risks

MDPI

Article

Multivariate Spectral Backtests of Forecast Distributions under Unknown Dependencies

Janine Balter [1,†] and Alexander J. McNeil [2,*]

[1] Deutsche Bundesbank, 40212 Düsseldorf, Germany; janinba@web.de
[2] School for Business and Society, University of York, York YO10 5DD, UK
* Correspondence: alexander.mcneil@york.ac.uk
† The opinions expressed here are our own and do not reflect the views of the Deutsche Bundesbank or its staff.

Abstract: Under the revised market risk framework of the Basel Committee on Banking Supervision, the model validation regime for internal models now requires that models capture the tail risk in profit-and-loss (P&L) distributions at the trading desk level. We develop multi-desk backtests, which simultaneously test all trading desk models and which exploit all the information available in the presence of an unknown correlation structure between desks. We propose a multi-desk extension of the spectral test of Gordy and McNeil, which allows the evaluation of a model at more than one confidence level and contains a multi-desk value-at-risk (VaR) backtest as a special case. The spectral tests make use of realised probability integral transform values based on estimated P&L distributions for each desk and are more informative and more powerful than simpler tests based on VaR violation indicators. The new backtests are easy to implement with a reasonable running time; in a series of simulation studies, we show that they have good size and power properties.

Keywords: backtesting; risk management; value-at-risk; model validation; Basel regulations

check for updates

Citation: Balter, Janine, and Alexander J. McNeil. 2024. Multivariate Spectral Backtests of Forecast Distributions under Unknown Dependencies. *Risks* 12: 13. https://doi.org/10.3390/risks12010013

Academic Editors: Tak Kuen Ken Siu and Hailiang Yang

Received: 19 December 2023
Revised: 5 January 2024
Accepted: 12 January 2024
Published: 17 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In this paper, we propose a multi-desk extension of the spectral backtest of Gordy and McNeil (2020). Our test is designed to address the requirement that banks should implement backtests of their risk models at the desk level. Although banks are required to report the results of single-desk value-at-risk backtests at the desk level, our proposal is for a test that simultaneously assesses the quality of all desk-level risk models, with respect to outcomes in the tail.

There are a number of potential advantages of our test. First, it may detect deficiencies in desk models that are missed when trading desk data are aggregated into a single portfolio. Aggregation will result in the loss of information and the netting effect across desks will potentially mask situations where an under-estimation of risk in one trading desk model is compensated for by the over-estimation of risk in another trading desk model. Moreover, when desks are tested jointly, we effectively increase the amount of data with respect to a single test at the level of the trading book, which improves the testing power. Furthermore, by embedding our test in the spectral framework, we are able to test the performance of risk models at a range of confidence levels, and not simply at a single level, as in standard VaR backtesting. To place our proposal in context, we give a brief review of the development of backtesting for the trading book in the following paragraphs.

Over the last three decades, a method of testing value-at-risk (VaR) predictions using historical data, known as backtesting (Jorion 2007), has become the industry standard for the validation of internal models of market risk. Under the previous regulatory frameworks of Basel II and Basel II.5 (see BCBS 2006, 2011), backtesting required the comparison of one-day-ahead VaR forecasts at the 99% probability level with the ex-post realised losses and was based on so-called VaR violation indicators. Most proposed backtests for the

accuracy of VaR forecasts are of a univariate nature and are based on univariate time series of VaR violations for a single portfolio, such as the entire trading book of a bank.

Christoffersen (1998) defined criteria, known as the unconditional coverage and independence hypotheses, that should be satisfied by credible VaR forecasts. The former is the requirement that the expected number of violations of VaR forecasts at probability level α over n time periods should be $n(1-\alpha)$, while the latter is the requirement that such violations should occur independently in time. The criteria can also be combined to obtain the conditional coverage hypothesis. An early approach to verifying the unconditional coverage hypothesis can be found in Kupiec (1995), while an approach to testing for the independence of the sequence of VaR violation indicators was presented by Christoffersen (1998). Tests of the independence hypothesis generally require an assumption about the dependence structure under the alternative hypothesis and can have very limited power to detect departures from independence caused by other forms of serial dependence. Moreover, as pointed out in Campbell (2007), joint tests of the unconditional coverage and independence hypothesis are not automatically preferable to separate testing; poorly performing VaR models, which violate only one of the two hypotheses, are less likely to be detected by a joint test than by two separate tests.

Since a reasonable model of a profit-and-loss (P&L) distribution should deliver acceptable VaR estimates at a range of probability levels, the revised regulatory requirements for the measurement of market risk (BCBS 2019)[1] emphasise that backtesting should be carried out at multiple probability levels beyond the usual 99% level. In these new regulations, the expected shortfall risk measure plays an important role in the formula used to determine the required capital for market risk in the trading book. Since the expected shortfall is defined as the mean of the losses that are greater than the VaR at level α, a good estimate of the expected shortfall requires a P&L forecast that is accurate for VaR estimation at a range of α values in the tail of the distribution. Kratz et al. (2018) proposed a simultaneous multinomial test of VaR estimates at different α levels and suggested that such a test could be viewed as an implicit backtest of the expected shortfall. Alternative multilevel VaR tests were also developed by Campbell (2007), who proposed a Pearson's chi-squared test for goodness of fit, and Pérignon and Smith (2008), who developed a likelihood-ratio test generalising the unconditional coverage test of Kupiec (1995).

Testing the accuracy of VaR forecasts at more than one α level implies that we move away from a simple assessment of the validity of VaR to a more thorough assessment of the forecast of the P&L distribution from which the VaR forecast is calculated. If the P&L distribution is adequately estimated, the resulting VaR estimates must be accurate for every α level in $[0,1]$. In the extreme case that we test at every α level, we backtest the complete P&L distribution. One advantage of backtesting the forecast of the P&L distribution (or a region theoreof) is that we exploit much more information in comparison to using the series of VaR violation indicators at a single α level. The latter may take only the values one or zero, depending on whether a VaR violation has occurred or not, and, for typical α levels in the region of 99%, violations are rare and the resulting indicator data are sparse.

Backtests of the forecast of the P&L distribution can be based on realised probability-integral transform (PIT) values. These are transformations of the realised value of P&L by the cumulative distribution function of the model used to forecast P&L at the previous time point. An ideal forecaster, i.e., a forecaster in the sense of Gneiting et al. (2007) who has knowledge of the correct model, would produce independent and uniformly distributed PIT values. Berkowitz (2001) proposed the first backtest of this kind based on the transformation of realised PIT values to a normal distribution under the null hypothesis.

