite parameter space Θ. This framework takes into account that density forecasts are often constructed using parametric models and allows for dynamic misspecification as defined, for instance, by Corradi and Swanson (2006a). For the time being, we treat θ_0 as known, but we also discuss how we can take estimation uncertainty into account.

Consider a sample $\{y_t, \Omega_{t-1}\}_{t=1}^{n}$ of which the first R observations can potentially be used to estimate θ_0 and the remaining P observations are used to evaluate the predictive densities generated by $F_{Y_t}(y|\Omega_{t-1}, \theta_0)$. We are interested in testing whether the model $F_{Y_t}(y|\Omega_{t-1}, \theta_0)$ is correctly specified in the sense that

$$H_0 \; : \; F_{Y_t}(y|\Omega_{t-1}, \theta_0) = G_{Y_t}(y|\mathfrak{F}_{t-1}) \quad \text{a.s. } \forall \; y \quad \text{in} \quad R^d, \; \forall \; t = R+1, \; \dots \; , T. \tag{1}$$

We call a density forecast satisfying Equation (1) properly calibrated. To specify the null exactly, assumptions need to be made about whether θ_0 has to be estimated and about whether dynamic misspecification can be ignored, should be controlled for, or should jointly be tested. In Table 1 we provide an overview about the assumptions made in the literature.

In the univariate case, H_0 implies that the probability integral transform (PIT), given by $U_t = F_{Y_t}(Y_t)$, is uniformly distributed between 0 and 1 (see, e.g.,Gneiting & Katzfuss, 2014). This fact can be used to test for proper density calibration (e.g., Dawid, 1984; Diebold et al., 1998).

Unfortunately, matters are more complicated in the multivariate case because the distribution of the multivariate PITs of Y_t under the null is unknown, in general, for $d > 1$ (see, e.g.,Genest & Rivest, 2001). In essence, the task then is to reduce the multivariate problem to a univariate one by using suitable transformations. The commonly used approach is based on the factorization of the joint densities into the product of conditional densities. Let $F_{Y_i}(y|\Omega_{t-1}, \theta_0)$ denote the marginal (conditional) CDF for the ith element of Y_t and denote by $F_{Y_i|Y_{i-1},\dots,Y_1}(y|Y_{i-1,t}, \; \dots \; , Y_{1,t}, \Omega_{t-1}, \theta_0)$ the conditional distribution of $Y_{i,t}$ given $Y_{i-1,t}, \; \dots \; , Y_{1,t}$, and by $F_{Y_i}(y|\Omega_{t-1}, \theta_0)$ and $f_{Y_i|Y_{i-1},\dots,Y_1}(y|Y_{i-1,t}, \; \dots \; , Y_{1,t}, \Omega_{t-1}, \theta_0)$ the corresponding PDFs. Rosenblatt (1952) shows that the sequences of *conditional PITs* for the elements of Y_t given by

$$U_t^1 = F_{Y_1}(Y_{1,t}), \; U_t^{2|1} = F_{Y_2|Y_1}(Y_{2,t}), \; \dots \; , U_t^{d|1,\dots,d-1} = U_t^{d|1:d-1} = F_{Y_d|Y_{d-1},\dots,Y_1}(Y_{d,t}) \tag{2}$$

are independent of each other and distributed $\mathcal{U}(0,1)$. The next step is to obtain a univariate testing problem based on this vector of PITs.

A commonly used approach is to transform the vector-valued random variable Y_t into a scalar random variable and to compute PITs for this transformed random variable. The computation of the conditional PITs is often an intermediate step in such transformations. 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}. The PIT of W_t is given by

$$U_t^W = F_{W_t}(W_t). \tag{3}$$

A particular class of such transformations is based on the conditional PITs defined by Equation (2); that is, $W_t = g_t(Y_t) = \tilde{g}_t \left[U_t^1(Y_t), \; U_t^{2|1}(Y_t), \; \dots \; , \; U_t^{d|1,\dots,d-1}(Y_t) \right]$.

In general, testing H_0 is equivalent to testing whether $U_t^W \sim \mathcal{U}(0,1)$. In the absence of dynamic misspecification, the PITs are also independently distributed across time under H_0; that is, $U_t^W \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$.

Different transformations $g_t(\cdot)$ have been considered in the literature. Clements and Smith (2000) propose evaluating density forecasts based on the product of the conditional PITs corresponding to one particular permutation of the variables. In this case, the transformation function $g_t(\cdot)$ is given by

$$\mathrm{CS}_t = g(Y_t) = \prod_{i=1}^{d} U_t^{i|1:i-1}, \tag{4}$$

where we define $U_t^{1|1:0} = U_t^1$ and assume that it is implicitly understood that the statistic depends on the dimension d. Ko and Park (2013) explain why tests based on CS_t 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 is given by

$$\mathrm{KP}_t = g(Y_t) = \prod_{i=1}^{d} \left(U_t^{i|1:i-1} - 0.5 \right). \tag{5}$$

Diebold et al. (1999) achieve the reduction of dimension somewhat differently by stacking all conditional PITs. More formally, if we let

$$S_t = \left[U_t^{d|1:d-1}, \ \dots \ , U_t^1 \right]', \tag{6}$$

then $S = \left[S'_{R+1}, S'_{R+2}, \ \dots \ , S'_n \right]'$ constitutes a vector of variables that are uniformly distributed under H_0.

2.2 | The order of variables

So far, we have implicitly assumed that there exists a natural order of variables from 1 to d. This, of course, is not true, as already mentioned in most papers on the topic (Clements & Smith, 2002; Diebold et al., 1999; Hong & Li, 2005; Ishida, 2005). Ordering the elements in Y_t in a different way will generally lead to different results because the Rosenblatt transform in Equation (2) clearly depends on the order of the variables. Consequently, the outcome of a hypothesis test will depend on the selected order. 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 orders and only report the outcome with the largest or smallest p-value. 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 made.

We use the following notation for different permutations of the variables. Let $\{\pi_k\}$, for $k = 1, \ \dots \ , d!$, 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 kth permutation. Then, the conditional PITs under permutation π_k are given by

$$\begin{aligned}
U_t^{\pi_k(1)} &= F_{Y_{\pi_k(1)}}(Y_{\pi_k(1),t}), \\
U_t^{\pi_k(2)|\pi_k(1)} &= F_{Y_{\pi_k(2)}|Y_{\pi_k(1)}}(Y_{\pi_k(2),t}), \\
&\ \ \vdots \\
U_t^{\pi_k(d)|\pi_k(1):\pi_k(d-1)} &= F_{Y_{\pi_k(d)}|Y_{\pi_k(d-1),t}, \ \dots \ , Y_{\pi_k(1),t}}(Y_{\pi_k(d),t}).
\end{aligned} \tag{7}$$

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

The next proposition shows that tests based on the established 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.

Proposition 1. *Test statistics $T(\pi_k)$ based on $\{CS_t\}_{t=1}^n$, $\{KP_t\}_{t=1}^n$ and on the stacked transformation $\{S_t\}_{t=1}^n$ are order invariant if and only if under H_0 the variables $Y_{1,t}, \ \dots \ , Y_{d,t}$ are independent—that is, when $f_{Y_t}(Y_t) = f_{Y_1}(Y_{1,t}) \times \ \dots \ \times f_{Y_d}(Y_{d,t})$.*

In the next section, we first discuss a transformation that is based on the Rosenblatt transformation and can be order invariant under less restrictive conditions, and we derive new transformations that are always order invariant.

