International Journal of Forecasting 39 (2023) 1078–1096

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Static and dynamic models for multivariate distribution forecasts: Proper scoring rule tests of factor-quantile versus multivariate GARCH models

Carol Alexander [a,b,*], Yang Han [a], Xiaochun Meng [a]

[a] University of Sussex, United Kingdom
[b] PHBS Business School, Peking University, China

ARTICLE INFO

Keywords:
Bagging
Continuous ranked probability score
Energy score
Factor quantile regression
Historical simulation
Multivariate density
Forecast
Variogram score

ABSTRACT

Many static and dynamic models exist to forecast Value-at-Risk and other quantile-related metrics used in financial risk management. Industry practice favours simpler, static models such as historical simulation or its variants. Most academic research focuses on dynamic models in the GARCH family. While numerous studies examine the accuracy of multivariate models for forecasting risk metrics, there is little research on accurately predicting the entire multivariate distribution. However, this is an essential element of asset pricing or portfolio optimization problems having non-analytic solutions. We approach this highly complex problem using various proper multivariate scoring rules to evaluate forecasts of eight-dimensional multivariate distributions: exchange rates, interest rates and commodity futures. This way, we test the performance of static models, namely, empirical distribution functions and a new factor-quantile model with commonly used dynamic models in the asymmetric multivariate GARCH class.

1. Introduction

Many static and dynamic models are employed to forecast Value-at-Risk (VaR) and other quantile-related metrics used in financial risk management. Many surveys of the literature have been published during the last twenty years, most recently by Nieto and Ruiz (2016), who provide a comprehensive review of the main methodological and empirical developments in (univariate) VaR models and their backtesting. Developing tractable models for forecasting the entire distribution has attracted little academic attention, especially in a multivariate setting. However, the problem is important because accurate distribution forecasting is fundamental for the success of two important types of financial problems. These are asset pricing, including the valuation of derivative products[1] and the optimization of portfolio allocations when the decision-maker utility function and/or the distribution of the asset returns preclude the existence of an analytic solution.[2]

A very popular topic for academic research is testing the accuracy of quantile forecasts from the Bollerslev

* Corresponding author at: University of Sussex, United Kingdom.
E-mail address: c.alexander@sussex.ac.uk (C. Alexander).

[1] See, for instance, Semenov (2008), Chiang and Tsai (2019) and Zhou et al. (2019).

[2] In portfolio optimization a forecast of the entire multivariate distribution for asset returns is required to calculate the investor's expected utility, see Ebens et al. (2009), Lwin et al. (2017), Thomann (2021) and Grant and Satchell (2020). There are many other studies, but we have selected these to point out that asset managers commonly employ static models based on historical simulation because they are much simpler than dynamic models. Also, the academic literature in this area tends to focus more on modelling the decision-maker utility than on the underlying multivariate returns process. See Birge (2007) and Resta (2012) for reviews.

https://doi.org/10.1016/j.ijforecast.2022.04.004

(1986) Generalised Autoregressive Conditional Heteroscedasticity (GARCH) model and its numerous variants (see, for example, Kuester et al. 2006 and Orhan and Köksal 2012). Yet, there is little evidence for the industry's widespread adoption of such models, particularly in a large-scale multivariate setting. Indeed, when commercial banks and other financial institutions report market risks, they tend to favour simple static models such as historical simulation.[3] The popularity of this approach to one-step-ahead forecasts is also supported by its robustness.[4] More recently Danielsson et al. (2016) show that, for predicting quantiles, it remains unclear whether a complex dynamic model in the GARCH class outperforms one that is based on the idea that the next period joint distribution of the variables can be well approximated by their joint historical distribution, as in Semenov (2008). Nevertheless, most academic research on the forecasting accuracy of quantile-based risk metrics centres on dynamic models in the GARCH family.

This paper examines the accuracy of simple, static models that are typically favoured by financial institutions for predicting, not just quantile, but an entire multivariate distribution. Can static models produce more accurate forecasts than complex dynamic models that are receiving the most attention in academic literature? To answer this question, we consider a semi-parametric extension of historical simulation that generates a multivariate distribution using a parametric copula with empirical distribution function (EDF) marginals (Patton, 2009). In addition, a substantial methodological section of this paper introduces a semi-parametric model for estimating multivariate distributions where marginals are derived from factor model quantile regressions (Koenker & Bassett, 1982) and the dependence structure is modelled using a conditional copula (Patton, 2006). We call it the *Factor Quantile* (FQ) model. These static models are relatively easy for less-quantitative managers to comprehend, they scale naturally to very large dimensions, and calibration is fast. These static models would be a natural candidate for adoption by the industry because they are simpler, quicker and more robust than multivariate GARCH models. We need to show that they produce forecasts that are at least as accurate as the most common multivariate GARCH models.

To this end, we report a very comprehensive study that is the first extensive empirical evaluation of forecasting accuracy using the model confidence set approach of Hansen et al. (2011) based on several proper multivariate scoring rules. Previously developed in meteorology and other branches of atmospheric science (Jolliffe & Stephenson, 2003; Keune et al., 2014), we apply these rules to assess the accuracy of daily forecasts for three different financial systems: exchange rates, interest rates and commodity futures. First, we test univariate distribution forecasts using the weighted conditional ranked probability score proposed by Gneiting and Ranjan (2011), which has the advantage of allowing different weight functions to focus on specific parts of the distribution. Then we apply the energy and variogram scores to measure the accuracy of multivariate distribution forecasts, see Gneiting et al. (2008) and Scheuerer and Hamill (2015). In each case, we assess the relative accuracy of the entire set of distribution forecasts considered in our empirical study through the equivalence test and elimination rules of the model confidence set of Hansen et al. (2011).

This way, we compare the forecasting performance of EDF models with two latent factor versions of the FQ model and with popular multivariate GARCH models, including the DCC-GARCH model of Engle (2002) with the exponential GARCH conditional variance specification of Nelson (1991) and Student-t innovations. We use the Gaussian copula to reduce complexity and so that the EDF and FQ models scale naturally and easily to higher dimensions. However, the GARCH models are more challenging to scale, and for this reason, we consider eight-dimensional multivariate distributions, not higher dimensions. Our empirical study examines daily forecasts for eight USD exchange rates using data from 1999–2018; a term structure of US interest rates with data from 1994–2018; and eight Bloomberg investable commodity indices from 1991–2018.

The following Section 2 sets out our work in the context of the recent literature on multivariate distribution forecasts and multivariate quantile models. Section 3 introduces the general FQ methodology, illustrating this in a simple bivariate example. We provide the motivation for various choices, such as the quantile grid and introduce two variants of the model with latent factors. Section 4 presents our empirical study. Finally, Section 5 summarises and concludes the paper. All the code (in Python/MATLAB) and all three data sets used in this paper are available from the authors on request.

2. Relevant literature

Forecasts of random variables should take the form of distributions to account for the randomness of the predicted event and prediction uncertainty. Yet most prior research in finance limits empirical evaluation to point forecasts, often linked to quantiles and typically of a univariate distribution. As a result, extensive empirical applications, even to univariate distribution forecasts, are hard to find. This is possibly due to their computational complexity. Some applications of multivariate scoring rules can be found in the literature. These are mostly limited to weather ensemble forecasts, or they use the multivariate logarithmic score (Diks et al., 2014, 2010) that has been criticised for its heavy penalty on low probability events which limits its practical application (Gneiting & Raftery, 2007; Selten, 1998). There are a few recent applications of proper scoring rules to financial or economic data. These studies are far more limited than ours, and they are limited to univariate distributions over a single out-of-sample period, see Panagiotelis and Smith

[3] See Pristsker (2006), Berkowitz et al. (2011), Prorokowski and Prorokowski (2014), Scheller and Auer (2018) and many others. A survey by Pérignon and Smith (2010) reported that almost 75% of banks forecasted VaR using historical simulation in their sample.

[4] Cont et al. (2010) introduce a rigorous framework for studying this feature, showing that historical VaR is more robust than sophisticated risk metrics based on parametric models estimated by maximum likelihood.

(2008), Ravazzolo and Vahey (2014) and Alexander et al. (2019).

Elliott and Timmermann (2016) emphasise that a forecast is an economic decision that should be evaluated using a loss function. They survey some elementary tests for univariate forecasts based on loss differentials. These include Diebold and Mariano (1995) who develop out-of-sample tests that compare errors of point forecasts, and Giacomini and White (2006) who extend these tests to multi-step point, interval or entire (univariate) distribution forecasts. Since then, several other papers have investigated scoring rules applied to univariate forecasts. Bao et al. (2007) advocate using the Kullback–Leibler information criterion, which is derived from the logarithmic score. Amisano and Giacomini (2007) compare density forecasts using a weighted likelihood ratio test, but this is not a proper scoring rule. Gneiting and Raftery (2007) advocate using the Continuous Ranked Probability Score (CRPS). Gneiting and Ranjan (2011) extend the CRPS to adopt the weighting approach of Amisano and Giacomini (2007) so that evaluation can be focused on a specific area of the distribution, such as the tail or centre. Boero et al. (2011) find that ranked probability scores have better discriminatory power than logarithmic or quadratic scores.

Concerning the literature relevant to our proposed factor-quantile model, many empirical studies apply the quantile regression model of Koenker et al. (1978) to predict financial data. However, most examine the accuracy of a few specific (typically extreme) quantiles and not the entire distribution. For instance, Ma and Pohlman (2008) and Meligkotsidou et al. (2019) examine the predictability of stock returns and realized volatilities by lagged economic variables, and Hua and Manzan (2013) use realized volatilities to predict quantiles of stock and bond returns comparing their quantile-regression results with four different univariate versions of GARCH. Gaglianone and Lima (2012) apply quantile regression to predict the distribution of U.S. unemployment rates, using a single-factor model with an exogenous consensus forecast based on forecast averaging. Similarly, Bunn et al. (2016) use quantile factor models with exogenous forecasts of factors to predict the spot electricity price. Cenesizoglu and Timmermann (2008) use a model with a single lagged predictor variable to forecast quantiles and estimate the distribution by fitting a crude step function introduced by Koenker and Bassett (1982). Other papers, such as Hagfors et al. (2016) apply quantile regressions with multiple factors but use only in-sample diagnostics to examine the model fit. Some papers use very short time series, e.g., Koenker and Bassett (2010), and/or compare quantile regression with the benchmark, models which may be inadequate for the data.

Concerning the various attempts to model multivariate quantiles, Chakraborty (2003) proposes to minimize a loss function that is a straightforward multivariate equivalent of the standard loss function used in univariate quantile regression, introduced by Koenker et al. (1978). However, this doesn't allow the estimation of an associated distribution function because it is only based on the notion of geometric multivariate quantiles. Similarly, Hallin et al. (2010) use the half-space depth contours of Tukey (1974) which are not equivalent to an associated distribution function. By contrast, insisting on the equivalence between the quantile function and a well-defined multivariate distribution, Chavas (2018) proposes that a multivariate q-quantile is a set $\mathbf{c}$ corresponding to the q-contour of the multivariate distribution F, i.e., $F(\mathbf{c}) = q$. This must reflect the general properties of q-quantiles, e.g., $F(\mathbf{c})$ is always non-decreasing. However, the q-contours need not be convex, so F need not have a unique inverse. Chavas (2018) assumes that quantiles are linear functions of exogenous variables. He only derives statistical properties of the quantile estimator when conditional distributions of the endogenous variables are independent.

Several recent papers also examine new ways to predict the financial variables we consider, albeit only by point forecasts. For the US interest rate term structure see Almeida et al. (2017); for USD exchange rates see Greenaway-McGrevy et al. (2018); and for commodity futures see Zolotko and Okhrin (2014), amongst many others. Finally, the voluminous literature on multivariate GARCH forecasting in financial markets is summarised by Silvennoinen and Terasvirta (2009) and Zakamulin (2015). Most studies only consider in-sample specification tests, except for Laurent et al. (2012), who apply the model confidence set of Hansen et al. (2011) and the Hansen (2005) tests for the superior predictive ability to GARCH covariance forecasts, not to multivariate returns distributions.

3. Factor quantile models

Deriving multivariate distribution forecasts from a system of common factor quantile regressions presents many challenges. The basic problem is that there is no unique way to invert a multivariate distribution function and no inherent ordering of quantiles in multiple dimensions. So, unlike the univariate case, even the definition of a multivariate quantile is not unique. Alternative definitions support different techniques for estimating multivariate quantile regressions, not all of which identify distribution functions. The motivation for FQ models is to circumvent these problems entirely, deriving a multivariate distribution by applying a conditional copula to marginals generated from univariate factor model quantile regressions.

The fundamental steps are easily understood in three stages. (i) For each dependent variable, we estimate conditional prediction quantiles at a range of levels in (0, 1) using univariate quantile regression on multiple common factors. (ii) For a given realisation of common factors and each dependent variable, fit a conditional distribution to the quantiles estimated in (i). (iii) Impose a dependence structure on the conditional marginals using a conditional copula.