The spectral tests of Gordy and McNeil (2020) are constructed by transforming PIT values with a weighting function, referred to as a kernel, and are available in both unconditional coverage and conditional coverage variants. Their form is very flexible and they include many of the previously proposed approaches to backtesting as special cases. Under the spectral backtesting philosophy, the risk modelling group or banking regulator can choose the kernel to apply weight to the area of the forecast distribution that is of

primary interest—for example, a region in the area around the 99% quantile. Among the tests subsumed in this framework are the test of Kupiec (1995), which corresponds to a kernel equivalent to the Dirac measure at a specific probability level α, and the test of Berkowitz (2001). The spectral risk measure test of Costanzino and Curran (2015) and the expected shortfall test of Du and Escanciano (2017) are further special cases obtained by choosing a kernel truncated to tail probabilities.

While the test of Gordy and McNeil (2020) can be viewed as an absolute test of model adequacy for a single candidate model, a weighted approach to forecast comparison in specific areas of a forecast distribution was proposed by Amisano and Giacomini (2007). Their weighted likelihood ratio test is a relative test that compares the performance of two competing density forecasts and is based on the weighted averages of the logarithmic scoring rule, where the weights are chosen according to the preferences of the risk modelling group. Diks et al. (2011) proposed similar tests but based them on conditional likelihood and censored likelihood scoring rules, while Gneiting and Ranjan (2011) applied the (quantile-weighted) continuous ranked probability score instead of the logarithmic score.

In contrast to the rich variety of tests available for the backtesting of a univariate series of density forecasts, or comparing competing sets of forecasts, the literature on multivariate backtesting is much more sparse. Multivariate extensions of univariate VaR backtests have not been widely developed, although Berkowitz et al. (2011) already sketched a number of ideas. We are aware of two papers by Danciulescu (2016) and Wied et al. (2016) that deal with the multivariate backtesting of VaR, in the sense of backtesting VaR forecasts for several portfolios (or desks) simultaneously.

Danciulescu (2016) suggests a test for the unconditional coverage hypothesis and a test for the independence hypothesis using multivariate portmanteau test statistics of the Ljung–Box type, which are applied to the multivariate time series of VaR violation indicator variables. The test for the independence hypothesis simultaneously tests for the absence of cross- and autocorrelations in the multivariate time series of VaR violation indicator variables up to some finite lag K. In Wied et al. (2016), multivariate tests are proposed for the detection of clustered VaR violations, which would violate the conditional coverage hypothesis. They argue that their test can detect the clustering of VaR exceedances for a single desk, which would indicate that the probability of VaR violations is varying over time, as well as the clustering of VaR exceedances across desks at different lags, which would cast doubt on the assumption of independent VaR violations for different desks at different time points.

Unfortunately, due to the sensitivity of the data concerned, there are very few empirical studies of bank-wide P&L backtesting and even fewer studies of desk-level data. One exception is Berkowitz et al. (2011), who analysed daily realisations from the P&L distribution and daily forecasts of VaR (calculated by the widely used historical simulation method) for each of four separate business lines at a large international commercial bank. In this study, various univariate tests, including the Markov test of Christoffersen (1998) for the conditional coverage hypothesis, the CaViaR test for autocorrelation of Engle and Manganelli (2004), and the unconditional coverage hypothesis test of Kupiec (1995), were used to backtest the VaR for each trading desk separately. Moreover, the tests were assessed based on their finite sample size and power properties in a Monte Carlo study. The authors found that the VaR models for two out of the four business lines were rejected due to volatility clustering and that the model of a third business line was rejected by the unconditional coverage hypothesis test of Kupiec (1995).

The remainder of our paper is organised as follows. In Section 2, we explain the testing approach based on PIT values and recapitulate the main details of the spectral test of Gordy and McNeil (2020). We then show how this may be extended to obtain multivariate spectral tests based on a single spectrum or multiple spectra in Section 3. In Section 4, we carry out a simulation study to analyse the size and power of the proposed tests for different amounts of data, different numbers of desks, different choices of kernels, and different deviations from the null hypothesis. Concluding remarks are found in Section 5.

2. Spectral Backtests of Forecast Distributions

2.1. Realised PIT Values

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space and let $(L_t)_{t\in\mathbb{N}}$ be a time series defined on this space representing losses (positive numbers) and profits (negative numbers) on a portfolio of risky assets. For every $t \in \mathbb{N}$, define the conditional loss distribution function F_t by

$$F_t(x_t \mid x_{t-1}, \ldots, x_1) = \mathbb{P}(L_t \leq x_t \mid L_{t-1} = x_{t-1}, \ldots, L_1 = x_1)$$

and assume that F_t is an absolutely continuous function of x_t. If we now define the process $(U_t)_{t\in\mathbb{N}}$ by $U_1 = F_1(L_1)$ and $U_t = F_t(L_t \mid L_{t-1}, \ldots, L_1)$ for $t \geq 2$, then the result of Rosenblatt (1952) implies that $(U_t)_{t\in\mathbb{N}}$ is a sequence of independent and identically distributed (iid) standard uniform variables. Note that this holds in general without any assumption of stationarity for $(L_t)_{t\in\mathbb{N}}$. This is important because the time-varying nature of the functions F_t may come both from the time-varying nature of the underlying risk factors affecting the portfolio value (for example, equity prices and interest rates) as well as changes in the composition of the portfolio caused by rebalancing and capital inflows and outflows.

We assume that, at each time t, a risk modelling group at a bank or other financial institution makes a forecast $\widehat{F}_t$ of F_t based on the information available to it up to time $t-1$, which includes $\{L_1, \ldots, L_{t-1}\}$ and possibly additional information. The corresponding random variables P_t obtained by setting $P_t = \widehat{F}_t(L_t \mid L_{t-1}, \ldots, L_1)$ for $t \in \mathbb{N}$ are referred to as the realised PIT values. For an omniscient or ideal forecaster in the sense of Gneiting et al. (2007), which is a forecaster who possesses extra information about the exact form of the function F_t at every time point, it would follow that $\widehat{F}_t = F_t$ and the realised PIT values would be iid uniform. For an ordinary mortal forecaster, we effectively hold them to this ideal standard and test the extent to which the realised PIT values satisfy the iid uniform assumption.

We assume that model validation is carried out by a regulator who only has access to the time series of realised PIT values (P_t) and has no knowledge of the forecasting models $(\widehat{F}_t)$ that were used.