2.3 | New transformations

The first transformation that we propose leads to order-invariant test statistics under less restrictive conditions and forms the basis for additional transformations that always lead to order-invariant tests. Consider the following transformation, which is based on the squares of inverse normal transforms of the PITs for one particular permutation of the data:

$$Z_t^2 = \sum_{i=1}^{d} \left[\Phi^{-1} \left(U_t^{i|1:i-1} \right) \right]^2, \tag{8}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution.

Proposition 2. *Test statistics $T(\pi_k)$ based on $\left\{ Z_t^2 \right\}_{t=1}^n$ are order invariant if under H_0 $Y_t \sim \mathcal{N}(\mu, \Sigma)$—that is, when Y_t follows a multivariate normal distribution with mean vector μ and covariance matrix Σ or when $Y_{1,t}, \ldots, Y_{d,t}$ are independent under H_0.*

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. The proof of Proposition 2 in the Supporting Information Appendix 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).

Ideally, however, we would like to obtain a transformation that is order invariant in general. A transformation that fulfills this criterion is similar in structure to Z_t^2 but considers the sum over all distinct conditional PITs. Consider all possible permutations π_k for $k = 1, \ldots, d!$ and the corresponding sequences of conditional PITs defined by Equation (7). The number of distinct PITs is $d \times \sum_{k=0}^{d-1} \binom{d-1}{k} = d \times 2^{d-1}$. Let γ_i^k, for $k = 1, \ldots, 2^{d-1}$ be the set of all sets of conditioning variables corresponding to all distinct conditional PITs for $Y_{i,t}$. Then the suggested transformation has the form

$$Z_t^{2*} = \sum_{i=1}^{d} \sum_{k=1}^{2^{d-1}} \left[\Phi^{-1} \left(U_t^{i|\gamma_i^k} \right) \right]^2. \tag{9}$$

Since all distinct conditional PITs enter into this transformation and, thus, the initial order of the variables in Y_t is irrelevant, order invariance is clearly ensured for any test statistic based on Z_t^{2*}.

When d increases, the number of terms entering Z_t^{2*} can become prohibitively large. In this case, it appears sensible to use an order-invariant transformation for which the number of terms grows only linearly with d. We propose to consider only such conditional PITs corresponding to each $Y_{i,t}$ that are conditional on all other variables. We think that this choice has some merits since, on the one hand, the considered subset of conditional PITs contains rich information about the dependence structure of the elements of Y_t while, on the other hand, also containing information about the margins. Denoting those conditional PITs by $U_t^{i|-i}$, the new transformation is given by

$$Z_t^{2\dagger} = \sum_{i=1}^{d} \left[\Phi^{-1} \left(U_t^{i|-i} \right) \right]^2. \tag{10}$$

2.4 | Distribution of transformations

To test H_0 based on the transformations, we need to know their distribution under the null hypothesis as indicated by Equation (3). The transformation S_t is simply a vector of independent uniformly distributed random variables under H_0. Clements and Smith (2000) derive the distribution of CS_t for $d = 2, 3$, and Ko and Park (2013) provide the distribution of KP_t for $d = 2$. In the Supporting Information Appendix, we derive the distributions of CS_t and KP_t for arbitrary d.

Next, we derive the distributions of the new transformations. We distinguish two cases. In the general case, we do not make any assumptions about the distribution of Y_t, except that it is continuous. The corresponding results include, for instance, cases in which H_0 implies non-Gaussian parametric distributions of Y_t or its distribution is not available analytically, so that the conditional PITs have to be calculated numerically. For the special case of normally distributed Y_t, we show that the distributions of Z^{2*} and $Z^{2\dagger}$ become much more tractable.

2.4.1 | Distributions of new transformations: General case

As shown in Section 2.1, the different conditional PITs for one particular permutation are independent. Therefore, H_0 implies that $Z_t^2 \sim \chi_d^2$, where χ_d^2 denotes the chi-squared distribution with d degrees of freedom. Denoting its CDF by $F_{\chi_d^2}$, the random variable $U_t^{Z^2} = F_{\chi_d^2}\left(Z_t^2\right)$ is distributed $\mathcal{U}(0,1)$ under H_0.

The transformation Z_t^{2*} is similar to Z_t^2. However, due to the fact that the summands in Equation (9) are not independent in general, Z_t^{2*} no longer follows a χ^2 distribution under H_0. The same argument applies in the case of $Z_t^{2\dagger}$. However, we can straightforwardly obtain the distributions of the transformations by Monte Carlo simulation as long as it is possible to generate random draws from the density model under H_0.

The following algorithm describes how the distributions of $Z_t^\bullet \in \left\{Z_t^{2*}, Z_t^{2\dagger}\right\}$ can be approximated numerically to compute $U_t^{Z^\bullet}$. We would like to stress that this algorithm is exclusively used to approximate this distribution for a given parameter value that can be either θ_0 or an estimate $\hat{\theta}$.[2]

1. Generate M conditional forecasts, $y_t^{(m)}$, based on the model under H_0; that is, draw repeatedly from the conditional predictive densities $f_{Y_t}(y)$.
2. Given $f_{Y_t}(y)$, construct $u_{t,(m)}^{i|\gamma_i^k}$, $\forall i,k$, for $m = 1, \ldots, M$ along the lines described in Section 2.3.
3. Compute the corresponding inverse PITs as $\Phi^{-1}\left(u_{t,(m)}^{i|\gamma_i^k}\right)$.
4. Based on the set of $\Phi^{-1}\left(u_{t,(m)}^{i|\gamma_i^k}\right)$, compute $z_{t,(m)}^\bullet$ using Equation (9) or 10, respectively.
5. Compute $u_t^{Z^\bullet} = \Pr\left(Z_t^\bullet < z_{t,(m)}^\bullet\right)$ 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.

If H_0 holds, $U_t^{Z^\bullet}$ is distributed $\mathcal{U}(0,1)$ for M sufficiently large. The validity of this simulation approach is straightforward as we simulate directly from the (parametric) null distribution and only apply continuous transformations.

2.4.2 | Distributions of new transformations: Gaussian case

Under the assumption that Y_t is normally distributed, the distributions of Z_t^{2*} and $Z_t^{2\dagger}$ are available analytically and do not need to be simulated. In this case, the terms $\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right)$ jointly follow a multivariate normal distribution. However, since their marginal distributions are not independent, the transformations do not follow a chi-squared distribution but a mixture of chi-squared distributions, where the weights depend on the dependence structure of the $\Phi^{-1}\left(U_t^{i|\gamma_i^k}\right)$. For Z_t^{2*}, we obtain the following result:

Proposition 3. *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 nonzero 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*}, where γ_i^k for $k = 1, \ldots, 2^{d-1}$ denotes all subsets of the set $\{1, \ldots, i-1, i+1, \ldots, d\}$. A typical entry of R_{Z^*} is given by*

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

where the $\Sigma_{r,c}$ $(r,c \in \{i,\gamma_i^k\})$ are scalars, vectors, and matrices containing those elements of Σ 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.

The distribution of $Z_t^{2\dagger}$ in the Gaussian case is given by the following corollary:

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 $Z_1, \ldots, Z_d$ and $\lambda_1, \ldots, \lambda_d$ 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 $i = 1, \ldots, d$ entering $Z_t^{2\dagger}$. A typical entry of $R_{Z^\dagger}$ is given by*

[2]We show in the Supporting Information Appendix how parameter uncertainty is accounted for in the latter case at another step of the testing procedure.

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

where the index $-i$ denotes all rows/columns of Σ except for the ith one.

Note that, of course, $Z_t^2\sim\chi_d^2$ continues to hold under H_0 in the Gaussian case.