This way, the FQ model generates a multivariate distribution, conditional on the common factors, with marginals derived from quantile regressions and a flexible dependence structure imposed by the choice of the copula. This algorithm is very fast and flexible, and because the quantile regressions are univariate, it scales very easily as the

dimension of the system increases. By comparison, multivariate quantile regression approaches, such as those proposed by Chakraborty (2003) or Chavas (2018), require a vast data set and are much more computationally intensive.

The starting point of our model description is a standard linear factor model

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \qquad t = 1, \ldots, T,$$

where: $\mathbf{y}_t = (\mathrm{y}_{1t}, \ldots, \mathrm{y}_{nt})'$ and $\mathbf{x}_t = (\mathrm{x}_{1t}, \ldots, \mathrm{x}_{mt})'$ denote the time t values of n dependent variables and m common factors, respectively; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ is the vector of intercepts, and $\mathbf{B}$ is the matrix of factor sensitivities, both assumed constant; and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})'$ is a multivariate error process. Further, we assume that the data $\{\mathbf{y}_t\}_{t=1}^T$ are generated by a stochastic process $\mathbf{y}$ with stationary conditional joint distribution $F|\mathbf{x}$ and conditional marginal distributions $F_1|\mathbf{x}, \ldots, F_n|\mathbf{x}$.

Macroeconomic, fundamental and statistical factor models were introduced by Ross (1976), Fama and French (1993) and Connor et al. (2012) respectively. Applications to predicting stock portfolios, interest rates, exchange rates and economic variables have been considered by many authors, including Patton (2006), Coroneo et al. (2016), Duan and Miao (2016), and Wellmann and Trück (2018). These apply standard estimation techniques, such as ordinary least squares, but then forecasts are limited to inferences on the means and variances of the dependent variables, conditional on each factor. By contrast, we use factor quantile regressions, which allow the explanatory variables to affect the dependent variables differently for each τ-quantile, and estimation can trace out the conditional distribution of each dependent variable as τ ranges from 0 to 1. Thus, to capture this flexibility, we extend the contemporaneous quantile-regression framework of Gaglianone and Lima (2012) to multiple factors as follows:

$$\mathbf{y}_t = \boldsymbol{\alpha}^{(\tau)} + \mathbf{B}^{(\tau)}\mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(\tau)}, \qquad t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{\varepsilon}_t^{(\tau)}$ are quantile-dependent error processes, $\boldsymbol{\alpha}^{(\tau)}$ are the intercepts and $\mathbf{B}^{(\tau)}$ are the matrices of quantile regression coefficients. We can view $\mathbf{y}_t^{(\tau)} = \boldsymbol{\alpha}^{(\tau)} + \mathbf{B}^{(\tau)}\mathbf{x}_t$ as the vector containing the τ-quantile of each element of $\mathbf{y}_t$, conditional on $\mathbf{x}_t$.

Motivated by the relatively weak fit of forecast models using lagged explanatory variables, especially when multiple quantiles are considered, we shall assume a contemporaneous relationship between dependent and explanatory variables. In the studies of Cenesizoglu and Timmermann (2008) and Zhu (2013), most of the lagged economic predictors for the stock and bond returns are not statistically significant in the quantile regressions. By contrast, Bunn et al. (2016) utilize contemporaneous information in their quantile model, which performs well against asymmetric GARCH models with non-normal innovations.

Thus, our starting point is similar to a multi-factor generalisation of the quantile regressions in Gaglianone and Lima (2012). To derive forecasts for our dependent variables, we shall need to set values for our independent variables. In general, FQ models may use any consistent set of values $\mathbf{x}^*$ that accounts for the dependency structure between the explanatory variables. Assuming such a set is available, we can estimate the quantile regressions (1) using historical data for $t = 1, \ldots, T$, for some pre-defined set $\mathbb{Q}$ of $\tau \in (0, 1)$, and then predict each conditional quantile as:

$$\hat{\mathbf{y}}^{(\tau)}|\mathbf{x}^* = \hat{\boldsymbol{\alpha}}^{(\tau)} + \hat{\mathbf{B}}^{(\tau)}\mathbf{x}^*. \tag{2}$$

Next, consider the quantiles at the levels in $\mathbb{Q}$ and focus on the ith element of $\mathbf{y}$. If $\mathbb{Q}$ outlines a sufficiently dense grid, the shape of the entire conditional distribution function $F_i|\mathbf{x}^*$ of y_i can be estimated through $\{(\tau, \hat{\mathrm{y}}_i^{(\tau)}|\mathbf{x}^*) : \tau \in \mathbb{Q}\}$. The optimal node positions depend on $F_i|\mathbf{x}^*$ and should focus on parts where the distribution is expected to be irregular. Since fitting the tails of the distribution is more of a challenge than fitting the centre, nodes concentrated around the tails are beneficial.

Multiple methods have been applied to interpolate a continuous distribution from the estimated quantiles: Koenker and Bassett (1982) use a step function that assigns the value of the next smallest quantile in $\tau \in \mathbb{Q}$. This method is adapted by Cenesizoglu and Timmermann (2008) and Pedersen (2015); kernel density estimations, e.g., with Gaussian or Epanechnikov kernel, can be employed as in Koenker and Bassett (2010) and Gaglianone and Lima (2012); or shape-preservation can be maximized using the Piecewise Cubic Hermite Interpolating Polynomials (PCHIP) algorithm of Fritsch and Carlson (1980). We explain our reasons for using the third alternative in the next section.

Given a vector $\mathbf{x}^*$ of values for the common factors, denote the interpolated conditional distribution functions by $\hat{F}_i|\mathbf{x}^*$, for $i = 1, \ldots, n$. The probability integral transform variables are uniformly distributed if the forecast is probabilistically calibrated and will only be independent if the residuals $\varepsilon_i|\mathbf{x}^* = F_i - \hat{F}_i|\mathbf{x}^*$ are independent which may be not the case unless the factor model perfectly represents the regressand without any missing variables or similar problems. Otherwise, we capture dependence using an extension of Sklar's theorem to conditional copulas due to Patton (2006) which represents a joint conditional distribution in terms of a unique conditional copula defined by

$$\hat{F}\left(\mathbf{y}|\mathbf{x}^*\right) = C\left(\hat{F}_1\left(y_1|\mathbf{x}^*\right), \ldots, \hat{F}_n\left(y_n|\mathbf{x}^*\right) \middle| \mathbf{x}^*\right), \tag{3}$$

where C denotes the conditional copula, which is a multivariate distribution function with marginal distributions that are uniform on [0, 1]. This way, any conditional marginals can be transformed into a valid multivariate distribution provided the copula is conditioned on the same variables as the marginal distributions. As Patton (2013) points out, this multi-stage approach results in a multivariate model without the challenges associated with simultaneous estimations in high dimensions.

To summarize, the general methodology of FQ models proceeds as follows:

Stage 1 Estimate quantile regressions for τ-quantiles where $\tau \in (0, 1)$ are pre-specified by a grid $\mathbb{Q}$ of quantile levels;

Stage 2 For a given vector $\mathbf{x}^*$ for the common factors, interpolate over the estimated conditional quantiles to obtain each conditional marginal $\hat{F}_1|\mathbf{x}^*, \ldots, \hat{F}_n|\mathbf{x}^*$;

Stage 3 Use a conditional copula and apply (3) to obtain the joint conditional distribution.

3.1. A simple illustration of a conditional factor quantile model

Consider the case where dependent variables are excess stock returns r_{it} with $i = 1, \ldots, n$ and the factor model is the two-factor Capital Asset Pricing Model (CAPM) introduced by Kraus and Litzenberger (1976). Through the inclusion of a quadratic term in the excess market return r_{tM}, the two-factor CAPM captures different sensitivities to positive and negative returns and allows the systematic risk of a stock to be related to skewness, as in Harvey and Siddique (2000). The quantile regressions are:

$$r_{it} = \alpha^{(\tau)} + \beta^{(\tau)} r_{tM} + \gamma^{(\tau)} r_{tM}^2 + \varepsilon_{it}^{(\tau)},$$
$$t = 1, \ldots, T \text{ and } i = 1, \ldots, n. \tag{4}$$

All conditional quantiles of the quadratic CAPM in (4) are calibrated on data from 03 January 2000 to 28 June 2018. The market return is on the S&P500 index, and all distributions are conditional on the realized S&P500 return on 29 June 2018. All data are of daily frequency, and we use 2000 observations to fit the quantile regression models.

We start by illustrating the selection of the quantile grid, which depends on the choice of interpolation. To see this, let us compare the properties of three alternative interpolation methods reviewed in the previous section. We estimate quantile regressions for returns on the stock Apple with the S&P500 as market factor and two different quantile grids, one where $|\mathbb{Q}| = 9$ and another where $|\mathbb{Q}| = 500$. When $|\mathbb{Q}| = 500$, we use an equidistant grid. For the other, we use $|\mathbb{Q}_9| = \{0.001, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.999\}$ which has more nodes in the extremes. This is to increase the domain of the estimated function without the need for extrapolation. Quantile regression is likely to yield a high sampling error for these extreme nodes because there are fewer data points in those percentiles. By taking the monotonicity requirement for quantiles and the hit-or-miss accuracy of ad-hoc extrapolation into account, additional nodes in the tails should benefit from the accuracy of the estimated distribution. Fig. 1 compares the results for (i) the step function introduced by Koenker et al. (1978) and applied by Cenesizoglu and Timmermann (2008), on the left in orange, (ii) the Epanechnikov kernel advocated by Gaglianone and Lima (2012) in the middle in green, and (iii) the PCHIP interpolation, on the right in blue.

The quantile grid with cardinality 500 produces similar distributions for all three methods. These are indistinguishable in a Kolmogorov–Smirnov test at a significance level of 1%. However, with $|\mathbb{Q}| = 9$, the shape-preserving interpolation fits much better than the kernel or the step function, which yield vastly different distributions depending on the choice of $\mathbb{Q}$. Only the shape-preserving interpolation produces similar results for both grid sizes. The smaller the grid size, the further apart the quantiles and the lower the chance that estimated quantiles exhibit non-monotonic behaviour. To quantify the additional quantile grid requirements of the kernel and the step function, we sample from distributions with varying equidistant quantile grids and compare them with the estimation based on $|\mathbb{Q}| = 500$ through a Kolmogorov-Smirnov test. Table 1 lists the p-values. The kernel requires a grid with between 30 and 40 points, and the step function requires a grid with between 40 and 50 points to achieve a similar distribution at the 1% significance level. However, the shape-preserving interpolation with $|\mathbb{Q}| = 9$ yields a function that a Kolmogorov–Smirnov test cannot distinguish from the one based on $|\mathbb{Q}| = 500$ at a significance level of 1%. The lower cardinality requirement of the shape-preserving interpolation is especially relevant in practice since it leads to major computational improvements. It is much faster to use fewer quantiles to specify the quantile regressions, and then apply the shape-preserving interpolation, than it is to run the numerous quantile regressions that would be required to obtain accurate conditional distributions with either the kernel or the step function approach to interpolation. Hence, in the rest of this paper, we shall use the much faster and more accurate shape-preserving algorithm for interpolating all conditional distributions. Fritsch and Carlson (1980) show that the PCHIP method preserves monotonicity defined by the estimated quantiles, so, provided the estimated quantiles do not cross, the PCHIP distribution will be well-defined.

Table 1
Kolmogorov–Smirnov p-values of distribution comparison (Apple).

$\lvert\mathbb{Q}\rvert$	Step function	Epanechnikov kernel
10	0.0027	0.2562
20	0.4493	0.9154
30	0.8110	0.9855
40	0.9885	0.9996
50	0.9997	0.9996

The quantiles for the return of Apple are calculated with the quadratic CAPM (4) and data from 03 January 2000 to 28 June 2018. We model the market return through the returns of the S&P500 index and condition all distributions on the realized S&P return from 29 June 2018.