2.2. Spectral Tests

The spectral tests of Gordy and McNeil (2020) are based on transformations of the realised PIT values through the level exceedance indicator function $Y_t(u) = I_{\{P_t \geq u\}}$ according to

$$W_t = \int_{[0,1]} I_{\{P_t \geq u\}} \mathrm{d}\nu(u), \; t = 1, \ldots, n \tag{1}$$

where ν is a probability measure on $[0,1]$, which is referred to as the kernel measure.

The risk modelling group or model validation group can choose this measure to weight a set of quantile levels according to their preferences for forecast model performance. For example, if the preference is for model performance around the 99th percentile, the kernel measure would typically be chosen to place weight in a sub-interval of $[0,1]$ containing the value 0.99. In the most extreme case, by choosing the Dirac measure at the value $\alpha = 0.99$, the spectral transformation of the realised PIT value in (1) would simply satisfy

$$W_t = Y_t(\alpha) = I_{\{\widehat{F}_t(L_t) \geq \alpha\}} = I_{\{L_t \geq \widehat{F}_t^{-1}(\alpha)\}}.$$

In other words, it would yield the indicator variable for a VaR exceedance at level α. The paper of Gordy and McNeil (2020) contains examples of discrete and continuous probability measures ν. In the continuous case, we have $W_t = \int_0^1 g(u) I_{\{P_t \geq u\}} \mathrm{d}u$, where g is the probability density corresponding to the measure ν; in this case, we usually assume that ν is supported on a strict sub-interval $[\alpha_1, \alpha_2]$ of $[0,1]$, which is referred to as the kernel window.

Monospectral tests are based on the time series $(W_t)_{t\in\mathbb{N}}$. Gordy and McNeil (2020) also propose bispectral and, more generally, m-spectral tests in which a set of distinct kernel measures $\nu_1, \ldots, \nu_m$ is considered. In this case, the data consist of vectors $(\boldsymbol{W}_t)$, where $\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,m})'$ and each component $W_{t,j}$ is calculated using the probability measure ν_j and a spectral transformation of the form given in (1).

We can consider hypotheses about the series $(\boldsymbol{W}_t)$ that generalise the notions of unconditional coverage, independence, and conditional coverage promoted by Christoffersen (1998): a test of the unconditional coverage hypothesis is a test that the distribution of the $\boldsymbol{W}_t$ values is the one implied by the uniformity of the realised PIT values; an independence test is a test of the hypothesis that the $\boldsymbol{W}_t$ values are independent of lagged realised PIT values $P_1, \ldots, P_{t-1}$ (which implies that the series $(\boldsymbol{W}_t)$ is iid). A test of the conditional coverage hypothesis is a combined test of both of these hypotheses.

In this paper, our focus is on unconditional tests. The generalisation of unconditional spectral tests to conditional spectral tests is extensively discussed in Gordy and McNeil (2020) and it is clear from the analysis in that paper that conditional extensions are more powerful at detecting the presence of serial dependence in PIT values—for example, serial dependence resulting from unmodelled volatility effects. In this paper, we concentrate on tests that address the hypothesis of the uniformity of PIT values and how these may be generalised to the multi-desk case.

Gordy and McNeil (2020) consider different styles of test based on data $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ for the unconditional coverage hypothesis, including likelihood ratio tests and Z-tests. We will extend the latter to the multi-desk situation due to their tractability, ease of implementation, and good performance in the single portfolio setting. A Z-test appeals to the asymptotic normality of the sample average $\overline{\boldsymbol{W}}_n = n^{-1}\sum_{t=1}^{n}\boldsymbol{W}_t$ and tests whether

$$T_n = n\big(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\big)'\Sigma_W^{-1}\big(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\big) \sim \chi^2_m \tag{2}$$

where $\boldsymbol{\mu}_W$ and Σ_W are the mean vector and covariance matrix of $\boldsymbol{W}_t$ when the underlying realised PIT value P_t is standard uniform. In the case of a monospectral test, we consider the simpler test of whether the following holds:

$$\tilde{T}_n = \frac{\sqrt{n}(\overline{W}_n - \mu_W)}{\sigma_W} \sim N(0,1). \tag{3}$$

In the case where the probability measures ν_j have absolutely continuous distribution functions G_j with densities g_j, we obtain simple formulas for the moments of $W_{t,j} = \int_0^1 g_j(u)I_{\{P_t\geq u\}}\mathrm{d}u$ taking the form

$$\mu_{Wj} = \mathbb{E}(W_{t,j}) = \int_0^1 g_j(u)(1-u)\mathrm{d}u \tag{4}$$

$$\Sigma_{Wj,k} = \mathbb{E}(W_{t,j}W_{t,k}) = \int_0^1 \big(g_j(u)G_k(u) + g_k(u)G_j(u)\big)(1-u)\mathrm{d}u \tag{5}$$

and when $m = 1$, these reduce to $\mu_W = \mu_{W1}$ and $\sigma_W^2 = \sigma_{W1}^2 = \Sigma_{W1,1}$. This is the case that we will extend in this paper. The main practical consideration is that the density functions g_j should permit the moments in (4) and (5) to be easily evaluated. We will consider parametric densities g_j supported on an interval $[\alpha_1, \alpha_2]$ in the neighbourhood of a standard VaR confidence level ($\alpha = 0.99$) such that g_j is either constant in the interval or increasing, to give greater weight to more extreme quantiles; the exact forms used may be found in Section 4.1. This approach is shown to work well in Gordy and McNeil (2020).

2.3. Monospectral versus Bispectral Tests

Monospectral tests can be useful in carrying out one-sided tests of the null hypothesis that $\mathbb{E}(W_t) \leq \mu_W$ (where μ_W is the value obtained when PITs are uniform) against the alternative that $\mathbb{E}(W_t) > \mu_W$; in the case of a VaR exception test, this amounts to testing

that $\mathbb{E}(Y_t(\alpha)) \leq 1-\alpha$ versus $\mathbb{E}(Y_t(\alpha)) > 1-\alpha$ or, in other words, the null hypothesis that the VaR exceedance probability is no larger than the desired value $(1-\alpha)$.

Bispectral or multispectral tests are two-sided tests and, in our view, should generally be preferred to the one-sided monospectral tests. In assessing the quality of forecasts, the systematic overestimation of risk is as suggestive of a deficiency in the forecast model as is systematic underestimation. Moreover, at the desk level, if the net position in some risky asset class changes from long to short, then a different tail of the P&L distribution becomes relevant and an overestimate of risk may become an underestimate.