2.5 | Tests for proper calibration

In this section, we describe how we can construct tests of H_0 based on the transformations derived in the previous section. Depending on the formulation of the null hypothesis, this involves either testing that $U_t^W\sim\mathcal{U}(0,1)$ or jointly testing $U_t^W\sim\mathcal{U}(0,1)$ and independence across time. We delegate the discussion of the second case to the Supporting Information Appendix and focus here on the first case in which we use Neyman's (1937) smooth test.

2.5.1 | Known parameters and no dynamic misspecification

In this subsection we assume θ_0 is known and $\Omega_{t-1}=\mathfrak{F}_{t-1}$. The null hypothesis is $U_t^W\sim\mathcal{U}(0,1)$.[3] Many tests can be used in this context. We follow Bera and Ghosh (2002) and De Gooijer (2007), who advocate testing uniformity with Neyman's (1937) smooth test. Ko and Park (2019) also formally derive the smooth test in a context similar to ours and establish its asymptotic properties.

To understand Neyman's smooth test, consider the alternative family of smooth distributions:

$$s(u)=b_0\exp\left(\sum_{i=1}^{k}b_i\psi_i(u)\right),\quad u\in[0,1],\tag{11}$$

with b_0 a normalization constant and ψ_i the orthonormal Legendre polynomials. Testing uniformity (and hence H_0) against all distributions nested in Equation (11) boils down to testing $b_i=0$ for all $i=1,\dots,k$. Here we consider the first four Legendre polynomials, but in principle one could also determine the number of polynomials in a data-driven fashion as suggested by Ledwina (1994) and applied, for instance, by Lin and Wu (2017) and Ko and Park (2019).

A score test is easily computed as follows. Denoting the vector of (log-)scores of Equation (11) by $\xi_t=\left[\psi_1\left(U_t^W\right),\dots,\psi_4\left(U_t^W\right)\right]'$, it follows that under the null hypothesis $\frac{1}{\sqrt{P}}\sum_{t=R+1}^{n}\xi_t\xrightarrow{d}N(0,I_4)$, where I_4 is the 4×4 identity matrix. The Neyman smooth test statistic is then given by $\text{NST}=P^{-1}\left[\sum_{t=R+1}^{n}\xi_t\right]'\left[\sum_{t=R+1}^{n}\xi_t\right]$, which follows a χ_4^2 distribution under H_0. This result, however, only holds when the model parameters are known and when there is no dynamic misspecification. In the next subsection we describe the adjustments necessary for relaxing those assumptions.

2.5.2 | Estimated parameters and accounting for dynamic misspecification

Parameter uncertainty and dynamic misspecification are often relevant in practice when parametric forecast models are used and the data generating process (DGP) of the variables to be forecast (including the true values of the relevant parameters) is unknown to the forecaster. Ignoring both issues will, in general, lead to oversized tests in an out-of-sample evaluation. Thus we now assume estimates of the parameters, $\hat{\theta}$, are obtained using a $\sqrt{T}$-consistent estimator and $\Omega_{t-1}\subset\mathfrak{F}_{t-1}$.

Building on ideas in Chen (2011), we adjust Neyman's smooth test by relying on results in West (1996) and West and McCracken (1998) to derive suitable tests in the presence of parameter uncertainty *and* potential dynamic misspecification. A similar approach based on the ideas in Chen has been proposed by Lin and Wu (2017). Recall that we split our n observations into R in-sample observations, which we use to estimate the parameters, and P out-of-sample observations, which we use to evaluate the forecast model. Let $\hat{\xi}_t=\left[\psi_1\left(\hat{\mathrm{U}}_t^W\right),\dots,\psi_4\left(\hat{\mathrm{U}}_t^W\right)\right]'$ denote the Legendre polynomials in the *estimated* PITs of the (univariate) transformed series W_t, where $\hat{\mathrm{U}}_t^W$ has been computed using the in-sample parameter estimates.

[3]This approach can also be used if autocorrelation is not of concern and/or the tested densities are not model based (for instance, because they are obtained from a survey).

The following results rely on assumptions 1–5 in West and McCracken (1998). The moment conditions of interest are given by the vector ξ_t defined above and we consider the situation that the model is estimated by maximum likelihood estimation, thus satisfying their assumption 2 under stationarity. The conditional densities $f_{Y_t}(y|\Omega_{t-1},\theta)$ need to be twice continuously differentiable. Furthermore, the expected moment function $\mathbb{E}(\xi_t)$ must be continuously differentiable with respect to θ. Suitable mixing conditions in assumption 4 in West and McCracken ensure the applicability of suitable limit theorems.

It follows that under H_0 the elements of $\hat{\xi}_t$ are no longer independently distributed with unit variance as above, but

$$\frac{1}{\sqrt{P}}\sum_{t=R+1}^{n}\hat{\xi}_t \xrightarrow{d} N(0,\Sigma), \tag{12}$$

where (using the notation in (Chen, 2011)

$$\Sigma = S^* - \eta_1\left(D^*A^{-1}C' + CA^{-1}D^{*\prime}\right) + \eta_2\left(CA^{-1}B^*A^{-1}C'\right). \tag{13}$$

Given the score function $s_t = \frac{\partial}{\partial\theta_0}\ln f_t(y|\Omega_{t-1},\theta_0)$, the elements of Σ are given by $A = E\left(\frac{\partial}{\partial\theta_0}s_t\right)$, $B = E\left(s_ts_t'\right)$, $C = E\left(\frac{\partial}{\partial\theta_0}\xi_t\right)$, $D = E\left(\xi_ts_t'\right)$, $S^* = \sum_{k=-\infty}^{\infty}E\left(\xi_t\xi_{t-k}'\right)$, $B^* = \sum_{k=-\infty}^{\infty}E\left(s_ts_{t-k}'\right)$, and $D^* = \sum_{k=-\infty}^{\infty}E\left(\xi_ts_{t-k}'\right)$. The constants η_1 and η_2 are determined by the sampling scheme (fixed, rolling, or recursive) used to estimate the parameters and the limiting value of the ratio of in-sample and out-of-sample observations $\lambda = \lim_{n\to\infty}P/R$; see Chen (2011) for the precise formulas.

In order to avoid evaluation of the matrices A and C (the latter of which may be particularly tedious to obtain), we use the fact that the equalities $A+B=0$ and $C+D=0$ continue to hold even under dynamic misspecification, even though in this case they cannot be interpreted as (generalized) information matrix equalities; see White (1994). Thus we can rewrite Equation (13) as

$$\Sigma = S^* - \eta_1\left(D^*B^{-1}D' + DB^{-1}D^{*\prime}\right) + \eta_2\left(DB^{-1}B^*B^{-1}D'\right). \tag{14}$$

The matrices B and D can be estimated straightforwardly by their sample counterparts. In contrast, $S^*, B^*,$ and D^* need to be estimated by an appropriate estimator that is autocorrelation consistent. While, in principle, the widely used HAC estimator by Newey and West (1987) could be used, we found results in finite samples to be better (in terms of size) if we use the quadratic spectral estimator proposed by Andrews (1991). Neyman's smooth test statistic is then given by

$$\text{NST} = P^{-1}\left[\sum_{t=R+1}^{n}\hat{\xi}_t\right]'\hat{\Sigma}^{-1}\left[\sum_{t=R+1}^{n}\hat{\xi}_t\right], \tag{15}$$

which follows a χ^2_4 distribution under H_0. In addition, we can consider two intermediate cases. In the absence of dynamic misspecification, it holds that $B^* = B, D^* = D,$ and $S^* = I_4$; thus Equation (13) simplifies to $\Sigma = I_4 + (\eta_2 - 2\eta_1)DB^{-1}D'$. In the absence of parameter uncertainty, we obtain $\Sigma = S^*$.