Next, we estimate quantile regressions (4) for the quantile grid with $|\mathbb{Q}| = 9$ to another US stock, Procter and Gamble (P&G), over the same time period. Interpolating allows for a visual comparison of the conditional distributions and densities of Apple and P&G, depicted in Fig. 2. Both distributions and densities exhibit irregularities that are difficult to capture with alternative parametric estimations. We use these conditional marginal distributions and fit conditional joint distributions with a Gaussian, a Gumbel and a Clayton copula. The density contours of the conditional joint densities are illustrated in Fig. 3. These show slight but noticeable differences depending on the copula choice. We prefer the Gumbel

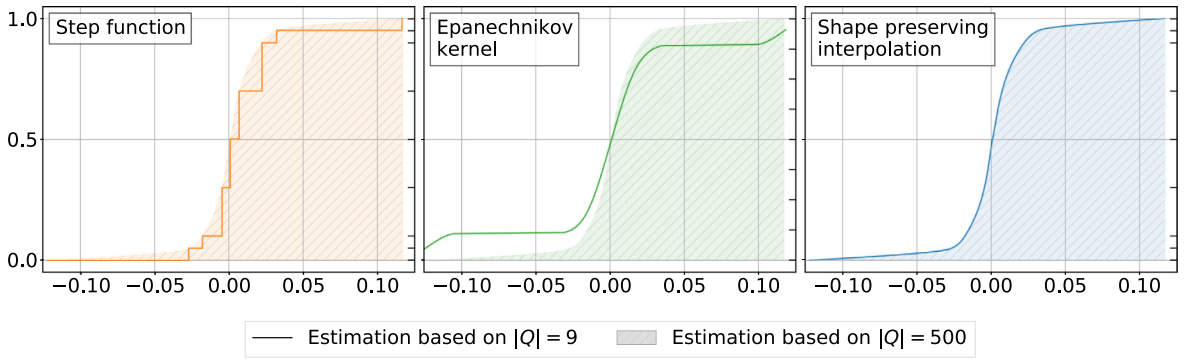

Fig. 1. Distribution estimates with varying quantile grids (Apple). Conditional distributions for the return on Apple based on a grid of equidistant quantile levels with $|\mathbb{Q}| = 500$ (shaded area) are compared with distributions based on a grid with $|\mathbb{Q}| = 9$ (solid line). The step function and the shape-preserving interpolation utilize the smaller quantile grid with a focus on the tails while the kernel estimation uses equidistant nodes as illustrated with the rugs on the right-side axis since this yields better estimations.

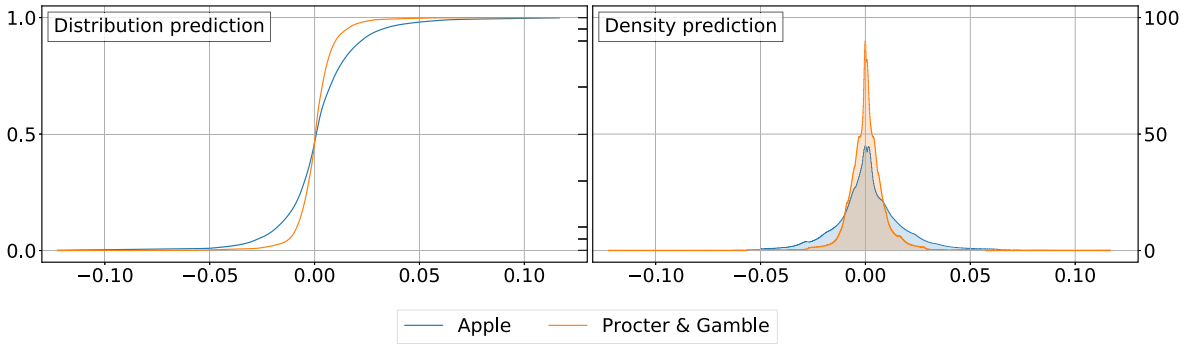

Fig. 2. Conditional distribution and density forecasts (Apple and P&G). The conditional marginal distribution and corresponding density for two US stock returns are generated with a FQ model based on the quadratic CAPM in (4). For the calibration, we use the quantile grid with $|\mathbb{Q}| = 9$ as illustrated with the rugs on the right-side axis of the left figure.

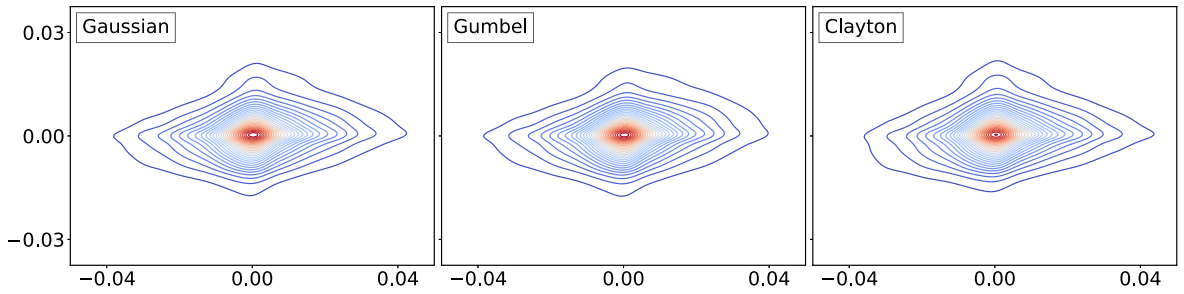

Fig. 3. Density contours of the joint conditional density forecasts (Apple and P&G). We use maximum likelihood estimation on the stock returns to derive the optimal parameters for the Gaussian and Archimedean copulas. This yields $\rho = 0.1988$ for the Gaussian copula and $\theta = 1.1590$ or $\theta = 0.2690$ for the Gumbel and Clayton copula respectively. Apple returns are on the horizontal axis and PG returns are on the vertical axis.

copula for this conditional joint distribution based on the standard information criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). It is also worth pointing out that the Archimedian copulas only have one parameter to model dependency structure, regardless of the number of dimensions of the multivariate distributions. Therefore, even though the Archimedian copulas can be easy to estimate, they inevitably become less and less accurate as the portfolio size increases.

We should emphasize that the dependency structure between the conditional marginal distributions of FQ models depends on the conditional copula. The quantile regression models (1) share the same predictor variables $\mathbf{x}$, but this does not affect conditional rank correlation metrics such as Kendall's τ or Spearman's ρ. Of course, changes in values of $\mathbf{x}$ affect all dependent variables simultaneously, so the unconditional dependency depends on the common factor structure as well as the copula.

3.2. Static latent factor quantile models

In this section, we develop the FQ model when the common factors in Eq. (1) are latent variables corresponding to the first principal components of the covariance matrix of $\mathbf{y}$. Following Stock and Watson (2002), many papers on quantile regression employ principal components derived from the covariance matrix of a set of exogenous predictor variables. Manzan (2015) empirically evaluates the predictive power of principal components of a large number of exogenous macroeconomic indicators when used to augment the Koenker and Xiao (2006) autoregressive model for quantiles. Maciejowska et al. (2016) generalize the quantile regression averaging approach by Nowotarski and Weron (2015) with principal components to avoid the ex-ante model selection. Quantile regression averaging involves applying quantile regression with a set of individual point forecasts as independent variables and the observed value of the predicted variable as the dependent variable. By contrast, we are interested in the case that the latent factors are *endogenous*, in the sense that the principal components are derived from the covariance matrix of the dependent variables alone. This endogenous approach was first employed by Connor and Korajczyk (1993) who used asymptotic results on principal components to determine the appropriate number of factors for explaining returns on US stocks. The endogenous approach is desirable because it does not require extra data other than the dependent variables.

In the following we consider a time series sample $\mathbf{y}_t = (\mathrm{y}_{1t}, \ldots, \mathrm{y}_{nt})'$ for $t = 1, \ldots, T$ on n dependent variables. Latent factor quantile models are also applicable to cross-sectional data because the factors are entirely derived from the eigenvectors of a sample covariance matrix $\boldsymbol{\Sigma}$. Although the FQ approach could equally be applied to cross-sectional data, we assume a time series setting, and we further assume $\mathbb{E}(\mathbf{y}_t) = \mathbf{0}$ for all t without loss of generality. This assumption is common for financial returns.

Denote the matrix with jth column equal to the jth eigenvector by $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$, having ordered these columns so that $\mathbf{w}_k$ is the unit eigenvector corresponding to λ_k, the kth largest eigenvalue of $\boldsymbol{\Sigma}$. Set $\mathbf{p}_t = \mathbf{W}'\mathbf{y}_t = (\mathrm{p}_{1t}, \ldots, \mathrm{p}_{nt})'$ so that p_{kt} is the kth principal component and its in-sample variance λ_k decreases as k increases. Because it is orthogonal, $\mathbf{W}' = \mathbf{W}^{-1}$, so inverting $\mathbf{W}$ yields the full principal component representation of the original system in terms of uncorrelated latent variables as $\mathbf{y}_t = \mathbf{W}\mathbf{p}_t$. Typically, the number of factors m is selected so that a large fraction, but not all, of the total variance is explained, i.e., $m < n$. Then a statistical factor model based on m endogenous principal component factors is an approximate representation $\mathbf{y}_t \approx \mathbf{W}_m\mathbf{x}_{mt}$ where $\mathbf{W}_m$ denotes the first m columns of $\mathbf{W}$ and $\mathbf{x}_{mt} = (\mathrm{p}_{1t}, \ldots, \mathrm{p}_{mt})'$. In the literature, it is often assumed that a principal component representation based on the first m components approximates the original data by omitting the variation captured by the last $n - m$ components, in effect disregarding any importance this variation has for forecasting.

Now consider the choice of latent variables. There is a trade-off between setting m small enough to ignore the variation that is regarded as unimportant and large enough to capture sufficient variation in the system to be informative for forecasts. Rules of thumb exist (such as taking m to be large enough to capture at least 95% of the variation and the other 5% being assigned to the information that is not useful for forecasts), but this is essentially a matter of empirical design which we discuss in more detail in Section 4.2. Also, as in any factor model, forecasts for $\mathbf{y}_t$ are conditional on some predetermined values $\mathbf{x}_m^*$ for $\mathbf{x}_m$. Because our endogenous latent factors are contemporaneous, $\mathbf{x}_m$ is unknown. One approach is to fit a dynamic model for $\mathbf{x}_m$ to obtain a forecast $\mathbf{x}_m^*$, which can then be fed into the factor quantile models. However, this approach defeats the simplicity of the factor quantile models. In this paper, we consider a static approach, where we use the unconditional distribution induced by the factor quantile models as our predictive density.

To obtain the static predictive density, we first consider the "bootstrap aggregation", commonly abbreviated to *bagging*, via an algorithm proposed by Breiman (1996). When data $\mathbf{Z}$ are used in some model to obtain a distribution $\hat{F}$ the meta-algorithm generates B bootstrap samples $\mathbf{Z}^1, \ldots, \mathbf{Z}^B$, each having the same pre-defined size, by drawing from $\mathbf{Z}$ with replacement. Then the CDF forecast based on bagging is the arithmetic average $\hat{F}^{\text{bag}} := B^{-1}\sum_{b=1}^{B}\hat{F}^b$, where $\hat{F}^b$ is the forecast based on data $\mathbf{Z}^b$. To proceed, we assume the quantile vector $\hat{\mathbf{y}}^{(\tau)}$ follows a multivariate Gaussian distribution. The mean and variances of this multivariate Gaussian distribution can be computed as follows. Let us assume the parameter estimators are denoted as $\hat{\boldsymbol{\alpha}}_m^{(\tau)}$ and $\hat{\mathbf{B}}_m^{(\tau)}$, where m indicates that these parameters are associated with m principle components. The assumption $\mathbb{E}(\mathbf{y}) = \mathbf{0}$, together with our construction based on the principal component analysis, implies that $\mathbb{E}(\mathbf{x}_m) = \mathbf{0}$. Hence, Eq. (2) induces the following mean of each prediction quantile:

$$\mathbb{E}\left(\hat{\mathbf{y}}^{(\tau)}\right) = \hat{\boldsymbol{\alpha}}_m^{(\tau)}. \tag{5}$$

Further, since the principal components are uncorrelated with each other, the estimated conditional covariance between the τ_k- and τ_l-quantiles is:

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(\hat{\mathbf{y}}^{(\tau_k)}, \hat{\mathbf{y}}^{(\tau_l)}\right) &= \mathbb{C}\mathrm{ov}\left(\hat{\boldsymbol{\alpha}}_m^{(\tau_k)} + \hat{\mathbf{B}}_m^{(\tau_k)}\mathbf{x}_m^*, \hat{\boldsymbol{\alpha}}_m^{(\tau_l)} + \hat{\mathbf{B}}_m^{(\tau_l)}\mathbf{x}_m^*\right)\\ &\approx \hat{\mathbf{B}}_m^{(\tau_k)}\mathrm{diag}(\lambda_1, \ldots, \lambda_m)\hat{\mathbf{B}}_m^{(\tau_l)'}.\end{aligned} \tag{6}$$

Eqs. (5) and (6) can be used to define the multivariate Gaussian distribution for the quantile predictions of different levels for each variable. Then, we can obtain the bootstrap samples $\mathbf{Z}^1, \ldots, \mathbf{Z}^B$ from this multivariate Gaussian distribution.

For each bagging draw, we generate the distribution forecast $\hat{F}^b$, for $b = 1, \ldots, B$. Then we average them, setting

$$\hat{F}^{\text{bag}} = B^{-1}\sum_{b=1}^{B}\hat{F}^b,$$

Pseudocode for the bagging algorithm is shown in Algorithm 1.