One means of understanding the difference between a monospectral and a bispectral test is that the former only tests for the consistency of weighted PIT values with the first moment of W_t under the null hypothesis, while the latter tests for the consistency of two different statistical functionals and thus effectively captures the misspecification of two moments. This can be particularly beneficial when one of the chosen kernels fails to detect deviations of the distribution of the realised PIT values from a uniform distribution. Spectral tests are weighted measures of the discrepancy between the empirical distribution of PIT values and the uniform distribution within the kernel window. It can emerge that this discrepancy results in the crossing of the distribution functions and that this crossing is undetected by a single weighting scheme but is revealed by using two such schemes. Further discussion of this phenomenon is found in Gordy and McNeil (2020).

3. Multi-Desk Spectral Backtests

3.1. Framework for Multi-Desk Backtests

We generalise the notation to consider d desks or sub-portfolios. For $i = 1, \ldots, d$, let $L_{t,i}$ represent the loss attributed to sub-portfolio or trading desk i at time t, let $\widehat{F}_{t,i}$ denote the forecast model for $L_{t,i}$ constructed using information up to time $t-1$, and let $P_{t,i} = \widehat{F}_{t,i}(L_{t,i} \mid L_{t-1,i}, \ldots, L_{1,i})$ be the corresponding realised PIT value.

If the risk management function of desk or sub-portfolio i is modelling its risks well, we expect that the univariate PIT time series $(P_{t,i})$ will behave like an iid uniform series. We expect cross-correlation between the series at lag zero, since portfolio or desk models are usually built using the same underlying methodology and the same correlated risk factor data, but we do not expect cross-correlation at other lags. This leads us to adopt the hypothesis that the PIT vectors $(\boldsymbol{P}_t)$ given by $\boldsymbol{P}_t = (P_{t,1}, \ldots P_{t,d})'$ are iid random vectors with uniform marginal distributions.

We first note that a simple but crude method of extending the single-portfolio spectral testing methodology to a multi-desk situation is through the Bonferroni correction (see, for example, Dunn 1959). Suppose that we calculate either monospectral test statistics $\tilde{T}_{n,i}$ based on (3) or multispectral test statistics $T_{n,i}$ based on (2) for each desk i. Let the p-values for the d desks be $p_1, \ldots, p_d$. The null hypothesis that the desks are collectively delivering acceptable forecasts of P&L would be rejected in a test of size no greater than β if $\min\{p_1, \ldots, p_d\} \leq \beta/d$. However, it is well known that the Bonferroni correction leads to a loss of statistical power and so this is not a recommended procedure in the backtesting context.

We now consider how the spectral methodology can be adapted to construct a joint test over all desks. In the following sections, we use the notation

$$W_{t,i,j} = \int_{[0,1]} I_{\{P_{t,i} \geq u\}} \mathrm{d}\nu_j(u)$$

to refer to the spectral-transformed realised PIT value for desk i at time t using kernel ν_j.

3.2. The Monospectral Case

We first consider a single spectral transformation of the PIT values of the form $W_{t,i,1} = \int_{[0,1]} I_{\{P_{t,i} \geq u\}} \mathrm{d}\nu(u)$ using the spectrum ν, which leads to a multi-desk generalisation of the test in (3). Under the assumption of uniform PIT values, each of the $W_{t,i,1}$ satisfies $\mathbb{E}(W_{t,i,1}) = \mu_W$ and $\mathrm{var}(W_{t,i,1}) = \sigma_W^2$, where these moments take the same values as in (3).

Let $\overline{W}_n = n^{-1}\sum_{t=1}^{n} Z_t$, where $Z_t = d^{-1}\sum_{i=1}^{d} W_{t,i,1}$ is the average of the spectrally transformed, realised PIT values over all desks at time t. We note that this is a generalisation of the notation used in (3) since, when $d = 1$, we have $\overline{W}_n = n^{-1}\sum_{t=1}^{n} W_{t,1,1}$, which is the average weighted realised PIT value over time for a single desk and a single spectrum. Under the uniform assumption, $\mathbb{E}(\overline{W}_n) = \mathbb{E}(Z_t) = \mathbb{E}(W_{t,i,1}) = \mu_W$. However, the variance of $\overline{W}_n$ satisfies $\mathrm{var}(\overline{W}_n) = n^{-2}\sigma_Z^2$, where $\sigma_Z^2 = \mathrm{var}(Z_t)$ and $\sigma_Z \neq \sigma_W$ unless $d = 1$; in general,the unknown dependencies between desks must be considered in evaluating σ_Z. Thus, the monospectral multi-desk test relies on generalising (3) to test whether

$$\tilde{T}_n = \frac{\sqrt{n}(\overline{W}_n - \mu_W)}{\sigma_Z} \sim N(0,1) \tag{6}$$

where σ_Z is a parameter that is not fully identified under the null hypothesis and must be replaced by a suitable estimate.

A natural first approach is to use the unbiased and consistent estimator given by $\hat{\sigma}_Z^2 = \frac{1}{n-1}\sum_{t=1}^{n}(Z_t - \overline{W}_n)^2$, but this does not work well in practice. Occasional large values of Z_t have a tendency to inflate $\hat{\sigma}_Z^2$ and reduce the size of the test statistic in (6), leading to the undersizing of the test. We have explored some alternative estimators that are better at controlling the size, and the best-performing estimator is based on a method that will be referred to as the correlation estimation (CE) method in the results. The CE method is based on the observation that σ_Z^2 may be written as

$$\sigma_Z^2 = \mathrm{var}(Z_t) = \frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\mathrm{cov}(W_{t,i,1}, W_{t,j,1}) = \frac{\sigma_W^2}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\rho_{ij}$$

where ρ_{ij} denotes the correlation between the spectrally transformed realised PIT values for desks i and j. If $\widehat{R}_{\boldsymbol{W}}$ denotes the correlation matrix of the data $\{(W_{t,1},\ldots,W_{t,d})',\ t = 1,\ldots,n\}$, we can use the consistent estimator

$$\hat{\sigma}_Z^2 = \frac{\sigma_W^2}{d^2}\mathbf{1}'\widehat{R}_{\boldsymbol{W}}\mathbf{1}. \tag{7}$$

The intuition is that by estimating a bounded quantity like correlation empirically while using the values of σ_W^2 derived under the null hypothesis, we obtain an estimator that is relatively robust to outlying observations of $(W_{t,1},\ldots,W_{t,d})'$, or, in other words, observations where multiple desks show extreme PIT values. In contrast, if we estimate empirical variances and covariances, we observe the tendency for inflated estimates to dilute the size and power of the test.