3 | MONTE CARLO SIMULATIONS

We use Monte Carlo simulations to analyze how severe the size and power distortions caused by data mining can be in the case of the order-dependent approaches and how the size and power of the tests based on the different transformations compare. Here we consider the problem of testing the null hypothesis of a multivariate normal distribution for the baseline case of known parameters and no dynamic misspecification and when relaxing these two assumptions. In the Supporting Information Appendix we consider the following additional settings: (i) testing the null hypothesis of a multivariate t distribution; (ii) testing the null hypothesis of a multivariate generalized autoregressive conditional heteroskedasticity (GARCH) model; and (iii) jointly testing the null hypothesis of proper calibration and independence using the generalized autocontour approach of González-Rivera and Sun (2015).

3.1 | Simulation setup

Assume that the DGP under the null hypothesis is given by $y_t = \varepsilon_t$ with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Sigma)$. The $d\times d$ covariance matrix Σ is constructed such that all elements of y_t have unit variances $\left(\sigma_i^2 = 1 \text{ for } i = 1,\ldots,d\right)$ and the correlation between any

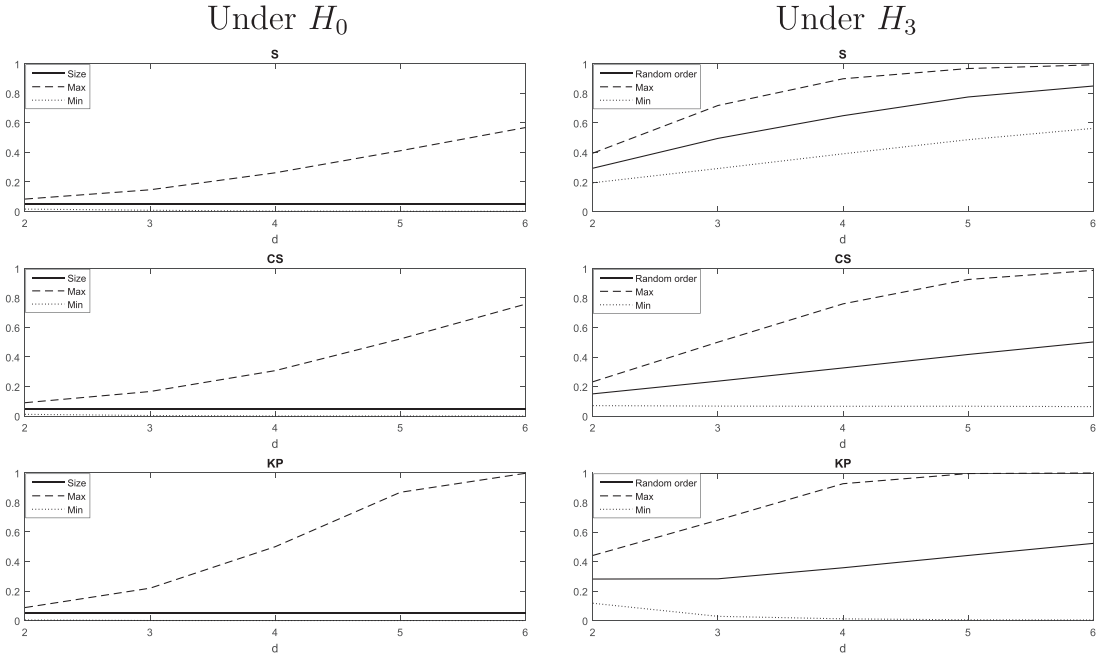

FIGURE 1 Scope for data mining: (a) under H_0; (b) under H_3. In the left-hand panel the solid lines show the nominal size of 5% and the dashed (dotted) lines show the rejection frequency that we obtain when we always choose that permutation for which we obtain the the highest (lowest) test statistic. In the right-hand panel the solid lines show the power that is obtained when the tests are applied properly and the dashed (dotted) lines show the rejection frequency that we obtain when we always choose that permutation for which we obtain the the highest (lowest) test statistic. The plots in the top panels refer to the stacked transformation by Diebold et al. (1999), those in the middle panels use the product transformation by Clements and Smith (2000), and those in the lower panels use the location-adjusted transformation Ko and Park (2013) [Colour figure can be viewed at wileyonlinelibrary.com]

two elements of y_t is equal to 0.5 $\left(\rho_{ij} = 0.5 \text{ for all } i \neq j\right)$. We consider different dimensions between $d = 2$ and $d = 50$ and (predictive) sample sizes of $P = \{50, 200\}$. Throughout the paper, we use 10,000 iterations for our Monte Carlo simulations. For the case of known parameters and no dynamics misspecification we consider four alternative DGPs that imply different deviations from H_0:

- *Alternative 1* (H_1): The innovations are generated from a multivariate normal distribution with $\sigma_i = 1.1$ and $\rho_{ij} = 0.5$.
- *Alternative 2* (H_2): The innovations are generated from a multivariate normal distribution with $\sigma_i = 1.0$ and $\rho_{ij} = 0.4$.
- *Alternative 3* (H_3): The innovations are generated from a multivariate t distribution with 8 degrees of freedom, with $\sigma_i = 1.0$ and $\rho_{ij} = 0.5$.
- *Alternative 4* (H_4): The innovations are generated from a multivariate Gaussian constant conditional correlation (CCC)-GARCH(1, 1) model, which we parametrize such that the unconditional covariance matrix is equal to that under H_0 and with GARCH parameters $(\omega, \alpha, \beta) = (0.05, 0.1, 0.85)$.

When allowing for estimated parameters and dynamic misspecification, we generate data by $y_t = 0.5y_{t-1} + \varepsilon_t$ with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$.[4] To simulate dynamic misspecification, we generate predictive densities ignoring the autocorrelation in the conditional mean. When the parameters are estimated, we consider only alternative 3 from the above list of alternatives. In addition, we consider a similar alternative with a more substantial deviation from H_0:

- *Alternative 5* (H_5): The innovations are generated from a multivariate t distribution with 4 degrees of freedom, with $\sigma_i = 1.0$ and $\rho_{ij} = 0.5$.

We also consider alternative 4, but control for the dynamic misspecification that is present in higher moments. Additionally, we consider the same alternative with dynamic misspecification in the mean:

[4]Results assuming a dynamic moving average structure, $y_t = 0.8\varepsilon_{t-1} + \varepsilon_t$, are very similar and not reported below.

- *Alternative 6* (H_6): The same as H_4 but with actual conditional mean dynamics given by the autoregressive homoskedastic model.

To estimate the model parameters, we consider a fixed estimation scheme and set $\lambda = 1/4$ to determine the size of the estimation sample $R = P/\lambda$.

3.2 | Potential for data mining

In this section we present results that show whether considering different permutations of the data can have a serious impact on the outcomes of the tests that are not order invariant. The idea is the following: A researcher who wants to discredit (support) the hypothesis that a particular model produces good density forecasts could, in principle, search across all permutations and select the one which yields the highest (lowest) test statistic. We present results for H_0 and H_2 based on $P = 100$; results are similar for other settings and are available upon request.