Algorithm 1: Factor Quantile Model with Bootstrap Aggregation

Input : Grid $\mathbb{Q}$ of quantile levels with $0 < \tau < 1$ for all $\tau_1, ..., \tau_q \in \mathbb{Q}$ with $q = |\mathbb{Q}|$; Observations on $\mathbf{y}_t$ for $t = 1, \ldots, T$;

Output: Unconditional multivariate distribution $\hat{F}_{T+1}$ of $\mathbf{y}_{T+1}$;

Use observations to calculate the first $m \leq n$ principal components $\mathbf{x}_t = (p_{1t}, \ldots, p_{mt})$ where m is determined by the target for the variance explained;

for $i = 1, \ldots, n$ **do**

  Estimate the factor quantile regressions

$$y_{it} \leftarrow \alpha_{im}^{(\tau_k)} + \mathbf{B}_{im}^{(\tau_k)}\mathbf{x}_t + \varepsilon_{it}^{(\tau_k)}$$

  which yields the parameters $\hat{\alpha}_{im}^{(\tau_k)}$ and $\hat{\mathbf{B}}_{im}^{(\tau_k)}$ for each $\tau_k \in \mathbb{Q}$; where i and m indicate that these parameters are associated with $y_{i,t}$ and using m principle components, respectively.

  Compute mean and covariance matrix for the quantiles as

$$\hat{\boldsymbol{\mu}}_i \leftarrow \left(\hat{\alpha}_{im}^{(\tau_k)} : \tau_k \in \mathbb{Q}\right),$$

$$\hat{\mathbf{V}}_i \leftarrow \left(\hat{\mathbf{B}}_{im}^{(\tau_k)}\mathrm{diag}(\lambda_1, \ldots, \lambda_m)\hat{\mathbf{B}}_{im}^{(\tau_l)'}\right)_{kl}$$

  for $b = 1, \ldots, B$ **do**

    Draw one q-dimensional sample $\mathbf{q}_b \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i\right)$;

    Interpolate $\mathbf{q}_b$ through shape-preserving interpolation to a distribution $\hat{F}_{T+1}|\mathbf{q}_b$;

  end

  Sample from $\hat{F}_{T+1,i}|\mathbf{q}_1, \ldots, \hat{F}_{T+1,i}|\mathbf{q}_B$ and aggregate samples to an estimate of $\hat{F}_{T+1,i}$, the unconditional distribution function of $y_{T+1,i}$ with an empirical distribution function;

end

Generate the multivariate distribution with the marginal distributions and a copula.

$$\hat{F}_{T+1}(\mathbf{y}) \leftarrow C\left(\hat{F}_{T+1,1}(y_1), \ldots, \hat{F}_{T+1,n}(y_n)\right);$$

We can also summarise the algorithm in the following stages:

Stage 1 Given observations at times $t = 1, \ldots, T$ for n stationary zero-mean stochastic variables $\mathbf{y}$ take the spectral decomposition of their covariance matrix $\boldsymbol{\Sigma}$ and thereby select the first m principal components for common factors, denoted $\mathbf{x}_m$;

Stage 2 Using the same sample, estimate quantile regressions of the form (1) for each τ-quantile in turn, where $\tau \in (0, 1)$ are pre-specified by a grid $\mathbb{Q}$ of $(0, 1)$;

Stage 3 We sample $\left(\hat{y}_k^{(\tau_1)}, \ldots, \hat{y}_k^{(\tau_q)}\right)$ from the multivariate Gaussian distribution, whose mean and variance are specified in Eqs. (5) and (6), we then apply shape-preserving interpolation to construct a marginal distribution and then take a sample size N from this distribution. Repeat this sampling B times. These $N \times B$ observations are combined to form the conditional marginal distribution $\hat{f}_k$ for each element k of the dependent variable;

Stage 4 Select a copula function to obtain the multivariate distribution forecast.

Bootstrap sampling is extremely fast, so we can set N and B to be very large numbers. For instance, in the empirical study of the next section, we set $N = 100{,}000$ and $B = 250$ so that 25 million samples are taken from each conditional marginal during the bagging algorithm. Alternative methods such as kernel density estimation could also be applied to aggregate the $N \times B$ observations to a distribution. However, this is unnecessary when $N \times B$ is large.

A simple alternative to bagging is to use only the expected value (5) and ignore the covariances (6). When the FQ models use the first few principal components, there exists considerable variation about the point forecast, which is the unconditional expectation of the quantiles, because the higher principal components have the greatest variation. So instead, this "alpha" latent FQ version employs the last $n-m$ principal components in quantile regressions. This way, the intercept captures a point forecast about which there is much less variation. The statistical properties described above remain valid as the common factors remain uncorrelated, but now the intercepts $\hat{\boldsymbol{\alpha}}_{n-m}^{(\tau)}$ capture an expected value with little variation: the covariance (6) is minimal because $(\lambda_{m+1}, \ldots, \lambda_n)$ consists of the smallest eigenvalues. This is not a new idea. Following Jensen (1968), using the intercept to encompass the remaining variation not explained by factors is now widely applied to the performance evaluation of portfolio managers.

4. Empirical results

We compare FQ models with benchmark models that are commonly applied to systems of financial and economic variables: (i) two asymmetric Student-t multivariate GARCH(1,1) models and (ii) a Gaussian copula with empirical marginals. These have been selected as (i) the family of parametric dynamic models that best capture the salient properties of financial time series, i.e., volatility clustering, skew and heavy tails, asymmetric response to shocks, and (ii) a copula that is amenable to high-dimensional systems and also performs well in previous forecast exercises (Patton, 2012, 2013). Of course, there are many models available, but including further models would provide information that detracts from the clear messages of this paper. Also, note that, since GARCH models do not scale well to higher dimensions, we have limited the dimensions of the systems selected in our empirical study. That is the only reason we have not

considered very large systems. FQ models scale easily and naturally to higher dimensions and retain very fast calibration times.

Let F denote the data generation process. A scoring rule is proper if the expected score is minimized when the forecaster issues the probabilistic forecast F, rather than another distribution $G \neq F$, and it is strictly proper if this minimum is unique. See Gneiting and Raftery (2007) for further discussion. Since the goal of a probabilistic forecast is to maximize the sharpness of the distribution forecast, subject to calibration, we focus our assessment on proper scoring rules which address both calibration and sharpness simultaneously (Winkler, 1996). Also, as recommended by Gneiting et al. (2008) and Scheuerer and Hamill (2015) we utilize multiple univariate and multivariate proper scores.

We compare two versions of our latent FQ model with two standard econometric models for predicting systems of exchange rates, term structures of interest rates and commodity future indices. We assess the model's accuracy using univariate and multivariate proper scoring rules. Section 4.1 begins with a specification of the proper scoring rules and briefly describes the benchmark models. Then Section 4.2 details the data used for this empirical study and outlines the model calibration. Section 4.3 presents results obtained using the weighted CRPS for univariate distribution forecasts, and Section 4.4 summarises results for multivariate distribution forecasts using the energy and variogram scores with different parameters. Many results cannot be reported in detail, but they are available from the authors on request, along with the data and code used to generate these results.

4.1. Empirical design

Scoring rules are a type of distance measure between a predictive distribution and an observation. They can be used to compare the predictive performance of competing models. In the class of densities with finite first moments, the weighted CRPS is a strictly proper scoring rule that is easy to compute and very flexible. It compares distribution forecasts by focusing on certain regions of interest, such as the centre or the tails. Introduced by Matheson and Winkler (1976), it is the recent work of Gneiting and Ranjan (2011) that drew attention to this score and the need for proper scoring rules applied to univariate distribution forecasting. Given a forecast distribution F of an unknown data generation process and a realization y from this unknown process, the weighted CRPS is defined as

$$C_w(F, y) = 2 \int_0^1 \left(\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha \right) \left(F^{-1}(\alpha) - y \right) w(\alpha) \, \mathrm{d}\alpha,$$

where $w(\alpha)$ is a weight function which specifies a focus on particular parts of the distribution. Gneiting and Ranjan (2011) recommend using $w(\alpha) = 1$ for the entire distribution, $w(\alpha) = \alpha(1-\alpha)$ for the centre, $w(\alpha) = \alpha^2$ for the left tail, $w(\alpha) = (1-\alpha)^2$ for the right tail and $w(\alpha) = (2\alpha - 1)^2$ for both tails of the distribution.

For ranking multivariate distribution forecasts with proper scoring rules we consider the energy score (Székely, 2003) which generalizes the kernel representation of CRPS specified by Gneiting et al. (2008) and the variogram score is proposed by Scheuerer and Hamill (2015) via a completely different construction principle. To define these scores we require the following notation: Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be an observation of the n-variate random vector $\mathbf{Y}$ and let F be a forecast of the multivariate distribution of $\mathbf{Y}$. The energy score is defined as

$$\mathrm{ES}(F, \mathbf{y}) = -\frac{1}{2}\mathbb{E}_F \left(\|\mathbf{Y} - \mathbf{Y}'\| \right) + \mathbb{E}_F \left(\|\mathbf{Y} - \mathbf{y}\| \right)$$

where $\| \cdot \|$ denotes the Euclidean norm and $\mathbf{Y}$ and $\mathbf{Y}'$ are independent random vectors with distribution $F \in \mathcal{F}$, the class of Borel probability measures such that $\mathbb{E}_F(\|\mathbf{Y}\|)$ is finite. Székely (2003) proves that the energy score is strictly proper relative to $\mathcal{F}$. The variogram score of order p is defined as

$$\mathrm{VS}_p(F, \mathbf{y}) = \sum_{i,j=1}^{n} \left(|y_i - y_j|^p - \mathbb{E}_F \left(|Y_i - Y_j|^p \right) \right)^2$$

where Y_i and Y_j are the ith and jth component of a random vector with distribution F. The score is proper relative to the class of the probability distributions for which the $2p$-th moments of all components are finite. The inclusion of the variogram score is especially important since the energy score is not sensitive to misspecification of correlations (Pinson & Girard, 2012).

To rank the performance of the competing models, we employ the Model Confidence Set (MCS) of Hansen et al. (2011) based on the three proper scores above. Given a loss function and an initial set $\mathcal{M}^0$ containing all competing models, MCS applies a sequential equivalence test and an elimination rule to apply when this test is rejected. For some pre-specified α, MCS returns a set of superior models $\mathcal{M}^*_\alpha$ that includes the best models in $\mathcal{M}^0$, in the sense that their performance cannot be distinguished with equivalence tests at a confidence level of $1 - \alpha$.

Consider a finite set $\mathcal{M}$ with models indexed by $i = 1, \ldots, N$ and a loss function L, so that L_{it} is the loss of model i for a forecast at time t. Then for $i, j = 1, \ldots, N$ and $t = 1, \ldots, T$ we define $d_{ij,t} := L_{it} - L_{jt}$ and $\mu_{ij} := \mathbb{E}(d_{ij,t})$. To test $\mathrm{H}_0^{\mathcal{M}} : \mu_{ij} = 0$ for all i, j versus $\mathrm{H}_A^{\mathcal{M}} : \mu_{ij} \neq 0$ for some $i \neq j$ the MCS test statistic is

$$T_{\mathcal{M}} := \max_{i,j \in \mathcal{M}} \left| \overline{d_{ij}} / \sqrt{\hat{\sigma}^2} \right| \qquad \text{where}$$

$$\overline{d_{ij}} = T^{-1} \sum_{t=1}^{T} d_{ij,t},$$

and $\hat{\sigma}^2$ is the bootstrapped estimate of the variance of $d_{ij,t}$. Since the distributions of $T_{\mathcal{M}}$ are non-standard, they have to be estimated through a bootstrap procedure and, as suggested by Hansen et al. (2011), this should avoid the estimation of high-dimensional covariance matrices. To this end we employ a block–bootstrap where the block-length is determined by the maximum number of significant parameters during the fitting of an autoregressive model on the relative performance variable.

If the hypothesis of equal predictive ability is rejected we then identify the worst model e using the elimination rule $e = \text{argmax}_i \left\{ \sup_{j \in \mathcal{M}} \overline{d_{ij}} / \sqrt{\hat{\sigma}^2} \right\}$, and repeat the testing procedure with the updated model set $\mathcal{M} \backslash \{e\}$. Otherwise, we set $\mathcal{M}^* = \mathcal{M}$. This way, forecasting accuracy can be assessed by the frequency that each model remains in the final set $\mathcal{M}^*_\alpha$. The number of models in the MCS increases as we decrease α, just like the size of a confidence interval. We follow Hansen et al. (2011) and most empirical work since, using $\alpha = 0.25$ and 0.1 to generate the 75% and the 90% MCS, the former being a sub-set of the latter.