In general, we expect positive dependence between desks. The above estimator can occasionally give realised values that point to negative dependence, even in simulation experiments where we know that there is a positive correlation between desks. We find that better results are obtained by setting a floor for the estimator at the value that would be obtained when the desks are independent, i.e., $\sigma_Z^2 = \sigma_W^2/d$. Thus, we work, in practice, with the estimator $\tilde{\sigma}_Z^2 = \max(\sigma_W^2/d, \hat{\sigma}_Z^2)$, where $\hat{\sigma}_Z^2$ is given by (7).

3.3. The Bispectral Case

Recall from Section 2.2 that a multispectral test is based on m probability measures $\nu_1,\ldots,\nu_m$. In this section, we will consider $m = 2$ continuous measures with densities g_1 and g_2 so that, for each desk i and time t, we have a pair of spectrally transformed PIT values given by $(W_{t,i,1}, W_{t,i,2})$. Let $\boldsymbol{Z}_t$ be given by $\boldsymbol{Z}_t = (Z_{t,1}, Z_{t,2})'$, where $Z_{t,j} = \frac{1}{d}\sum_{i=1}^{d} W_{t,i,j}$ is the desk average for measure j and $j = 1, 2$.

Consider the statistic $\overline{\boldsymbol{W}}_n = n^{-1}\sum_{t=1}^{n}\boldsymbol{Z}_t$, which generalises the statistic used in (2) to the case of more than one desk. The mean of $\overline{\boldsymbol{W}}_n$ under the null hypothesis is given by

$$\boldsymbol{\mu}_W = \mathbb{E}(\overline{\boldsymbol{W}}_n) = \mathbb{E}(\boldsymbol{Z}_t) = (\mu_{W1}, \mu_{W2})'$$

where the μ_{Wj} are exactly as calculated in (4). The covariance matrix of $\overline{\boldsymbol{W}}_n$ is given by $\text{var}(\overline{\boldsymbol{W}}_n) = n^{-1}\Sigma_Z$, where

$$\Sigma_Z = \begin{pmatrix} \text{var}(Z_{t,1}) & \text{cov}(Z_{t,1}, Z_{t,2}) \\ \text{cov}(Z_{t,1}, Z_{t,1}) & \text{var}(Z_{t,2}) \end{pmatrix}. \tag{8}$$

The application of the multivariate Central Limit Theorem Van der Vaart (2000) leads to a test of the form

$$T_n = n\big(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\big)'\Sigma_Z^{-1}\big(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\big) \sim \chi^2_2\,. \tag{9}$$

Note that in the case of a single desk ($d = 1$), we have $\Sigma_Z = \Sigma_W$ and we recover the test statistic T_n given in (2) in Section 2.2.

As in the case of the monospectral test, we have to consider how to estimate Σ_Z under the null hypothesis when we have no information about dependencies across desks. We use a method that is the natural extension of CE. Specifically, we decompose the variances and covariances in (8) into a variance part, which is known under the null hypothesis, and a part depending on the unknown correlation across desks.

Let $\boldsymbol{V}_{t,j} = (W_{t,1,j}, \ldots W_{t,d,j})'$ denote the vector of W-values across all d desks at time t under weighting function g_j, for $j = 1, 2$. Since $Z_{t,j} = d^{-1}\mathbf{1}'\boldsymbol{V}_{t,j}$, we have

$$\begin{aligned} \text{var}(Z_{t,j}) &= \frac{1}{d^2}\mathbf{1}'\,\text{var}(\boldsymbol{V}_{t,j})\mathbf{1} = \frac{\sigma^2_{Wj}}{d^2}\mathbf{1}'R_{\boldsymbol{V}_j}\mathbf{1}, \quad j = 1, 2, \\ \text{cov}(Z_{t,1}, Z_{t,2}) &= \frac{1}{d^2}\mathbf{1}'\,\text{cov}(\boldsymbol{V}_{t,1}, \boldsymbol{V}_{t,2})\mathbf{1} = \frac{\sqrt{\sigma^2_{W1}\sigma^2_{W2}}}{d^2}\mathbf{1}'R_{\boldsymbol{V}_1\boldsymbol{V}_2}\mathbf{1} \end{aligned} \tag{10}$$

where $\sigma^2_{Wj} = \Sigma_{Wj,j}$ with $\Sigma_{Wj,j}$ as defined in (5) and where $R_{\boldsymbol{V}_j}$ and $R_{\boldsymbol{V}_1\boldsymbol{V}_2}$ denote correlation matrices. More specifically, $R_{\boldsymbol{V}_j}$ is the $d \times d$ correlation matrix for the W-values under measure v_j across all d desks; this is estimated by $\widehat{R}_{\boldsymbol{V}_j}$, the empirical correlation matrix of the vectors $\{\boldsymbol{V}_{t,j},\ t = 1, \ldots, n\}$. The matrix $R_{\boldsymbol{V}_1\boldsymbol{V}_2}$ is the $d \times d$ correlation matrix whose element (j, k) is the correlation between the W-value for desk j under measure 1 and desk k under measure 2; this is estimated by $\widehat{R}_{\boldsymbol{V}_1\boldsymbol{V}_2}$, the matrix whose element (j, k) is the sample correlation of the pairs $\big\{(W_{t,j,1}, W_{t,k,2}),\ t = 1, \ldots, n\big\}$.

4. Simulation Study

4.1. Design of the Study

We vary the following variables in the simulation study.

- Sample size n. This corresponds to the number of days used in the bank's backtesting exercise. The length of the backtesting period is typically small (one or two years of daily data corresponding to days on which markets are open) and so we consider $n = 250$ and $n = 500$.
- Number of desks d. This can be quite large in a bank with extensive trading operations and we consider the values $d = 50$ and $d = 100$.
- Copula C of PIT values across desks. We assume different dependence structures across desks by sampling PIT values with different copulas. In particular, we use the Gauss copula and the copula of a multivariate t distribution with 4 degrees of freedom. The latter case allows us to see how the properties of the spectral tests are affected by tail dependencies in the PIT data.
- Level of dependence ρ across desks. For simplicity, we assume that all desks are equi-dependent by setting the correlation matrix R of the Gauss and t copulas to be an equicorrelation matrix with common parameter ρ, which takes the values $\rho = 0$ or $\rho = 0.5$; more details are given below.

- Fraction of misspecified desks. In examining size and power, we vary the fraction of desks that use misspecified P&L distributions in their risk models.
- Specification of spectral test. We use a number of different monospectral and bispectral tests as detailed below.