The left-hand part of Figure 1 shows how severe data mining can be under the null hypothesis. The solid line indicates the nominal size of 5%, which, as we show below, is obtained when tests are applied properly (meaning that the order of variables is chosen randomly). The other lines refer to the rejection frequencies that we obtain for the tests based on S, CS, and KP, respectively, when we always choose the permutation for which we obtain the highest (lowest) test statistic. At the lower end of obtainable rejection rates, it is clearly possible virtually never to reject the null hypothesis for any dimension. On the other hand, the null hypothesis can be rejected much too frequently if one chooses those permutations that yield high test statistics. For $d = 2$ the scope for data mining is rather limited, with obtainable rejection rates being around 10%. However, 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 KP.

TABLE 2 Size and power: Known parameters and no dynamic misspecification

	$n = 50$						$n = 200$					
	S	**CS**	**KP**	Z_t^2	Z_t^{2*}	$Z_t^{2\dagger}$	**S**	**CS**	**KP**	Z_t^2	Z_t^{2*}	$Z_t^{2\dagger}$
Size												
$d = 2$	0.047	0.051	0.047	0.051	0.050	0.052	0.053	0.051	0.050	0.051	0.052	0.053
$d = 4$	0.049	0.050	0.048	0.050	0.048	0.051	0.051	0.050	0.051	0.045	0.051	0.049
$d = 6$	0.047	0.051	0.049	0.048	0.047	0.048	0.052	0.049	0.049	0.052	0.053	0.051
$d = 10$	0.047	0.046	0.049	0.049	0.052	0.053	0.053	0.051	0.051	0.050	0.062	0.053
$d = 20$	0.053	0.048	0.051	0.053	—	0.059	0.053	0.051	0.053	0.050	—	0.054
$d = 50$	0.048	0.049	0.048	0.051	—	0.051	0.054	0.049	0.046	0.053	—	0.055
Power against H_1												
$d = 2$	0.197	0.137	0.139	0.198	**0.199**	0.164	0.556	0.338	0.358	**0.596**	0.583	0.484
$d = 4$	0.323	0.164	0.184	**0.338**	0.335	0.289	0.859	0.435	0.487	**0.893**	0.882	0.822
$d = 6$	0.449	0.193	0.219	**0.482**	0.475	0.434	0.961	0.527	0.600	**0.977**	0.971	0.954
$d = 20$	0.925	0.384	0.497	**0.953**	—	0.941	**1.000**	0.915	0.978	**1.000**	—	**1.000**
Power against H_2												
$d = 2$	0.066	0.046	0.100	0.067	0.072	**0.106**	0.100	0.063	**0.244**	0.106	0.105	0.235
$d = 4$	0.135	0.060	0.106	0.149	0.175	**0.238**	0.377	0.114	0.199	0.429	0.513	**0.691**
$d = 6$	0.225	0.075	0.138	0.252	0.324	**0.373**	0.706	0.187	0.327	0.762	0.856	**0.915**
Power against H_3												
$d = 2$	0.107	0.077	0.080	0.183	**0.188**	0.156	0.299	0.172	0.218	0.544	**0.545**	0.439
$d = 4$	0.177	0.091	0.125	**0.481**	0.472	0.391	0.563	0.247	0.413	**0.970**	**0.970**	0.925
$d = 6$	0.264	0.114	0.173	**0.747**	0.736	0.670	0.752	0.344	0.619	**1.000**	**1.000**	0.998
Power against H_4												
$d = 2$	0.290	0.204	0.252	0.319	**0.320**	0.285	0.413	0.314	0.376	**0.477**	**0.477**	0.423
$d = 4$	0.353	0.216	0.169	**0.406**	0.402	0.350	0.491	0.316	0.285	**0.623**	0.619	0.545
$d = 6$	0.388	0.222	0.194	**0.457**	0.443	0.402	0.547	0.335	0.304	**0.712**	0.699	0.641

Note. Rejection frequencies of Neyman's smooth test based on the transformations introduced in Sections 2.1 and 2.3 for the null hypothesis of multivariate normality with $\sigma_i = 1$ for $i = 1, \ldots, d$ and $\rho_{ij} = 0.5$ for all $i \neq j$. All Monte Carlo simulations are based on 10,000 iterations. The alternative models deviate from the null in terms of wrong variances (H_1), wrong correlations (H_2), fat tails (H_3), and GARCH effects (H_4). The exact hypotheses are defined in Section 3.1. For each set of simulations, the highest power is set in bold. A dash indicates that we did not compute Z^{2*} because it would take too much computing time.

In the right-hand part of Figure 1, the solid lines indicate the power that is obtained 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. The range of obtainable rejection rates is considerable in all cases. For tests based on CS and KP, the lower line is very close to 0. This means that, even though the data are generated from a different DGP, a researcher would be able to purposely select permutations in such a way that H_0 is almost never rejected.

3.3 | Size and power

When studying the size and power of the test we distinguish the baseline case of known parameters and no dynamic misspecification, and the situation when this is relaxed.

3.3.1 | Known parameters and no dynamic misspecification

Table 2 shows the Monte Carlo results concerning the size and power for the different transformations under the assumption of known parameters and no dynamic misspecification. Focusing on the upper panel of the table, we see that none of the approaches suffers from notable size distortions. In all cases, the obtained actual sizes are very close to the nominal size of 5%.

The second panel of the table reveals that tests based on our three new transformations and on S have the best power when the alternative implies deviations of the variances (H_1). Tests based on our new transformations outperform the test based on S for large dimensions. Results for H_2 show that the three new approaches consistently outperform the tests based on established transformations when deviations from H_0 are specified in terms of the correlation structure of the multivariate density.

TABLE 3 Size and power: Estimated parameters and dynamic misspecification

	$P = 50$						$P = 200$					
	S	**CS**	**KP**	Z^2	Z^{2*}	$Z^{2\dagger}$	**S**	**CS**	**KP**	Z^2	Z^{2*}	$Z^{2\dagger}$
Size (original test)												
$d = 2$	0.252	0.236	0.154	0.124	0.124	0.127	0.240	0.221	0.153	0.123	0.124	0.122
$d = 4$	0.270	0.229	0.101	0.153	0.152	0.174	0.260	0.235	0.086	0.132	0.132	0.136
$d = 6$	0.295	0.239	0.102	0.182	0.182	0.252	0.261	0.223	0.085	0.138	0.137	0.157
Size (adjusted test)												
$d = 2$	0.039	0.020	0.038	0.039	0.041	0.037	0.055	0.063	0.066	0.059	0.061	0.057
$d = 4$	0.026	0.015	0.024	0.025	0.029	0.028	0.048	0.054	0.053	0.056	0.059	0.055
$d = 6$	0.010	0.006	0.010	0.014	0.018	0.014	0.043	0.058	0.041	0.047	0.049	0.050
Power against H_3												
$d = 2$	**0.050**	0.019	0.043	0.022	0.024	0.022	0.121	0.096	0.122	0.129	**0.130**	0.106
$d = 4$	**0.038**	0.013	0.024	0.007	0.007	0.009	0.170	0.092	0.136	**0.360**	0.351	0.262
$d = 6$	**0.022**	0.005	0.009	0.003	0.003	0.003	0.223	0.095	0.166	**0.648**	0.629	0.538
Power against H_5												
$d = 2$	**0.074**	0.016	0.042	0.015	0.016	0.013	0.471	0.266	0.424	**0.654**	0.649	0.541
$d = 4$	**0.088**	0.010	0.023	0.007	0.007	0.006	0.758	0.320	0.613	**0.979**	0.978	0.942
$d = 6$	**0.077**	0.007	0.010	0.004	0.004	0.002	0.883	0.348	0.757	**0.999**	**0.999**	0.997
Power against H_4												
$d = 2$	**0.141**	0.075	0.092	0.081	0.098	0.075	**0.327**	0.267	0.336	0.315	0.321	0.280
$d = 4$	**0.169**	0.065	0.045	0.049	0.065	0.051	**0.392**	0.248	0.254	0.373	0.363	0.323
$d = 6$	**0.152**	0.045	0.029	0.030	0.042	0.036	**0.448**	0.241	0.258	0.427	0.412	0.370
Power against H_6												
$d = 2$	**0.069**	0.019	0.047	0.045	0.049	0.043	0.239	0.200	**0.295**	0.244	0.248	0.217
$d = 4$	**0.066**	0.011	0.025	0.020	0.020	0.018	**0.297**	0.161	0.243	0.285	0.279	0.221
$d = 6$	**0.044**	0.006	0.011	0.005	0.007	0.008	**0.347**	0.148	0.233	0.319	0.295	0.247