Next, we define the two classes of established models used in our study. The multivariate GARCH models are from the family of Constant Conditional Correlation GARCH (CCC-GARCH) of Bollerslev (1990) and the Dynamic Conditional Correlation GARCH (DCC-GARCH) family of Engle (2002), both of which are widely used in literature. This choice is motivated by Hansen and Lunde (2005) who provide an extensive comparison of 330 univariate GARCH specifications, using the Hansen (2005) superior predictive ability data-snooping check, concluding that it is hard to beat an asymmetric GARCH(1,1) model with Student-t innovations. The symmetric version is sufficient for some exchange rates, and for some stocks, a Gaussian conditional distribution for the errors performs as well. However, these models are nested within our more general multivariate specification. Both CCC- and DCC-GARCH are based on the decomposition of the covariance matrix $\boldsymbol{\Sigma}_t$ of the asset returns with $\boldsymbol{\Sigma}_t = \mathbf{D}_t \mathbf{C}_t \mathbf{D}_t$, where $\mathbf{D}_t$ is a diagonal matrix of the time-varying univariate GARCH volatilities. To account for the well-documented asymmetric response of volatility to positive and negative shocks in returns, we employ the E-GARCH model of Nelson (1991) with Student-t innovations for the variances of each asset return y_{it}, thereby specifying the following data generation process:

$$\begin{aligned} y_{it} &= \mu + \varepsilon_{it}, \quad \varepsilon_{it} = \sigma_{it} z^{\nu}_{it}, \\ \log\left(\sigma^2_{it}\right) &= \kappa_i + \gamma_i \log\left(\sigma^2_{i,t-1}\right) \\ &\quad + \alpha_i \left[\frac{|\varepsilon_{i,t-1}|}{\sigma_{i,t-1}} - \mathbb{E}\left(\frac{|\varepsilon_{i,t-1}|}{\sigma_{i,t-1}} \right) \right] + \xi_i \left(\frac{\varepsilon_{i,t-1}}{\sigma_{i,t-1}} \right), \end{aligned} \tag{7}$$

where z^{ν}_{it} follows a Student-t distribution with ν_i degrees of freedom and $\kappa_i, \gamma_i, \alpha_i, \xi_i$ are the GARCH parameters. Bollerslev (1990) assumes that the correlation matrix $\mathbf{C}_t$ is not time varying and uses a constant correlation matrix in CCC-GARCH while Engle (2002) extends $\mathbf{C}_t$ in DCC-GARCH to a time-varying but non-stochastic matrix. The literature on empirical studies which compare the accuracy of different asymmetric univariate GARCH models does not agree on a single superior parametrisation. Therefore, it is unlikely that our results would change if we employed a different asymmetric model (e.g., GJR-GARCH). Besides this, our purpose here is not to test the accuracy of different GARCH models, it is to validate the use of FQ models relative to the standard models that are commonly used for predicting financial returns.

The other model class, which has established itself in the finance literature, are based on an Empirical Distribution Function (EDF). These will be combined into a multivariate distribution using a Gaussian copula with a historical correlation matrix estimated on the same data used for calibration. There are, of course, numerous alternative parametric choices for both marginals and copula, as described by Patton (2013). We opted for the EDF because it is simple to implement in practice. In addition, it has been shown that the EDF approach performs competitively compared with parametric approaches in the literature. Using EDF marginals based on the same data as the FQ marginals allows us to test the effectiveness of PCA factor models, in the context of quantile regressions, for reducing the noisy variation that could deteriorate the forecasting accuracy of models with EDF marginals. To summarize, we have selected a parsimonious set of alternative models and two different benchmark FQ parametrisations, and we will compare the models' predictive performance as measured by proper scoring rules. The margins and the entire multivariate predictive distributions are evaluated separately, which helps to quantify the benefit of accurately predicting the correlation structure.

4.2. Data and calibration

We use three eight-dimensional multivariate time series data sets on USD exchange rates, US interest rates and Bloomberg investable commodity indices. Within each set, we have selected variables to broadly represent the asset class: the exchange rates are those with the highest trading volume (excluding the Chinese Renminbi, which was pegged to the USD until recently); the interest rates span the term structure of US Treasury bonds from 6 months to 20 years; and the commodity indices are chosen to represent the energy, metals, softs and livestock sectors. The exchange rate and commodity index data are obtained from Thomson Reuters Datastream, and the interest rates data are downloaded from the US Treasury website. All series are daily and end on 31 December 2018. The start date varies with data availability, being 01 January 1991 for the commodity indices, 01 January 1994 for the interest rates and 01 July 1999 for the exchange rates. The exchange rate data started in 1999, and we need 2000 observations to calibrate the models. With approximately 250 trading days per year, we could not start the exchange rate evaluations until March 2007. We used the same sub-samples for all data sets for comparison. The models are re-calibrated daily, and the estimated parameters are used to generate one-day-ahead distribution forecasts. Then the fixed-size calibration sample is rolled forward one day, and the forecasts are repeated. We avoid data snooping by using a broad range of data sets with assets motivated by economic factors rather than the predictive prowess of our models. All parameters of the FQ models are chosen based on available ex-ante criteria. Additionally, we quantify the performance based on very long time series, further limiting the probability that any superior performance can be attributed to chance.

We calibrate the multivariate GARCH models using maximum likelihood estimators adapted from the

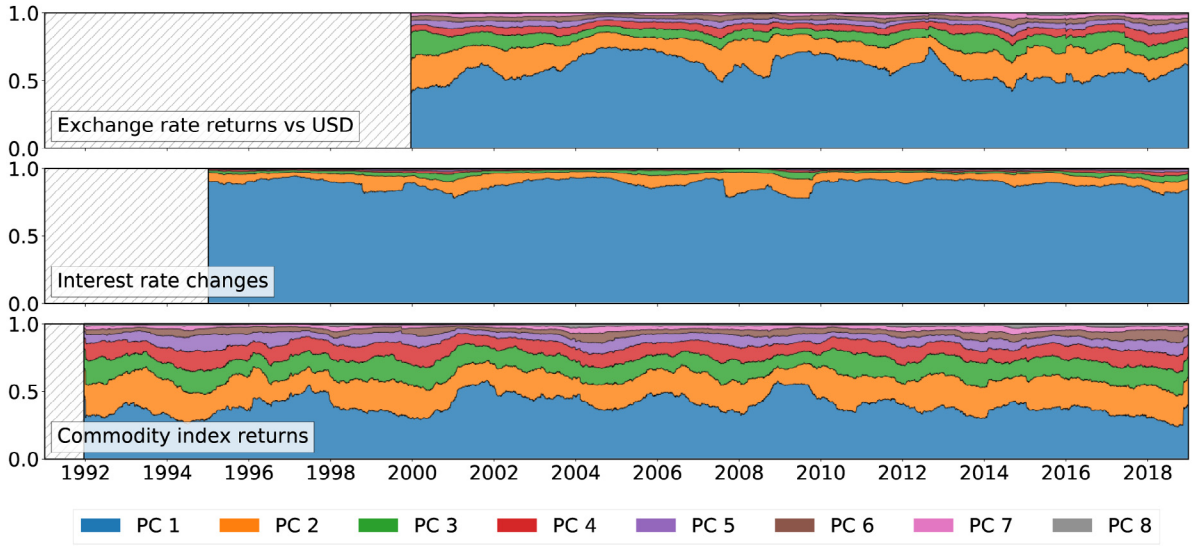

Fig. 4. Cumulative variance explained by the principal components. The variance explained by each principal component is derived by applying PCA on daily rolling windows of 250 data points. Each data set starts on a different date and the results begin approximately one year after the start date.

implementation in the Oxford MFE Toolbox by Sheppard (2013) to utilize E-GARCH with Student-t distributed innovations. We have replaced the univariate Gaussian GARCH(1,1) in the MFE toolbox code for CCC- and DCC-GARCH with Student-t E-GARCH. It is well known that multivariate GARCH models can have ill-conditioned likelihood functions that are hard to optimize unless the calibration sample has sufficient size. Therefore, we have selected daily 2000 returns for each time series for this calibration. We prefer to confine each set to eight dimensions to limit the computational complexity when estimating the multivariate GARCH models. This point is discussed in more detail at the end of this sub-section.

Regarding the FQ specifications, we apply the latent versions based on the last principal components (FQ-A) and bagging (FQ-B) with the same Gaussian copula as the EDF. Both specifications of our FQ model use the quantile grid based on $\mathbb{Q}_9$ from Section 3.1 for the regressions and employ the shape-preserving method for interpolating distribution functions. The estimated conditional quantiles exhibited no crossing behaviour on any data set with any of the calibration choices in our empirical study, indicating that our factor models are well-conditioned. However, if any issues with the non-monotonicity of quantiles arise in an application, many methods could be employed to ensure non-crossing quantiles. For instance, see Koenker et al. (2005), Chernozhukov et al. (2010), Rodrigues and Fan (2017) or Santos and Kneib (2020).

In FQ-B, we select $m = 4$ components as common factors for the exchange rates, $m = 2$ for the interest rates and $m = 6$ for the commodity indices. By depicting the cumulative variance explained by the rolling principal components over the available data period for each asset class, Fig. 4 motivates how these values of m are selected. On average, over the entire period shown, the four components explain 90% of the variation in the exchange rate data, the two components explain 95% of the variation in the interest rates, and the six components explain 95% of the variation in the commodity returns. Following the same reasoning, FQ-A uses $m = 4$ components as common factors for the exchange rates, $m = 6$ for the interest rates and $m = 2$ for the commodity indices.

Both GARCH models are restricted to long calibration periods to estimate the long-term variance and the stability of calibrated parameters. For consistent comparison with the GARCH models, which are not well-conditioned on smaller calibration sizes, we have also taken 2000 data points for the other models. But EDF and FQ models are likely to perform better on smaller calibration sizes. FQ models yield robust estimates with principal component factors even with a calibration size of 250. Here we present results for the FQ and EDF models only for calibration sizes of 2000 and 250, although others are available on request.[5] Table 2 summarises the models that we apply in the remainder of this study.

Despite the necessarily large calibration sizes, both multivariate GARCH models exhibit calibration issues because there are (at least) 40 parameters for an eight-dimensional time series. Hence, the likelihood surfaces are very challenging to optimize. Sometimes parameter estimates do not converge to sensible values, particularly for the commodities data. In such cases, we exchange erroneous parameters with the most recent unproblematic values, as illustrated by Fig. 5. Locating and correcting miscalibrated parameters requires manual attention, which prevents the full automation of multivariate GARCH models.

[5] We compared several calibration sizes between 250 and 2000, finding that, in general, the models with smaller calibration sizes performed better than those with larger ones. We included a calibration sample of size 2000 for comparison with the GARCH models because these could not be calibrated robustly on a calibration size of 250.

Table 2
Summary of models used in the empirical study.

Model	Marginals	Dependency	Calibration size
FQ-A_{250}	Alpha FQ	Gaussian copula	250
FQ-A_{2000}	Alpha FQ	Gaussian copula	2,000
FQ-B_{250}	Bagging FQ	Gaussian copula	250
FQ-B_{2000}	Bagging FQ	Gaussian copula	2,000
EDF_{250}	EDF	Gaussian copula	250
EDF_{2000}	EDF	Gaussian copula	2,000
CCC-GARCH	Student-t E-GARCH(1,1)	CCC	2,000
DCC-GARCH	Student-t E-GARCH(1,1)	DCC	2,000

This table summarizes the acronyms used to denote each model: FQ stands for factor quantile; EDF stands for empirical distribution function; CCC stands for constant conditional correlation, which is equivalent to the Gaussian copula calibrated to a sample size 2000; and DCC stands for the time-varying dynamic conditional correlation matrix.

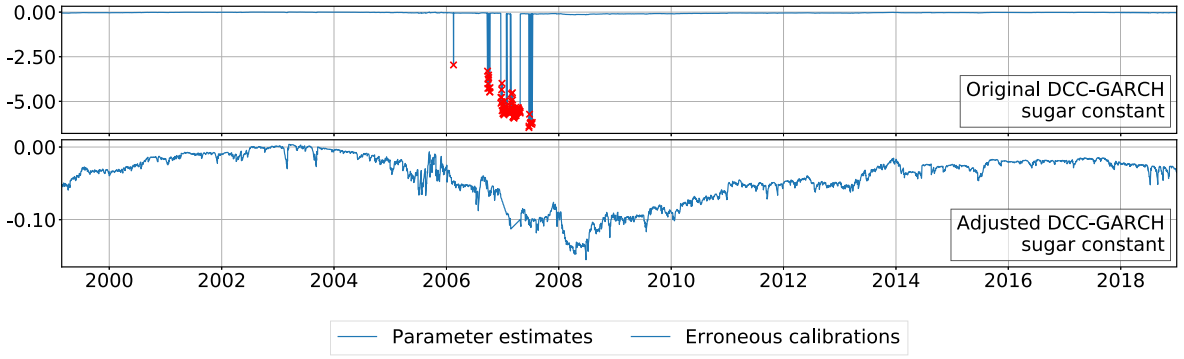

Fig. 5. Convergence issues with GARCH models (sugar). The parameter illustrated is the constant estimated for the sugar marginal in DCC-GARCH. The upper figure shows the parameter obtained using the adapted Oxford MFE toolbox and the lower figure shows the parameter after replacing erroneous calibrations with the most recent unproblematic value. Parameters that differ by a very large amount from previous estimations are classified as mis-calibrations and are marked by red crosses in the upper figure.