We generate n independent realisations of the vector $\boldsymbol{P_t} = (P_{t,1}, \ldots, P_{t,d})$ representing the PIT values at time t for desks $i = 1, \ldots, d$. The vectors $\boldsymbol{P_t}$ are drawn from the distribution

$$\boldsymbol{P_t} \sim C(G_1(u_1), \ldots, G_d(u_d)), \tag{11}$$

where C denotes the copula (Gauss or Student) and $G_1, \ldots, G_g$ are continuous univariate distribution functions, which are designed to capture the effects of correctly and incorrectly specified desk-level P&L models. The correlation matrix R, which is used to parameterise the Gauss copula and the Student t4 copula, is an equicorrelation matrix with parameter $\rho = 0$ or $\rho = 0.5$. Note that, in the former case, the Gauss copula yields a model where the simulated PIT values are independent across desks, while the t4 copula yields a model with dependencies. This is because, even with a correlation matrix R equal to the identity, the Student t4 copula still has tail dependence, which will tend to lead to very large or very small PIT values occurring together across a number of desks; see, for example, McNeil et al. (2015)[2].

For the marginal distributions G_i, we use a construction first developed in Kratz et al. (2018) and also used in Gordy and McNeil (2020). We set $G_i(u) = \Phi(F_i^{-1}(u))$, where Φ is the standard normal distribution function and where F_i is either the standard normal distribution or the distribution function of a univariate Student t4 distribution scaled to have variance one. If we wish to mimic a desk that is using a correctly specified desk model, we choose the normal and thus obtain $G_i(u) = u$, the distribution function of standard uniform. If we wish to mimic a desk that is using an incorrectly specified model, we choose the scaled t4 and obtain a distribution function G_i, which is supported on the unit interval $[0, 1]$ but is not uniform; on the contrary, it is the type of PIT value distribution that would be obtained if the desk were using a model (represented by Φ) that was lighter-tailed than the true P&L distribution (represented by F_i) and thus underestimated the potential for large losses (and large gains). It is important to note that the use of these two distributions is simply a device to generate PIT data that either satisfy or violate the null hypothesis; we do not claim that these distributions are in any sense the true distributions. Recall that the null hypothesis is that the PIT vectors $(\boldsymbol{P_t})$ are iid random vectors with uniform marginal distributions but an unknown dependence structure.

For the spectral tests, we selected kernels corresponding to the following three continuous weighting functions g defined on $[\alpha_1, \alpha_2]$:

(1) The uniform weighting function $g(u) = 1$;
(2) The linear weighting $g(u) = u$;
(3) The exponential weighting function $g(u) = \exp(k(u - \alpha_1)) - 1$ with $k = 1$.

The values α_1 and α_2 determine the kernel window. We choose $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$, which gives a symmetric interval around 0.99. Note that the functions are non-decreasing, placing more weight on more extreme outcomes (in the right tail).

For the multivariate monospectral Z-tests, the kernel functions listed above lead to three different tests, which we denote, respectively, by SP.U, SP.L, and SP.E. For the multivariate bispectral Z-tests, we will look at two different Z-tests, which combine the continuous kernel functions listed above. The test denoted SP.UL combines the uniform and linear weighting functions. The test denoted SP.UE combines the uniform and exponential weighting functions.

The simulation experiments described in the following sections were all performed using the R package `simsalapar`, which is a very flexible tool for conducting large-scale studies with a number of different dimensions (see Hofert and Mächler 2016). The tables that we provide show observed rejection rates for the null hypothesis in 1000 replications of the simulation experiment.

We use a colouring convention to help with the interpretation of these rejection rates and this is applied differently according to whether the results address the size or power of a test. Simulation results relating to size are colour-coded as follows: good results (observed size smaller or equal 6%) are coloured green; poor results (observed size in range [9–12%]) are coloured pink; very poor results (size above 12) are coloured red; all other values are uncoloured. Simulation results relating to power are colour-coded as follows: good results (observed power above 70%) are coloured green; poor results (observed power in range [30–10%]) are coloured pink; very poor results (power below 10%) are coloured red; all other values are uncoloured.

4.2. Evaluating the CE Method of Correcting for Inter-Desk Dependence

We first investigate the crucial CE method of correcting for the unknown dependence structure across desks. We consider two extreme situations—one in which all desks are correctly specified and one in which all desks are incorrectly specified. The former situation allows us to evaluate the size of the test, i.e., the probability of a significant test result when the null hypothesis holds. The latter situation is one that should certainly be picked up by any backtest with reasonable power.

Results for the monospectral tests are found in Table 1. These relate to one-sided tests of the null hypothesis, where we are interested in being able to detect the systematic underestimation of tail risk in the right tail of the loss distribution. The nominal level β of the tests is 0.05. The table shows the actual test rejection rates over 1000 replications for different backtest lengths n, desk numbers d, copulas C, distribution functions F_i, and correlation values ρ. The field CE shows whether the CE correction method has been used or not.

Table 1. Size and power of monospectral Z-tests when all desk models are correctly specified and when all desk models are misspecified.

			d	50				100			
			ρ	0		0.5		0		0.5	
Test	**CE**	n	$F_i\|C$	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**
SP.U	No	250	N	5.0	23.6	26.2	29.2	5.1	29.6	30.5	32.9
			t4	100.0	93.6	89.8	79.0	100.0	96.4	92.4	82.4
		500	N	5.1	24.7	27.5	32.9	5.4	29.4	32.5	36.7
			t4	100.0	99.4	97.7	91.1	100.0	99.7	98.6	93.4
	Yes	250	N	4.6	4.4	4.2	4.5	4.9	4.9	4.4	4.6
			t4	100.0	80.0	65.7	41.7	100.0	80.0	67.2	42.9
		500	N	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	100.0	95.9	87.3	65.8	100.0	97.3	88.9	66.3
SP.L	No	250	N	4.6	22.7	25.5	28.9	5.1	28.5	29.3	32.1
			t4	100.0	98.7	96.4	90.2	100.0	99.2	97.8	92.2
		500	N	5.1	25.5	26.6	34.2	5.5	30.4	32.4	37.7
			t4	100.0	100.0	99.7	97.4	100.0	100.0	99.9	98.4
	Yes	250	N	4.3	4.6	4.2	5.4	4.8	5.1	4.3	5.2
			t4	100.0	94.3	88.9	62.8	100.0	96.4	90.1	64.5
		500	N	5.0	3.3	5.1	5.0	5.5	3.5	4.6	5.2
			t4	100.0	99.9	98.1	86.0	100.0	99.8	98.7	86.3
SP.E	No	250	N	5.0	23.6	26.2	29.2	5.1	29.6	30.5	32.9
			t4	100.0	93.6	90.0	79.0	100.0	96.5	92.5	82.5
		500	N	5.1	24.7	27.5	32.9	5.4	29.4	32.5	36.8
			t4	100.0	99.4	97.7	91.3	100.0	99.7	98.6	93.5
	Yes	250	N	4.6	4.4	4.2	4.4	4.9	4.9	4.4	4.6
			t4	100.0	80.2	66.1	42.0	100.0	80.3	67.8	43.0
		500	N	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	100.0	96.1	87.4	66.1	100.0	97.5	89.1	66.4

The rows in which F_i is recorded as "N" address the question of size and we expect values close to the nominal level of the test 0.05. However, when the CE correction method is not implemented, it is clear that the spectral tests are oversized in all cases except where the desks are independent, which is the column corresponding to a Gauss copula and $\rho = 0$; in the case of a t4 copula with $\rho = 0$, the desks are still dependent and the tests are oversized. In the absence of correction for correlation, there are a number of results coloured red, indicating a complete inability to control the size. In contrast, when the CE correction is implemented, the results are coloured green in all cases.