Note. Rejection frequencies of Neyman's smooth test based on the transformations introduced in Sections 3.1 and 3.3 for the null hypothesis of multivariate normality. All Monte Carlo simulations are based on 10,000 iterations. The data are generated by VAR(1) models with innovations following a multivariate normal distribution (H_0), multivariate t distributions with 8 degrees of freedom (H_3), and 4 degrees of freedom (H_5). H_4 and H_6 correspond to Gaussian GARCH(1, 1) models without and with dynamic misspecification in the mean equation, respectively. H_3 and H_4 are defined in Section 3.1. For each set of simulations, the highest power is set in bold.

Turning to the power of the different tests for detecting misspecification of the kurtosis (H_3), we see that the new approaches outperform all established tests by a wide margin. Especially for $P = 50$ the results are stunning: The power of the new approaches exceeds that of even the best-performing established approach by a factor of more than two in many cases. Finally, the new transformations also have better power against GARCH effects (H_4). In light of the good power against H_3, this is expected given that this alternative leads to innovations that are unconditionally distributed with excess kurtosis.

3.3.2 | Estimated parameters and accounting for dynamic misspecification

In general, the results indicate that tests based on all transformations are substantially oversized if one does not adjust Neyman's smooth test (Table 3). Using the adjusted version described above yields correctly sized tests for $P = 200$. Furthermore, the results indicate that having to deal with estimated parameters and dynamic misspecification results in a considerable loss of power against H_3. Large evaluation samples seem to be necessary to detect this deviation from the null hypothesis reasonably well (especially for low-dimensional densities). Power increases considerably for $P = 200$ in the case of H_5, which implies a much stronger deviation from H_0. At the same time, the ranking of the competing tests remains largely unaffected under both alternatives, so that the new tests proposed in this paper continue to perform substantially better than established tests (for moderate to large samples). Interestingly, when GARCH effects are present, the test based on S outperforms other tests both when dynamic misspecification is present (H_6) and when it is not (H_4).

4 | APPLICATIONS

4.1 | Predicting the distribution of stock market returns

In this section we provide an application showing that using tests that are not order invariant offers room for data mining in many situations. 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 USA, Japan (JA), the UK, Australia (AU), and Germany (GE), which we obtained from Datastream. The sample spans the period from January 1978 to December 2018, for a total of 2,191 weekly returns. We consider eight different time periods of 4 years, for which we evaluate density forecasts. These (out-of-sample) evaluation periods are 1987–1990, 1991–1994, 1995–1998, 1999–2002, 2003–2006, 2007–2010, 2011–2014, and 2015–2018. For each period, the previous 10 years are considered as in-sample data to estimate the models of interest. The models are reestimated for each week using a recursive scheme.

Three competing models of increasing complexity are considered: (i) a Gaussian dynamic conditional correlation (DCC)-GARCH model (Engle, 2002); (ii) a time-varying t-copula with DCC-type dynamics and t-GARCH margins; and (iii) a time-varying t-copula with skewed-t-GJR-GARCH margins.

For the second model, the marginal models are the same as for the DCC, with the difference that the innovations follow a t distribution with ν_i degrees of freedom. The dependence between the t-distributed GARCH innovations is given by a t-copula with degrees of freedom ν_c and correlation matrix R_t. The evolution of the correlation matrix is the same as in the DCC model, but with the innovations transformed to have a t distribution with ν_c degrees of freedom. 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. 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.

The third model is even more flexible by assuming that the GARCH innovations follow the skewed t distribution of Hansen (1994) and by relying on the GJR-GARCH model of Glosten, Jagannathan, and Runkle (1993). The dependence is again given by the DCC-t-copula model.

For each model and each time period, we compute the Rosenblatt PITs and apply the established and new transformations described above. Recall that for non-Gaussian models the distribution of Z_t^{2*} and $Z_t^{2\dagger}$ is not known but can be computed numerically, as explained in Section 2.4.1. The null hypothesis of correctly predicted densities is then tested with Neyman's smooth test, accounting for parameter estimation and potential dynamic misspecification. We estimate the required long-run covariance matrices using a quadratic spectral kernel and automatic lag selection as proposed in Andrews (1991). For those tests that 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: USA, JP, UK, AU, GE) and, in brackets, the lowest and highest p-values across all permutations (Table 4).

The potential for data mining using the tests based on S, CS, and KP, respectively, is immense. For the majority of periods one can find permutations that reject and permutations that do not reject the null hypothesis of properly calibrated density

TABLE 4 Density forecast evaluation for stock market returns

	S	CS	KP	Z^2	Z^{2*}	$Z^{2\dagger}$
Gaussian DCC						
1987–1990	0.005 [0.000, 0.173]	0.016 [0.005, 0.365]	0.012 [0.000, 0.990]	0.034	0.003	0.004
1991–1994	0.000 [0.000, 0.000]	0.002 [0.000, 0.018]	0.000 [0.000, 0.002]	0.000	0.000	0.000
1995–1998	0.000 [0.000, 0.000]	0.001 [0.000, 0.030]	0.001 [0.000, 0.083]	0.001	0.000	0.001
1999–2002	0.003 [0.000, 0.028]	0.031 [0.010, 0.486]	0.016 [0.001, 0.570]	0.002	0.002	0.004
2003–2006	0.000 [0.000, 0.000]	0.001 [0.000, 0.002]	0.000 [0.000, 0.298]	0.000	0.001	0.000
2007–2010	0.002 [0.000, 0.049]	0.130 [0.004, 0.441]	0.004 [0.000, 0.447]	0.001	0.001	0.001
2011–2014	0.006 [0.000, 0.401]	0.407 [0.003, 0.796]	0.054 [0.001, 0.932]	0.004	0.039	0.128
2015–2018	0.062 [0.002, 0.316]	0.089 [0.000, 0.170]	0.874 [0.000, 0.961]	0.001	0.000	0.036
t-GARCH-tDCC-Cop						
1987–1990	0.083 [0.004, 0.282]	0.137 [0.091, 0.601]	0.655 [0.017, 0.988]	0.688 [0.486, 0.699]	0.7759	0.9627
1991–1994	0.008 [0.001, 0.095]	0.122 [0.009, 0.373]	0.000 [0.000, 0.250]	0.009 [0.006, 0.012]	0.0082	0.04
1995–1998	0.036 [0.000, 0.136]	0.005 [0.001, 0.122]	0.133 [0.004, 0.940]	0.017 [0.017, 0.055]	0.0117	0.0141
1999–2002	0.094 [0.012, 0.397]	0.045 [0.015, 0.544]	0.238 [0.011, 0.866]	0.056 [0.034, 0.079]	0.0582	0.0124
2003–2006	0.000 [0.000, 0.000]	0.003 [0.000, 0.007]	0.001 [0.000, 0.661]	0.001 [0.001, 0.001]	0.0005	0.001
2007–2010	0.299 [0.016, 0.478]	0.489 [0.004, 0.880]	0.098 [0.001, 0.976]	0.167 [0.100, 0.272]	0.0834	0.0606
2011–2014	0.154 [0.001, 0.817]	0.281 [0.005, 0.999]	0.390 [0.020, 0.993]	0.873 [0.535, 0.873]	0.8799	0.2113
2015–2018	0.437 [0.097, 0.940]	0.263 [0.019, 0.970]	0.726 [0.063, 0.991]	0.625 [0.489, 0.646]	0.4003	0.0708
st-GJR-tDCC-Cop						
1987–1990	0.000 [0.000, 0.002]	0.000 [0.000, 0.002]	0.003 [0.000, 0.409]	0.010 [0.005, 0.035]	0.0036	0.0003
1991–1994	0.000 [0.000, 0.014]	0.023 [0.000, 0.124]	0.000 [0.000, 0.133]	0.012 [0.006, 0.013]	0.0091	0.0534
1995–1998	0.027 [0.000, 0.275]	0.008 [0.001, 0.155]	0.169 [0.007, 0.966]	0.042 [0.041, 0.111]	0.0374	0.0296
1999–2002	0.016 [0.003, 0.091]	0.155 [0.017, 0.764]	0.031 [0.001, 0.543]	0.004 [0.004, 0.007]	0.0058	0.0044
2003–2006	0.000 [0.000, 0.000]	0.002 [0.000, 0.006]	0.004 [0.000, 0.431]	0.003 [0.002, 0.004]	0.0025	0.0025
2007–2010	0.092 [0.004, 0.529]	0.908 [0.087, 0.990]	0.282 [0.001, 0.989]	0.201 [0.181, 0.279]	0.1211	0.0816
2011–2014	0.862 [0.009, 0.961]	0.655 [0.022, 0.992]	0.645 [0.009, 0.986]	0.606 [0.583, 0.790]	0.555	0.2169
2015–2018	0.472 [0.021, 0.724]	0.775 [0.000, 0.956]	0.678 [0.007, 0.993]	0.267 [0.170, 0.410]	0.3771	0.196