We emphasise that FQ models are much easier to fully automate and by no means limited to eight dimensions. Thus, they are more amendable to the type of high-frequency trading that is common among hedge funds that often employ algorithms to re-balance portfolios every day. We have restricted this study to eight dimensions because the problems documented above with calibrating GARCH parameters are even further exacerbated. Further note that FQ models are much faster to calibrate than multivariate GARCH, even without dealing with the latter's convergence issues. In our empirical study, the FQ models were over 30% faster than CCC-GARCH and more than five times faster than DCC-GARCH. Also, our implementation of FQ models is based on Python, while the multivariate GARCH models use optimized MATLAB functions. As the efficiency of MATLAB is generally higher than that of Python scripts, we expect that the difference in speed would become even more pronounced when comparing the multivariate GARCH models to an optimized FQ algorithm.

4.3. Univariate forecasting accuracy results

Both Gneiting et al. (2008) and Scheuerer and Hamill (2015) emphasise the importance of testing the accuracy of univariate distribution forecasts derived from multivariate models. Applying multivariate tests alone is not sufficient because we require a model that forecasts accurate marginals and one that correctly captures the dependence between them. So in this section, we present the results of applying weighted CRPS to each model listed in the initial set $\mathcal{M}_0$ defined in Table 2 and then find the MCS derived from these scores.

Table 3, Table 4 and Table 5 present the average scores for the entire out-of-sample period when the CRPS is uniformly weighted, centre weighted and both-tail weighted, respectively. Smaller average scores are preferred. In each table the model with lowest average score has this score highlighted in blue. In some cases the average scores differ only in the third decimal place, but this can still make a difference to inclusion in the MCS. See, for instance, the results for 2-year interest rates in Table 3. We use one asterisk to indicate which models are included in the 90% MCS and two asterisks for models in the 75% MCS. The latter could include more than one model, but most tests identify a single model as the superior one. This suggests that our out-of-sample period is informative enough to select a best model unequivocally. For each of the 24 univariate data sets, the model(s) having lowest average score are highlighted in blue and the last row of each table reports the number of times a given model is included in the 90% MCS, divided by 24.

These average score results show that the most competitive of the proposed FQ models is for the interest rate

Table 3
Average values of uniformly-weighted CRPS.

Asset	GARCH		EDF		FQ-B		FQ-A	
	CCC	DCC	250	2000	250	2000	250	2000
Exchange rate returns ($\times 10^{-3}$)								
AUD	4.46	4.37**	4.46	4.49	4.45	4.55	4.43	4.49
CAD	3.25**	3.26	3.32	3.36	3.30	3.40	3.31	3.36
CHF	3.63	3.48**	3.52	3.53	3.56	3.57	3.50	3.52
EUR	3.35**	3.58	3.36	3.37	3.35**	3.40	3.34**	3.36**
GBP	3.89	3.18**	3.24	3.25	3.24	3.28	3.25	3.29
JPY	3.50**	3.53	3.52	3.53	3.52	3.57	3.49**	3.53
NZD	4.55**	4.74	4.66	4.70	4.64	4.73	4.64	4.70
SEK	4.27**	4.79	4.31	4.35	4.30*	4.39	4.28**	4.35
Interest rate changes								
6 month	1.33**	1.37	1.37	1.44	1.37	1.47	1.37	1.47
1 year	3.09	2.91	1.55**	1.63	1.56	1.63	1.56	1.64
2 year	2.68	2.67	2.42	2.50	2.42**	2.50	2.42*	2.51
3 year	3.12	3.23	2.76	2.82	2.75**	2.82	2.75**	2.83
5 year	3.17	3.19	3.16	3.20	3.14**	3.20	3.14**	3.21
7 year	3.49	3.51	3.26	3.29	3.24**	3.29	3.24**	3.30
10 year	3.51	3.51	3.11	3.14	3.10**	3.14	3.10**	3.14
20 year	3.47	3.51	3.01	3.03	2.99**	3.03	2.99**	3.03
Commodity index returns ($\times 10^{-3}$)								
Copper	8.60*	8.59**	8.68	8.81	8.67	8.92	8.62*	8.81
Corn	8.57**	8.95	8.81	8.88	8.82	8.95	8.77	8.89
Gold	6.21	5.84	5.70**	5.74	5.74	5.80	5.83	5.90
Live Cattle	5.06	4.99**	5.05	5.07	5.05	5.10	5.45	5.56
Nat. Gas	15.16	14.98**	15.25	15.26	15.22	15.37	15.15	15.25
Soybean	8.04	7.68**	7.85	7.88	7.84	7.95	7.83	7.88
Sugar	10.98	10.96	10.93	11.02	10.94	11.08	10.86**	11.01
WTI Oil	11.37**	11.66	11.46	11.53	11.44	11.60	11.38**	11.52
Summary	37.5%	29.2%	8.3%	0.0%	33.3%	0.0%	50.0%	4.2%

The table reports average uniformly weighted CRPS over the entire out-of-sample period for each model and each univariate data set. The models having the lowest average scores are highlighted in blue. Beside the score we indicate which models are in the 90% and 75% MCS, using * and **, respectively, the latter being a sub-set of the former. Note that when there is only one superior model in the 75% MCS, this will also be in the 90% MCS. The bottom row summarizes the percentages each model is in the 90% MCS for all 24 data sets. That is, the number of * or ** occurrences in each column, divided by 24.

returns. An explanation for why static models become particularly competitive for interest rates is that they have less conditional heteroscedasticity than the other data, and that most of the variation of interest rates is captured by just two components, as we have seen in Fig. 4. Another possible explanation for the competitive performance of FQ forecasts, relative to those generated by EDFs is a smoothing effect due to sampling 25,000,000 times from each predictive distribution for the FQ models, versus the available 250 data points in the calibration set for the EDF specification. It is also worth pointing out that our proposed FQ models outperform the static EDF approach. The MCS results based on proper univariate scoring rules also indicate favourable forecasting performance of both FQ specifications. They match or exceed the accuracy of more complicated GARCH models and significantly surpass the accuracy of copula models with EDF marginals for interest rates. The FQ models also perform competitively for the commodity indices.

We observe that FQ and EDF models based on 250 observations almost always outperform their counterparts with 2000 observations. This may be explained by violating the stationarity assumption for the data generating process over very long calibration windows. The EDF model performance is generally worse than that of both FQ models. It only performs well for 1-year interest rates and gold. This indicates that the use of latent principal component factors reduces the amount of unimportant variation in the observed historical data and produces significantly more accurate forecasts.

Table 6 summarizes the MCS inclusion rates for each model overall data and then in three sub-periods that are the same for each data set: (1) March 2007–December 2010; (2) January 2011–December 2014; and (3) January 2015–December 2018. As well as dividing the out-of-sample period, we report results for three different weighting of CRPS as before, i.e., uniformly, centre and both-tail weighted. The first sub-period is a little less than four years because the exchange rate data began in 1999, and we require 2000 observations to calibrate the models. The results in Table 6 demonstrate that forecasting accuracy varies strongly over time, especially for exchange rates that exhibit the most pronounced regime-specific behaviour. Most other studies in the literature evaluate models only on small samples, spanning limited time periods. Our sub-sample analysis shows that, while the

Table 4
Average values of centre-weighted CRPS.

Asset	GARCH		EDF		FQ-B		FQ-A	
	CCC	DCC	250	2000	250	2000	250	2000
Exchange rate returns ($\times 10^{-5}$)								
AUD	85.83	84.97**	85.94	86.20	85.84	87.09	85.44*	86.13
CAD	63.38**	63.50	64.31	64.70	64.04	65.39	64.05	64.65
CHF	68.93	67.43**	67.96	68.02	68.70	68.80	67.64*	67.93
EUR	64.85	67.26	65.12	65.16	64.87**	65.70	64.80**	65.04**
GBP	72.05	61.82**	62.58	62.71	62.63	63.18	62.58	63.12
JPY	67.58**	68.01	67.79	67.90	67.79	68.63	67.36**	67.85
NZD	88.96**	91.30	90.50	90.78	90.04	91.29	90.08	90.71
SEK	82.93**	90.28	83.48	83.90	83.37	84.57	83.06**	83.89
Interest rate changes ($\times 10^{-2}$)								
6 month	25.69**	26.07	25.91*	26.76	26.05	27.31	26.06	27.36
1 year	54.12	51.17	29.76**	30.69	29.85	30.90	29.86	30.98
2 year	49.72	49.55	46.88	47.86	46.81**	48.05	46.86	48.11
3 year	58.83	60.54	53.59	54.41	53.40**	54.41	53.42**	54.49
5 year	61.51	61.74	61.42	61.95	61.11**	61.97	61.12**	61.99
7 year	65.96	66.18	63.42	63.81	63.11**	63.85	63.09**	63.85
10 year	64.85	64.86	60.61	60.90	60.33**	60.99	60.33**	60.96
20 year	63.38	63.74	58.55	58.77	58.28**	58.87	58.29**	58.84
Commodity index returns ($\times 10^{-4}$)								
Copper	16.64**	16.63**	16.73	16.88	16.73	17.11	16.65**	16.88
Corn	16.72**	17.22	17.03	17.10	17.05	17.24	16.96	17.10
Gold	11.71	11.15	10.95**	11.01	11.03	11.13	11.06	11.15
Live Cattle	9.80	9.71**	9.78	9.80	9.79	9.84	10.16	10.28
Nat. Gas	29.41	29.18**	29.53	29.52	29.49	29.73	29.37	29.51
Soybean	15.44	14.95**	15.17	15.18	15.16	15.32	15.12	15.18
Sugar	21.22	21.20	21.19	21.28	21.22	21.40	21.07**	21.28
WTI Oil	22.06**	22.46	22.18	22.25	22.16	22.38	22.05**	22.24
Summary	33.3%	29.2%	12.5%	0.0%	29.2%	0.0%	54.2%	4.2%

The table reports average centre weighted CRPS over the entire out-of-sample period for each model and each univariate data set. The models having the lowest average scores are highlighted in blue. Beside the score we indicate which models are in the 90% and 75% MCS, using *and **, respectively, the latter being a sub-set of the former. Note that when there is only one superior model in the 75% MCS, this will also be in the 90% MCS. The bottom row summarizes the percentages each model is in the 90% MCS for all 24 data sets. That is, the number of * or ** occurrences in each column, divided by 24.

proportion of MCS that includes a given model depends on the sample, the FQ models are still highly competitive, provided they are calibrated on a small sample.

4.4. Multivariate forecasting accuracy results

For evaluating multivariate forecasting accuracy, we apply the energy score and the variogram scores with $p = 0.5, 1, 2$. These values of p were introduced by Scheuerer and Hamill (2015) and are considered typical choices (Jordan et al., 2019). Contrary to the CRPS results, the multivariate scoring rules encapsulate the accuracy for all eight marginals and their dependency into a single score that holistically quantifies the model's performance on a given data set.

Table 7 reports the average values obtained using different multivariate scoring rules. The results are obtained by applying each multivariate model to each asset class and deriving scores from the entire out-of-sample period. As before, preferred models have smaller average scores. The relative accuracy of a given model depends on the scoring rule applied. The energy and variogram scores differ in their recommendations, and none of the scores predominantly favour a specific model; rather, the preferred model depends on the data. Overall, the FQ-A models perform best in terms of average scores, and they are included in more of the 90% MCS. By comparison, there is a 50% inclusion rate of DCC-GARCH, which is much stronger than CCC-GARCH and all the EDF models.

The comparable performance of FQ models, even with a simple Gaussian copula, to DCC-GARCH, is especially relevant since DCC-GARCH is much more computationally intensive. As pointed out in Section 4.2, both FQ versions are at least five times faster and do not require additional attention to check for miscalibrated parameters. Notably, FQ models outperform EDF forecasts again despite sharing the same calibration window and copula. This demonstrates that the variation reduction through our latent factor model improves the accuracy of the distribution forecast considerably.

Compared to the univariate analysis, models with longer calibration windows perform better and are now present in the superior sets. This might be because the standard errors in the correlation matrix decrease as the sample size increases. Further, the performance of DCC-GARCH is much better in the multivariate comparison than in the prior univariate one, even for exchange rate

Table 5
Average values of both-tail-weighted CRPS.