The rows in which F_i is recorded as "t4" address the question of power, since all desks are misspecified. For each of the spectral tests, the power increases with both n and d, as we would expect. However, the power decreases with strengthening dependence across desks; the power for the case C = t4 and $\rho = 0$ is greater than for C = Gauss and $\rho = 0.5$, which in turn is greater than for C = t4 and $\rho = 0.5$. Increasing levels of dependence can be thought of as effectively reducing the number of independent desk results. Turning to the different spectral tests, the performance shows similar patterns but the SP.L kernel (linear weighting function) seems to give the highest power in this case.

In view of this first set of results, we will apply the CE correction method to the spectral tests in all further experiments.

4.3. Size and Power of Bispectral Tests

We now consider the two bispectral tests under the two scenarios of Table 1—all desks correctly specified and all desks misspecified. For the bispectral test, we also add results for $n = 1000$ backtests, corresponding to 4 years of data. Results are shown in Table 2. Note that the bispectral tests are two-sided tests.

Table 2. Size and power of bispectral Z-tests when all desk models are correctly specified and when all desk models are misspecified.

		d	50				100			
		ρ	0		0.5		0		0.5	
Test	n	$F_i \vert C$	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**
SP.UL	250	N	5.8	9.4	11.6	18.6	4.1	12.3	12.2	21.0
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	N	2.7	5.4	5.0	10.2	3.1	7.2	6.6	10.0
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1000	N	3.0	3.1	3.1	4.8	2.3	4.5	4.6	5.5
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SP.UE	250	N	6.6	11.2	12.7	20.2	4.9	14.0	14.5	22.4
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	N	3.0	6.2	5.4	10.6	3.6	7.7	6.8	10.7
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1000	N	3.1	3.0	3.3	5.2	2.5	4.7	4.8	6.0
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

While the power of these tests is perfect, the size properties are not as good as for the monospectral tests. This is particularly apparent in the case of a backtest of length $n = 250$; the situation improves for $n = 500$ and the size results are very good for $n = 1000$; if anything, there is evidence of undersizing. We interpret these results as showing that more data are required in order for the estimators of (10) to give an accurate correction for the unknown desk dependence structure in the bispectral case.

4.4. Evaluating the Effect of Desk Misspecification Rate

We now consider the more realistic situation where only a certain fraction of desks are incorrectly specified. We choose values of this fraction in the set of 25% and 50%. All the misspecified desks use the model in which F_i is t4. Results for monospectral tests are shown in Table 3.

Table 3. Power of monospectral Z-tests when a certain proportion of desks is misspecified.

		d	50				100			
		ρ	0		0.5		0		0.5	
Test	**Fraction**	$n\|C$	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**
SP.U	25	250	43.0	15.6	12.8	10.1	69.9	16.8	14.1	10.7
		500	66.7	22.2	18.3	13.4	90.4	24.0	19.6	13.1
	50	250	89.0	36.5	29.6	18.5	99.3	39.6	29.3	19.9
		500	99.3	56.2	43.1	27.8	100.0	58.9	43.5	27.7
SP.L	25	250	67.2	23.1	19.1	13.0	89.7	25.9	20.3	14.5
		500	88.8	34.7	28.1	18.7	99.3	37.8	30.8	19.7
	50	250	99.2	56.7	46.5	27.4	100.0	58.7	47.4	27.2
		500	100.0	80.0	68.7	43.3	100.0	82.1	70.2	44.5
SP.E	25	250	43.1	15.6	12.9	10.3	70.1	16.9	14.1	10.7
		500	66.8	22.3	18.3	13.4	90.4	24.2	19.6	13.1
	50	250	89.1	36.6	29.6	18.5	99.3	39.8	29.5	20.0
		500	99.3	56.6	43.4	27.9	100.0	59.1	43.7	28.0

When 25% of models are misspecified models, the power is rather low, except in the case where the desks are independent (i.e., have a Gauss copula and correlation $\rho = 0$). We attribute this to the intuition that independent data increase the effective sample size, whereas correlated data decrease it. The power increases with both n and d. As the misspecification rate increases, the power increases, as we would expect. The spectral test with a linear weighting function (SP.L) gives the most power, while the other two kernels are comparable.

The results for the bispectral tests SP.UL and SP.UE are reported in Table 4. In this case, we add results for $n = 1000$, since we have observed that bispectral tests typically require more data for the correlation correction (CE) method to give tests that are well sized. We also add results for a misspecification fraction of 10%.

The results are clearly better than for the monospectral tests and show the advantages of bispectral tests—by using two kernels, we can effectively test for the correct specification of more moments of the distribution of PIT values in the kernel window (see discussion in Section 2.3). The general observations are the same as for the monospectral tests; the power increases with both n and d and with weakening dependence across desks. While it would clearly be best to base backtests on 1000 observations, to obtain tests that are both well sized and powerful, reasonable results are obtained for $n = 500$ even when only 25% of the desks are using misspecified risk models. When only 10% of the desks use misspecified models, the power is clearly weaker, but the test does still have some ability to detect that a number of desks are delivering poor P&L estimates.

Ther is little to choose from between the SP.UL and SP.UE tests. While the latter seems to be slightly more powerful, it also tends to have slightly worse size properties, as seen in Table 2; we would tend to favour the former for samples of size $n = 500$.