Note. The table shows *p*-values corresponding to the different transformations introduced in Sections 2.1 and 2.3 using the adjusted version of Neyman's smooth test (Neyman, 1937), which accounts for parameter estimation and potential dynamic misspecification. The data are weekly MSCI stock index returns for the USA, Japan, the 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 that are not order invariant, the entries 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.

forecasts for any of the models. Note, however, that, in line with our results from Section 3.2, the range of *p*-values for tests based on S is, on average, smaller than for those based on CS and KP. Finally, 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 only moderate scope for data mining based on this transformation. In summary, we recommend evaluating the density forecasts based on Z^{2*} and $Z^{2\dagger}$, and possibly based on Z^2. The results based on the other transformations 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.

The results concerning the performance of the models are mixed and depend on the time period under study. However, a few things clearly stand out. First, the Gaussian DCC model is rejected by almost all tests for all time periods except the 2011–2014 period. Second, model specifications (ii) and (iii) perform much better, but are still rejected for some periods. Third, the most flexible specification (iii) does not consistently outperform specification (ii), confirming the known fact that model complexity may yield superior in-sample fit, but not necessarily a better forecasting performance. Using a 1% significance level, specifications (ii) and (iii) are rejected by the test based on Z^{2*} ($Z^{2\dagger}$) for 2 (1) and 4 (3) subsamples, respectively. When using a Bonferroni correction to address the fact that this is a case of multiple testing, specification (ii) is only rejected for the 2003–2006 period (based on both Z^{2*} and $Z^{2\dagger}$).[5] Thus, overall, the *t*-GARCH model with a time-varying *t*-copula can be recommended for modeling and predicting the joint density of weekly stock market returns.

[5]Since we apply the tests to eight different subsamples, a test at the 5% significance level should reject when the *p*-value is smaller than $0.05/8 = 0.0063$.

4.2 | Evaluating macroeconomic density forecasts

In a second application, we demonstrate how the newly developed tests can be applied in the area of macroeconomic forecasting. 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 gross domestic product price index (Δp_t), and the yield of 3-month Treasury bills (i_t). The data are downloaded from FRED and cover the sample from 1953:Q1 to 2018:Q4. We use the original specification with two lags and assuming a very flexible processes that governs 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. We estimate the model using Bayesian methods and follow the specification of priors as in Primiceri.[6]

We use a recursive scheme to generate density forecasts, $\hat{f}_{y_{t+h}}(y_{t+h}|\mathcal{F}_t)$, with forecast horizons $h = 1, \dots, 4$. The period between 1982:Q4+h and 2017:Q4+h is used as the evaluation sample. Thus we start by estimating the model using data until 1982:Q4 and constructing density forecasts for 1983:Q1, 1983:Q2, 1983:Q3, and 1983:Q4. Subsequently, we recursively add one observation to our estimation sample and shift the forecast period one quarter forward. In total, a sequence of 141 density forecasts is available for each forecast horizon.

The form of $\hat{f}_{y_{t+h}}(y_{t+h}|\mathcal{F}_t)$ is unknown. For $h = 1$ it follows a multivariate normal distribution conditional on the parameters of the model but not unconditionally. For $h > 1$, further deviations from a Gaussian distribution arise due to the fact that the conditional forecasts are nonlinear 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, which we obtain by keeping every 10th 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. 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.[7] Since the distributions of the random variables Z_t^{2*} and $Z_t^{2\dagger}$ are not known, we simulate their distribution by repeatedly computing Z_t^{2*} and $Z_t^{2\dagger}$ under H_0 in order to compute the corresponding PITs.

For comparison, we also check whether results differ if we use an approximation and assume that all conditional forecasts follow a multivariate normal distribution. 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 $\Sigma_{t+h} = (1/B)\sum_{b=1}^{B}\left(\hat{y}_{t+h}^{(b)} - \bar{y}_{t+h}\right)\left(\hat{y}_{t+h}^{(b)} - \bar{y}_{t+h}\right)'$.

Since for $h > 1$ the PITs will be subject to autocorrelation even under H_0 and without dynamic misspecification, we report results using the robust version of the Neyman smooth test. Test results for $h = 1$ and $h = 4$ are summarized in Table 5.

We first focus on the nonparametric predictive densities and the corresponding results in the right-hand panel. 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. 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 data mining can be a very serious problem in practice even for low-dimensional models.[8] In general, however, the null hypothesis is not rejected for most cases. Only the tests based on CS indicate for $h = 4$ and most permutations that the null hypothesis of proper calibration should be rejected.

The results in the left-hand panel correspond to the case of a normal approximation of the predictive density. They provide strong evidence against the null hypothesis of well-calibrated predictive densities. All order-invariant

[6]We use the corrected algorithm (Del Negro & Primiceri, 2015), which implies a different ordering of the Markov chain Monte Carlo steps. We use the "bvarsv" package for R to estimate the model.

[7]Since the data-driven determination of optimal bandwidths is computationally demanding, we do so only for every 12th period and keep the bandwidths fixed for all intermediate periods. We rely on least-squares cross-validation (Li, Lin, & Racine, 2013) to reoptimize the bandwidths. All nonparametric estimations are executed using the "np" package for R.

[8]The average (across all forecast horizons and transformations) standard deviation of the p-values is 0.12.