Asset	GARCH		EDF		FQ-B		FQ-A	
	CCC	DCC	250	2000	250	2000	250	2000
Exchange rate returns ($\times 10^{-3}$)								
AUD	10.24	9.74**	10.21	10.45	10.14	10.63	10.12	10.45
CAD	7.13**	7.19	7.47	7.75	7.42	7.86	7.46	7.69
CHF	8.69	7.83**	8.02	8.07	8.15	8.21	7.96	8.05
EUR	7.60	8.94	7.56	7.65	7.51**	7.74	7.50**	7.62
GBP	10.05	7.10**	7.33	7.46	7.34	7.55	7.42	7.64
JPY	8.00**	8.12	8.06	8.14	8.05	8.27	7.97**	8.13
NZD	9.95**	10.92	10.45	10.68	10.34	10.79	10.36	10.68
SEK	9.49**	11.80	9.67	9.97	9.61**	10.09	9.59**	9.97
Interest rate changes								
6 month	30.49**	32.34	32.94	36.92	32.95	37.48	32.91	37.79
1 year	92.42	86.57	36.25**	39.90	36.21**	39.76	36.19**	40.20
2 year	69.32	68.39	54.78	58.31	54.55**	58.17	54.55**	58.58
3 year	76.60	81.08	61.86	64.84	61.48**	64.54	61.41**	64.93
5 year	70.80	71.72	70.36	72.50	69.78**	72.38	69.73**	72.57
7 year	85.32	86.68	72.25	74.07	71.64**	74.01	71.63**	74.10
10 year	91.83	91.96	68.84	70.42	68.34**	70.41	68.26**	70.47
20 year	93.66	96.00	66.41	67.74	65.95**	67.78	65.89**	67.78
Commodity index returns ($\times 10^{-3}$)								
Copper	19.47	19.36**	19.83	20.54	19.78	20.81	19.65	20.54
Corn	18.82**	20.61	20.00	20.39	19.98	20.59	19.86	20.49
Gold	15.27	13.76	13.20**	13.40	13.27*	13.52	14.05	14.43
Live Cattle	11.38	11.09**	11.37	11.50	11.37	11.59	13.90	14.49
Nat. Gas	33.97	33.10**	34.35	34.46	34.27	34.74	34.03	34.44
Soybean	18.64	16.97**	17.85	18.05	17.78	18.22	17.80	18.03
Sugar	24.91	24.82	24.53	25.04	24.48	25.13	24.28**	25.00
WTI Oil	25.48**	26.78	25.87	26.27	25.73	26.42	25.60**	26.25
Summary	29.2%	29.2%	8.3%	0.0%	41.7%	0.0%	50.0%	0.0%

The table reports average both-tail weighted CRPS over the entire out-of-sample period for each model and each univariate data set. The models having the lowest average scores are highlighted in blue. Beside the score we indicate which models are in the 90% and 75% MCS, using * and **, respectively, the latter being a sub-set of the former. Note that when there is only one superior model in the 75% MCS, this will also be in the 90% MCS. The bottom row summarizes the percentages each model is in the 90% MCS for all 24 data sets. That is, the number of * or ** occurrences in each column, divided by 24.

returns where CCC-GARCH was included in more superior sets than DCC-GARCH. This suggests that the time-varying conditional correlation structure improves over the constant conditional correlation that requires strong assumptions that are not fulfilled for many assets.

Table 8 investigates the robustness of the results in Table 7 by separating the out-of-sample period into 3 sub-periods, as before. Here we only report how many of the four multivariate scoring rules include each model in the superior set. For instance, the number 3 for the CCC-GARCH applied to exchange rates for the sub-sample 2007–2010 indicates that 3 out of 4 scoring rules keep this model in the MCS when scores are derived only from this sub-sample. According to these criteria, the ranking depends on the sub-sample, and in most samples, the highest rank is given to DCC-GARCH or FQ-A_{250}.

5. Summary and conclusions

This paper contributes to the empirical analysis of proper multivariate scoring rules and introduces a new Factor Quantile (FQ) model class. The FQ models are flexible and semi-parametric. They are employed to generate multivariate distribution functions where marginals are derived from interpolations on quantiles estimated via factor-model regressions. Their dependence is selected by choosing a parametric conditional copula. It is not a dynamic model, but we demonstrate that its forecasts (based on the idea that their joint historical distribution can reasonably approximate the joint distribution of the variables) are at least as accurate as constant and dynamic conditional correlation models with Student-t asymmetric E-GARCH(1,1) marginals. Moreover, the FQ models have several advantages over multivariate GARCH models. These should make them attractive to banks and asset managers, or any other players involved in portfolio optimisation and multi-asset pricing, who aim to model and/or forecast large multivariate distributions of financial asset returns.

The class of FQ models is very flexible: they can be built on any factor model, and they can use any conditional copula. We have illustrated an application of the FQ model to bivariate stock returns using the asymmetric CAPM factor model with a Gumbel copula. However, in larger-dimensional systems, we strongly advocate using latent principal component factors, for which we have developed two alternative versions. One of them is very simple to implement, and the other requires the use of a bagging algorithm proposed by Breiman (1996).

Table 6
Comparison of univariate performance over time.

	GARCH		EDF		FQ-B		FQ-A	
	CCC	DCC	250	2000	250	2000	250	2000
Uniform								
Exchange rate returns								
2007 to 2010	0.625	0.375	0.25	0.25	0.25	0.25	0.25	0.25
2011 to 2014	0.125	0.00	0.125	0.00	0.75	0.00	0.875	0.00
2015 to 2018	0.50	0.625	0.25	0.25	0.25	0.125	0.25	0.00
Interest rate changes								
2007 to 2010	0.75	0.50	0.00	0.00	0.25	0.00	0.50	0.00
2011 to 2014	0.125	0.125	0.50	0.00	0.50	0.00	0.50	0.00
2015 to 2018	0.375	0.75	0.25	0.25	0.25	0.00	0.375	0.125
Commodity index returns								
2007 to 2010	0.50	0.375	0.00	0.25	0.625	0.00	0.75	0.50
2011 to 2014	0.125	0.125	0.375	0.125	0.625	0.25	0.625	0.00
2015 to 2018	0.50	0.75	0.125	0.125	0.50	0.125	0.375	0.25
Centre								
Exchange rate returns								
2007 to 2010	0.625	0.375	0.25	0.25	0.25	0.25	0.25	0.25
2011 to 2014	0.125	0.00	0.125	0.125	0.875	0.00	0.875	0.00
2015 to 2018	0.50	0.625	0.25	0.375	0.25	0.125	0.25	0.125
Interest rate changes								
2007 to 2010	0.50	0.50	0.00	0.00	0.25	0.00	0.50	0.00
2011 to 2014	0.125	0.00	0.50	0.00	0.625	0.00	0.625	0.00
2015 to 2018	0.375	0.75	0.25	0.25	0.25	0.00	0.375	0.125
Commodity index returns								
2007 to 2010	0.50	0.375	0.00	0.375	0.625	0.00	0.875	0.50
2011 to 2014	0.125	0.125	0.375	0.25	0.625	0.125	0.75	0.00
2015 to 2018	0.50	0.75	0.125	0.25	0.50	0.125	0.375	0.25
Both tail								
Exchange rate returns								
2007 to 2010	0.625	0.375	0.125	0.125	0.25	0.25	0.25	0.25
2011 to 2014	0.125	0.00	0.00	0.00	0.625	0.00	0.875	0.00
2015 to 2018	0.375	0.625	0.25	0.25	0.25	0.125	0.25	0.00
Interest rate changes								
2007 to 2010	0.50	0.50	0.00	0.00	0.25	0.00	0.50	0.00
2011 to 2014	0.25	0.125	0.50	0.00	0.625	0.00	0.625	0.00
2015 to 2018	0.25	0.75	0.25	0.25	0.25	0.00	0.375	0.125
Commodity index returns								
2007 to 2010	0.375	0.375	0.375	0.375	0.625	0.00	0.75	0.50
2011 to 2014	0.125	0.125	0.375	0.125	0.625	0.25	0.625	0.00
2015 to 2018	0.50	0.75	0.125	0.125	0.50	0.125	0.375	0.25

This table shows the proportion of cases that each model is included in the 90% MCS for each data set over different sample periods and based on different weighting for CRPS. In each case, this proportion is derived by counting the number of times the model is in the 90% MCS and dividing this by eight since there are eight variables in each asset class. The model that is included in most of the MCS, for each asset class and sample period, and each CRPS weighting, is highlighted in blue.

Our extensive empirical study forecasting exchange rates, interest rates and commodity futures is the first substantial study of multivariate distribution forecasting for financial asset returns. We assess the accuracy of forecasts using the MCS of Hansen et al. (2011) derived from the (strictly) proper energy score (Székely, 2003), the variogram score (Scheuerer & Hamill, 2015) and the weighted CRPS introduced by Gneiting and Raftery (2007). We highlight how both the scores, and the superior model sets depend on the asset class and the sample. These conclusions accord with Giacomini and White (2006), Machete (2013) and Elliott and Timmermann (2016), all of whom emphasise that there is no single superior approach: the best model depends on the statistical properties of the data and the economic properties of the variable being predicted.

The best univariate model for each data set depends on the weights used in the CRPS, but overall, these scores favour GARCH models for exchange rates and commodity indices and FQ models for interest rates. However, the results are also sample-specific. For instance, in the period 2011 to 2014, the FQ-A model was best for each data set, according to the tailed-weighted CRPS, but from 2015 to 2018, the GARCH-DCC model was best for each data set, according to the same scoring rule. The multivariate results, based on energy scores and variogram scores with different parameters, also depend on the scoring rule employed. The best models for exchange rates are FQ-A(250) according to the variogram scores with $p = 0.5$ and $p = 1$, and GARCH-DCC according to the variogram score with $p = 2$ and the energy score. For interest rates, the best models are GARCH-DCC, according to the variogram scores with $p = 0.5$ and $p = 1$, FQ-A(2000) according

Table 7
Average values for multivariate scores.

Asset	GARCH		EDF		FQ-B		FQ-A	
	CCC	DCC	250	2000	250	2000	250	2000
Exchange rate returns								
$VS_{0.5}$ $(\times 10^{-3})$	71.37	67.96**	70.96	74.98	74.08	80.69	67.67**	70.50*
$VS_{1.0}$ $(\times 10^{-4})$	22.13	21.11**	22.20	22.90	22.92	24.12	21.02**	21.44**
$VS_{2.0}$ $(\times 10^{-7})$	63.08**	62.48**	64.49	63.69	64.20	63.48	63.17**	62.68**
ES $(\times 10^{-3})$	12.77	12.73**	12.91	13.03	12.91	13.10	12.86	12.98
Interest rate changes								
$VS_{0.5}$	49.64	45.86**	64.41	75.50	66.49	78.22	66.62	77.85
$VS_{1.0}$ $(\times 10^{2})$	7.21	6.41**	9.17	10.66	9.11	10.50	9.15	10.44
$VS_{2.0}$ $(\times 10^{4})$	54.05	58.04	31.58	30.57	30.24	27.64	30.39	27.54**
ES $(\times 10^{-1})$	87.49**	87.89	87.90	89.94	87.61**	89.87	87.59**	89.87
Commodity index returns								
$VS_{0.5}$ $(\times 10)$	22.50	22.66	20.75	21.15	20.98	21.51	20.62**	20.98
$VS_{1.0}$	16.58	16.68	14.80	15.10	14.89	15.26	14.70**	14.98
$VS_{2.0}$ $(\times 10^{-4})$	96.80	96.52	79.10	79.77	79.05	79.68	78.50**	79.11*
ES	33.88	33.89	33.78	33.93	33.73**	33.93	33.83	34.00
Summary	16.7%	50.0%	0.0%	0.0%	16.7%	0.0%	58.3%	41.7%

Average variogram and energy scores for different multivariate models applied to the full out-of-sample period in each of the three multivariate data sets. The variogram scores are based on $p = 0.5$, 1 and 2 and results in each row are multiplied by a relevant power of 10 for ease of presentation, there is no comparison between rows. But within each row we compare the average score across the columns and, as before, depict the lowest score in blue and use * and ** to indicate that the model is in the 90% and 75% MCS, respectively. The bottom row summarizes the percentages each model is in the 90% MCS for all the data.

Table 8
Comparison of multivariate performance over time.

	GARCH		EDF		FQ-B		FQ-A	
	CCC	DCC	250	2000	250	2000	250	2000
Exchange rate returns								
2007 to 2010	1	4	0	3	1	2	1	3
2011 to 2014	1	2	0	1	1	0	1	1
2015 to 2018	1	0	0	3	1	3	4	3
Interest rate changes								
2007 to 2010	0	2	0	0	0	0	3	0
2011 to 2014	1	2	0	0	2	0	1	0
2015 to 2018	0	0	1	0	1	0	3	0
Commodity index returns								
2007 to 2010	0	2	1	0	2	0	4	2
2011 to 2014	1	3	0	0	2	0	2	0
2015 to 2018	0	0	4	1	3	1	3	4

This table lists the number of times each model is included in one of the 90% superior sets for the multivariate scores. Since we consider four different scoring rules, each model can be included at most four times. We again use blue to highlight the most successful model in each row. Column $(*)$ uses the entire out-of-sample periods, while columns (1), (2) and (3) are restricted to the sub-periods March 2007–December 2010, January 2010–December 2014 and January 2015–December 2018, respectively.

to the variogram score with $p = 2$, and GARCH-CCC according to the energy score. For commodity indices, the best models are: FQ-A(250) according to the variogram scores with $p = 0.5$, 1 and 2; and the FQ-B(250) according to the energy score. The reason why different rankings are obtained from different scoring rules applied to diverse data sets may be an interesting subject for further research.

Our empirical results suggest that the FQ-B models perform slightly worse than the FQ-A models. Since the latter is conceptually simpler and much easier to implement, these may be the preferred choice in practical applications. The bagging algorithm used in the FQ-B model is computationally complex, and it should only be applied in practice if it provides a clear improvement on other FQ models.

We conclude that the forecasting performance of latent factor FQ models generally exceeds the static model that is standard in the industry, i.e., historical simulation (the variant represented in this paper uses a Gaussian copula with EDF marginals). The forecasts generated by these FQ models also match or exceed the accuracy of standard dynamic forecasting models, represented here with multivariate GARCH. However, even though we have not taken the most advanced models in the class, the multivariate GARCH models still take over five times longer to optimize, require very large calibration samples, and exhibit difficulties with parameter convergence even

in eight dimensions. By contrast, FQ models scale naturally to high-dimensional systems while retaining very fast calibration times. They are much easier to automate fully than GARCH models. Therefore, they could be very attractive to hedge funds and other high-frequency traders in the industry, who commonly employ algorithms to re-balance portfolios every day.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alexander, C., Kaeck, A., & Sumawong, A. (2019). A parsimonious parametric model for generating margin requirements for futures. *European Journal of Operational Research*, *273*(1), 31–43.

Almeida, C., Ardison, K., Kubudi, D., Simonsen, A., & Vicente, J. (2017). Forecasting bond yields with segmented term structure models. *Journal of Financial Econometrics*, *16*(1), 1–33.

Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, *25*(2), 177–190.

Bao, Y., Lee, T.-H., & Saltoğlu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, *26*(3), 203–225.

Berkowitz, J., Christoffersen, P., & Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, *57*(12), 2213–2227.

Birge, J. (2007). Chapter 20 optimization methods in dynamic portfolio management. *Handbooks in Operations Research and Management Science*, *15*, 845–865.

Boero, G., Smith, J., & Wallis, K. F. (2011). Scoring rules and survey density forecasts. *International Journal of Forecasting*, *27*(2), 379–393.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *The Review of Economics and Statistics*, *72*(3), 498–505.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Bunn, D., Andresen, A., Chen, D., & Westgaard, S. (2016). Analysis and forecasting of electricity price risks with quantile factor models. *Energy Journal*, *37*(1), 101–122.

Cenesizoglu, T., & Timmermann, A. G. (2008). Is the distribution of stock returns predictable? Available at SSRN Abstract=1107185.

Chakraborty, B. (2003). On multivariate quantile regression. *Journal of Statistical Planning and Inference*, *110*(1–2), 109–132.

Chavas, J.-P. (2018). On multivariate quantile regression analysis. *Statistical Methods & Applications*, *27*(3), 365–384.

Chernozhukov, V., Fernández-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, *78*(3), 1093–1125.

Chiang, S., & Tsai, M. (2019). Valuation of an option using non-parametric methods. *Review of Derivatives Research*, *22*(3), 419–447.

Connor, G., Hagmann, M., & Linton, O. (2012). Efficient semiparametric estimation of the fama-french model and extensions. *Econometrica*, *80*(2), 713–754.

Connor, G., & Korajczyk, R. (1993). A test for the number of factors in an approximate factor model. *The Journal of Finance*, *48*(4), 1263–1291.

Cont, R., Deguest, R., & Scandolo, G. (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, *10*(6), 593–606.

Coroneo, L., Giannone, D., & Modugno, M. (2016). Unspanned macroeconomic factors in the yield curve. *Journal of Business & Economic Statistics*, *34*(3), 472–485.

Danielsson, J., James, K., Valenzuela, M., & Zer, I. (2016). Model risk of risk models. *Journal of Financial Stability*, *23*, 79–91.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Diks, C., Panchenko, V., Sokolinskiy, O., & van Dijk, D. (2014). Comparing the accuracy of multivariate density forecasts in selected regions of the copula support. *Journal of Economic Dynamics & Control*, *48*, 79–94.

Diks, C., Panchenko, V., & Van Dijk, D. (2010). Out-of-sample comparison of copula specifications in multivariate density forecasts. *Journal of Economic Dynamics & Control*, *34*(9), 1596–1609.

Duan, J.-C., & Miao, W. (2016). Default correlations and large-portfolio credit analysis. *Journal of Business & Economic Statistics*, *34*(4), 536–546.

Ebens, H., Kotecha, C., Ypsilanti, A., & Reiss, A. (2009). Introducing the multi-asset strategy index. *Journal of Alternative Investments*, *11*(3), 6–25.

Elliott, G., & Timmermann, A. (2016). Forecasting in economics and finance. *Annual Review of Economics*, *8*, 81–110.

Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, *20*(3), 339–350.

Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fritsch, F. N., & Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, *17*(2), 238–246.

Gaglianone, W. P., & Lima, L. R. (2012). Constructing density forecasts from quantile regressions. *Journal of Money, Credit and Banking*, *44*(8), 1589–1607.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, *17*(2), 211.

Grant, A., & Satchell, S. (2020). Investment decisions when utility depends on wealth and other attributes. *Quantitative Finance*, *20*(3), 499–513.

Greenaway-McGrevy, R., Mark, N. C., Sul, D., & Wu, J.-L. (2018). Identifying exchange rate common factors. *International Economic Review*, *59*(4), 2193–2218.

Hagfors, L. I., Paraschiv, F., Molnar, P., & Westgaard, S. (2016). Using quantile regression to analyze the effect of renewables on EEX price formation. *Renewable Energy and Environmental Sustainability*, *1*, 32.

Hallin, M., Paindaveine, D., & Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From L1 optimization to halfspace depth. *The Annals of Statistics*, *38*(2), 635–669.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365–380.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)? *Journal of Applied Econometrics*, *20*(7), 873–889.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, *55*(3), 1263–1295.

Hua, J., & Manzan, S. (2013). Forecasting the return distribution using high-frequency volatility measures. *Journal of Banking & Finance*, *37*(11), 4381–4403.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, *23*(2), 389–416.

Jolliffe, I. T., & Stephenson, D. B. (2003). *A practitioner's guide in atmospheric science*. Wiley, Chichester.

Jordan, A. I., Krüger, F., & Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, *90*(12), 1–37.

Keune, J., Ohlwein, C., & Hense, A. (2014). Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Monthly Weather Review*, *142*(11), 4074–4090.
Koenker, R., & Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica (Pre-1986)*, *50*(1), 43.
Koenker, R., & Bassett, G. (2010). March madness, quantile regression bracketology, and the hayek hypothesis. *Journal of Business & Economic Statistics*, *28*(1), 26–35.
Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.
Koenker, R., Chesher, A., & Jackson, M. (2005). *Quantile Regression*. Econometric Society Monographs.
Koenker, R., & Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, *101*, 980–990.
Kraus, A., & Litzenberger, R. H. (1976). Skewness preference and the valuation of risk assets. *The Journal of Finance*, *31*(4), 1085–1100.
Kuester, K., Mittnik, S., & Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, *4*(1), 53–89.
Laurent, S., Rombouts, J., & Violante, F. (2012). On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics*, *27*(6), 934–955.
Lwin, K., Qu, R., & MacCarthy, B. (2017). Mean-var portfolio optimization: A nonparametric approach. *European Journal of Operational Research*, *260*(2), 751–766.
Ma, L., & Pohlman, L. (2008). Return forecasts and optimal portfolio construction: A quantile regression approach. *The European Journal of Finance*, *14*(5), 409–425.
Machete, R. L. (2013). Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, *143*(10), 1781–1790.
Maciejowska, K., Nowotarski, J., & Weron, R. (2016). Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, *32*(3), 957–965.
Manzan, S. (2015). Forecasting the distribution of economic variables in a data-rich environment. *Journal of Business & Economic Statistics*, *33*(1), 144–164.
Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096.
Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., & Vrontos, S. D. (2019). Quantile forecast combinations in realised volatility prediction. *Journal of the Operational Research Society*, *70*(10), 1720–1733.
Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370.
Nieto, M., & Ruiz, E. (2016). Frontiers in var forecasting and backtesting. *International Journal of Forecasting*, *32*(2), 475–501.
Nowotarski, J., & Weron, R. (2015). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, *30*(3), 791–803.
Orhan, M., & Köksal, B. (2012). A comparison of GARCH models for var estimation. *Expert Systems with Applications*, *39*(3), 3582–3592.
Panagiotelis, A., & Smith, M. (2008). Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions. *International Journal of Forecasting*, *24*(4), 710–727.
Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review*, *47*(2), 527–556.
Patton, A. J. (2009). Copula-based models for financial time series. In T. Andersen, R. Davis, J.-P. Kreiss, & T. Mikosch (Eds.), *Handbook of financial time series* (pp. 767–785). Springer.
Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, *110*, 4–18.
Patton, A. J. (2013). Copula methods for forecasting multivariate time series. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting, Vol. 2* (B), (pp. 899–960). North Holland.
Pedersen, T. Q. (2015). Predictable return distributions. *Journal of Forecasting*, *34*(2), 114–132.
Pérignon, C., & Smith, D. (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance*, *34*(2), 362–377.
Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12–20.
Pristsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance*, *30*(2), 561–582.
Prorokowski, L., & Prorokowski, H. (2014). Comprehensive risk measure – current challenges. *Journal of Financial Regulation and Compliance*, *22*(3), 271–284.
Ravazzolo, F., & Vahey, S. P. (2014). Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics & Econometrics*, *18*(4), 367–381.
Resta, M. (2012). Portfolio optimization: New challenges and perspectives. *Recent Patents on Computer Science*, *5*(1), 59–65.
Rodrigues, T., & Fan, Y. (2017). Regression adjustment for non-crossing Bayesian quantile regression. *Journal of Computational and Graphical Statistics*, *26*(2), 275–284.
Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, *13*(3), 341–360.
Santos, B., & Kneib, T. (2020). Non-crossing structured additive multiple-output Bayesian quantile regression models. *Statistics and Computing*, *30*, 855–869.
Scheller, F., & Auer, B. (2018). How does the choice of value-at-risk estimator influence asset allocation decisions? *Quantitative Finance*, *18*(12), 2005–2022.
Scheuerer, M., & Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, *143*(4), 1321–1334.
Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–61.
Semenov, A. (2008). Historical simulation approach to the estimation of stochastic discount factor models. *Quantitative Finance*, *8*(4), 391–404.
Sheppard, K. (2013). Oxford MFE toolbox. Accessed: 2017-02-07.
Silvennoinen, A., & Terasvirta, T. (2009). Multivariate GARCH models. In T. Andersen, R. Davis, J.-P. Kreiss, & T. Mikosch (Eds.), *Handbook of financial time series* (pp. 201–229). Springer.
Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.
Székely, G. J. (2003). *E-Statistics: the energy of statistical samples. Vol. 3. No. 05* (pp. 1–18). Bowling Green State University, Department of Mathematics and Statistics Technical Report.
Thomann, A. (2021). Multi-asset scenario building for trend-following trading strategies. *Annals of Operations Research*, *299*(1), 293–315.
Tukey, J. W. (1974). Mathematics and the picturing of of data. In B. Sirakov, P. de Souza, & M. Viana (Eds.), *Proceedings of the international congress of mathematicians. Vol. 2* (pp. 523–531). World Scientific.
Wellmann, D., & Trück, S. (2018). Factors of the term structure of sovereign yield spreads. *Journal of International Money and Finance*, *81*, 56–75.
Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1–60.
Zakamulin, V. (2015). A test of covariance-matrix forecasting methods. *Journal of Portfolio Management*, *41*(3), 97–108.
Zhou, R., Li, J.-H., & Pai, J. (2019). Pricing temperature derivatives with a filtered historical simulation approach. *European Journal of Finance*, *25*(15), 1462–1484.
Zhu, M. (2013). Return distribution predictability and its implications for portfolio selection. *International Review of Economics & Finance*, *27*, 209–223.
Zolotko, M., & Okhrin, O. (2014). Modelling the general dependence between commodity forward curves. *Energy Economics*, *43*, 284–296.