TABLE 5 Tests for proper calibration of macroeconomic forecasts

	Normal approximation						Nonparametric densities					
	S	**P**	**P***	Z^2	Z^{2*}	$Z^{2\dagger}$	**S**	**P**	**P***	Z^2	Z^{2*}	$Z^{2\dagger}$
$h = 1$												
$u_t - \Delta p_t - i_t$	0.036	0.109	0.235	0.000	0.000	0.001	0.455	0.581	0.031	0.008	0.224	0.162
$u_t - i_t - \Delta p_t$	0.030	0.053	0.248				0.370	0.148	0.700	0.067		
$\Delta p_t - u_t - i_t$	0.033	0.097	0.073				0.483	0.574	0.008	0.005		
$\Delta p_t - i_t - u_t$	0.005	0.117	0.003				0.312	0.073	0.086	0.052		
$i_t - u_t - \Delta p_t$	0.003	0.091	0.036				0.291	0.139	0.401	0.070		
$i_t - \Delta p_t - u_t$	0.007	0.064	0.013				0.261	0.379	0.191	0.077		
$h = 4$												
$u_t - \Delta p_t - i_t$	0.046	0.025	0.457	0.014	0.015	0.083	0.356	0.001	0.820	0.433	0.744	0.686
$u_t - i_t - \Delta p_t$	0.158	0.225	0.212				0.248	0.179	0.024	0.631		
$\Delta p_t - u_t - i_t$	0.049	0.030	0.816				0.239	0.004	0.553	0.396		
$\Delta p_t - i_t - u_t$	0.487	0.380	0.866				0.394	0.000	0.566	0.418		
$i_t - u_t - \Delta p_t$	0.416	0.002	0.202				0.436	0.000	0.807	0.462		
$i_t - \Delta p_t - u_t$	0.445	0.002	0.329				0.483	0.022	0.684	0.576		

Note. The table shows *p*-values corresponding to tests based on different transformations for all possible permutations of the data. Results account for potential autocorrelation in the PITs. For those transformations that yield order-invariant test statistics, we only report one *p*-value.

transformations (with the exception of $Z^{2\dagger}$ for $h = 4$) yield *p*-values very close to 0 for both forecast horizons. For $h = 1$, the tests based on the other transformations mostly reject the null hypothesis for the majority of permutations. For $h = 4$, the CS-based tests tend to reject H_0, whereas tests based on S and KP do not, for the majority of permutations.

In general, we conclude that (i) data mining can be a very relevant issue in practice even in macroeconomic applications with a relative low number of variables, (ii) the VAR model with time-varying parameters proposed by Primiceri (2005) is able to generate well-calibrated multivariate density forecasts for the US economy, and (iii) the latter result holds true only for properly estimated predictive densities but not when using a Gaussian approximation.

5 | CONCLUSION

In this paper we derive order-invariant tests for proper calibration of multivariate density forecasts of arbitrary dimension. We demonstrate that distortions in rejection rates can be very large when established tests, which are not order invariant, are used for data mining. Furthermore, we show that the new tests have very good power against a wide range of deviations from the null hypothesis. We do not find that one of our new tests dominates the others in terms of power regardless of the alternatives. Therefore, we recommend using simultaneously the tests based on Z^{2*} and $Z^{2\dagger}$ whenever there is no strong prior about the nature of potential deviations from the specified null model.

We believe there is a wide range of other applications in various fields. First, the proposed methods are useful whenever properly calibrated density forecasts are crucial to form well-informed decisions (about production, investment, pricing, etc.) and will foster the use of multivariate density forecasts in situations in which decisions are based on loss functions that take more than one variable as input arguments. Our tests could, for instance, be used to assess the overall forecast performance of macroeconomic dynamic stochastic general equilibrium models used in central banks. Second, the proposed methods are useful to improve the specification of multivariate models taking higher moments into account; obvious applications of this kind are common in financial econometrics—for example, 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 will become more common also for demand management or in macroeconomics.

Our study leaves room for future research. First, especially for financial applications, it would be interesting to extend the analytical results of our paper, which are limited to the case of multivariate Gaussian processes under the null hypothesis, to more general settings. Second, it would be interesting to analyze why a multivariate predictive density is not properly calibrated by looking at the distribution of individual conditional PITs. Finally, it might be worthwhile to investigate whether powerful order-invariant tests can be constructed that are not based on the Rosenblatt transformation.

ACKNOWLEDGMENTS

A previous draft of this paper has been circulated under the title "Order invariant evaluation of multivariate density forecasts." We would like to thank Malte Knüppel, Fabian Krüger, Maik Wolters, and the participants of seminars at RWTH Aachen University, University of Cologne, University of Graz, Heidelberg University, University of Kiel, Maastricht University, University of Duisburg-Essen, and the University of Warwick for their helpful comments on earlier drafts of the paper. We also appreciate valuable comments on an earlier draft of this paper by two anonymous referees and the Editor Michael W. McCracken.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/datasets/dovern001/].

REFERENCES

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

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

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

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

Bera, A. K., & Ghosh, A. (2002). Neyman's smooth test and its applications in econometrics. In A. Ullah, A. T. K. Wan, & A. Chaturvedi (Eds.), *Handbook of applied econometrics and statistical inference*. New York, NY: Marcel Dekker, pp. 177–230.

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

Chen, Y. T. (2011). Moment tests for density forecast evaluation in the presence of parameter estimation uncertainty. *Journal of Forecasting*, *30*, 409–450.

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

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

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

Corradi, V., & Swanson, N. R. (2006b). Handbook of economic forecasting, *Predictive density evaluation*, Chapter 5, Vol. *1*. Amsterdam, Netherlands: Elsevier, pp. 197–284.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, *147*, 278–292.

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

Del Negro, M., & Primiceri, G. E. (2015). Time varying structural vector autoregressions and monetary policy: a corrigendum. *Review of Economic Studies*, *82*, 1342–1345.

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

Diebold, F. X., Hahn, J., & Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange. *Review of Economics and Statistics*, *81*, 661–673.

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

Glosten, L. R., Jagannathan, R., & 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*, 1779–1801.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Applications*, *1*, 125–151.

González-Rivera, G., & Sun, Y. (2015). Generalized autocontours: Evaluation of multivariate density models. *International Journal of Forecasting*, *31*, 799–814.

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

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

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

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

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

Ishida, I. (2005). Scanning multivariate conditional densities with probability integral transforms. (*CARF F-Series 045*). Tokyo, Japan: CARF, University of Tokyo.

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

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

Ko, S. I. M., & Park, S. Y. (2019). Multivariate density forecast evaluation: Smooth test approach. Hong Kong, China: Chinese University of Hong Kong. (*Working paper*).

Ledwina, T. (1994). Data driven version of the Neyman smooth test of fit. *Journal of the American Statistical Association*, *89*, 1000–1005.

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

Lin, J., & Wu, X. (2017). A sequential test for the specification of predictive densities. *Econometrics Journal*, *20*, 190–220.

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

Newey, W. K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica*, *53*, 1047–1070.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

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*, 821–852.

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

Rossi, B., & Sekhposyan, T. (2013). Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics*, *177*, 199–212.

Rossi, B., & Sekhposyan, T. (2019). Alternative tests for correct specification of conditional predictive densities. *Journal of Econometrics*, *208*, 638–657.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, *4*, 283–291.

Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, *30*, 415–443.

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

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*, 1067–1087.

West, K., & McCracken, M. (1998). Regression based tests of predictive ability. *International Economic Review*, *39*, 817–840.

White, H. (1994). *Estimation, inference and specification analysis*. Cambridge, UK: Cambridge University Press.

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Dovern J, Manner H. Order-invariant tests for proper calibration of multivariate density forecasts. *J Appl Econ*. 2020;35:440–456. https://doi.org/10.1002/jae.2755