Table 4. Power of bispectral Z-tests SP.UL and SP.UE when a certain proportion of desks is misspecified.

		d	50				100			
		ρ	0		0.5		0		0.5	
Test	**Fraction**	$n\|C$	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**	**Gauss**	**t4**
SP.UL	10	250	23.1	16.3	21.2	17.5	38.5	19.2	23.3	18.4
		500	32.7	14.6	18.1	11.4	57.4	18.7	23.8	10.9
		1000	56.5	24.3	28.6	9.2	87.2	27.7	40.0	9.8
	25	250	76.9	48.6	59.0	28.3	97.4	69.5	83.4	31.7
		500	95.0	66.5	80.9	31.3	100.0	86.2	96.1	37.6
		1000	100.0	92.4	97.4	49.3	100.0	98.6	99.9	55.6
	50	250	99.9	97.2	98.9	78.9	100.0	99.9	100.0	83.8
		500	100.0	99.9	100.0	92.7	100.0	100.0	100.0	94.9
		1000	100.0	100.0	100.0	99.5	100.0	100.0	100.0	99.9
SP.UE	10	250	25.3	17.5	22.5	18.5	42.4	19.9	24.7	19.2
		500	34.5	14.5	18.7	11.6	60.0	18.7	24.6	10.9
		1000	57.5	24.4	28.7	8.5	87.7	27.8	40.1	9.3
	25	250	79.9	51.1	61.6	28.8	98.0	70.7	85.2	31.7
		500	95.5	67.2	81.8	30.3	100.0	86.7	96.2	36.8
		1000	100.0	92.5	97.5	48.7	100.0	98.5	99.9	54.8
	50	250	99.9	97.7	99.0	79.2	100.0	99.9	100.0	84.2
		500	100.0	99.9	100.0	92.8	100.0	100.0	100.0	94.6
		1000	100.0	100.0	100.0	99.5	100.0	100.0	100.0	99.9

5. Conclusions

In this paper, we proposed multivariate spectral Z-tests, which can be used to simultaneously backtest trading desk models when the dependence structure of P&L across trading desks is unknown. Multivariate backtests are potentially more powerful than univariate backtests at the level of the trading book, since they exploit a greater amount of data; typical banks can have 50–100 trading desks. Moreover, a simultaneous backtest avoids the problem of aggregating and drawing inferences from a set of single-desk backtests in the presence of unknown dependencies between test results.

The tests that we have developed are a multivariate extension of the spectral tests proposed in Gordy and McNeil (2020). They take the form of a Z-test against a normal or chi-squared reference distribution and make use of realised PIT values as input variables. PIT values provide more data about the quality of desk models than indicator variables for VaR violations and their benefits are already being exploited in the USA. It is likely that other regulatory authorities will use this type of information more systematically in the future as well.

The multivariate spectral tests are designed in such a way that the risk modelling group can select a kernel or weighting scheme to emphasise the region of the estimated P&L distribution where model performance is most critical, typically a region in the tail representing large losses. The tests can also provide an indirect validation of the expected shortfall measure, which now has a prominent role in the revised regulation.

We suggested a method of controlling for the unknown dependencies between desks, based on estimating correlations. The resulting tests were generally well sized, although bispectral tests typically required samples of $n = 500$ PIT vectors (2 years of daily data), while monospectral tests only required $n = 250$ (one year of daily data). However, in realistic situations where only a minority of the used desk models underestimated the risks in their P&L distributions, bispectral tests gave much better power and should generally be preferred to their monospectral counterparts.

The performance of the proposed multivariate tests suggests that they are a valuable addition to a bank's validation framework. Since they take the form of Z-tests, they are easy to implement and have quick run times. In the event of a significant test result, a bank would be able to implement a post-hoc testing scheme on individual desks to see where the main problems lie.

Finally, the new multivariate backtests could also have an interesting application to backtesting in the banking sector as a whole, based on data for the trading book provided to the regulator by individual banks. Significant test results would be an indication of the potential for spillover effects across banks caused by trading activities and would contribute to the understanding of systemic risk.

Clearly, it would be of interest to apply our tests to actual reported data from banks. However, multi-desk data are not currently available to researchers due to the sensitivities surrounding bank risk reporting both in the EU and the USA. An anonymised study of real data at the trading-book level for US banks is provided by Gordy and McNeil (2020) and shows the advantages of the (univariate) spectral backtesting framework over simple VaR backtests; we would certainly expect that these advantages will carry over to the multivariate setting. We hope that, by setting out a methodology for multi-desk backtesting in this paper, we can help to promote the use of desk-level PIT values for model validation and encourage banks and regulators to allow some datasets of realised PIT values to enter the public domain and stimulate further research into the refinement of the methodology.

Author Contributions: Conceptualisation, J.B. and A.J.M.; methodology, J.B. and A.J.M.; software, J.B.; validation, J.B. and A.J.M.; formal analysis, J.B. and A.J.M.; investigation, J.B. and A.J.M.; writing—original draft preparation, J.B. and A.J.M.; writing—review and editing, J.B. and A.J.M.; visualisation, J.B. and A.J.M.; supervision, A.J.M.; project administration, J.B. and A.J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The code to reproduce the simulation experiments is available from the authors.

Conflicts of Interest: Author Janine Balter was employed by the company Deutsche Bundesbank. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Deutsche Bundesbank had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Notes

[1] The regulatory framework, known as the Fundamental Review of the Trading Book (FRTB), requires that the backtest compares whether the number of observed VaR exceptions is consistent with a 99% confidence level with the comment that other α levels should be added to the backtesting exercise; see BCBS (2019) [M32.5] and [M32.13].

[2] At the suggestion of a referee, we also tried repeating all analyses using a Gumbel copula to simulate the dependence structure across desks, since this is a typical copula for extreme values and also leads to PIT values with upper tail dependence. The results were comparable to the Gauss and t4 copulas, in terms of size and power properties, and we have not included them in the tables.

References

Amisano, Gianni, and Raffaella Giacomini. 2007. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics* 25: 177–90.

Basel Committee on Banking Supervision. 2006. International Convergence of Capital Measurement and Capital Standards. Available online: https://www.bis.org/publ/bcbs128.pdf (accessed on 30 March 2020).

Basel Committee on Banking Supervision. 2011. Revisions to the Basel II Market Risk Framework. Available online: https://www.bis.org/publ/bcbs193.pdf (accessed on 13 January 2021).

Basel Committee on Banking Supervision. 2019. Minimum Capital Requirements for Market Risk. Available online: https://www.bis.org/bcbs/publ/d457.pdf (accessed on 18 January 2021).

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

Berkowitz, Jeremy, Peter Christoffersen, and Denis Pelletier. 2011. Evaluating value-at-risk models with desk-level data. *Management Science* 57: 2213–27. [CrossRef]

Campbell, Sean D. 2007. A review of backtesting and backtesting procedures. *Journal of Risk* 9: 1–17. [CrossRef]

Christoffersen, Peter F. 1998. Evaluating interval forecasts. *International Economic Review* 39: 841–62. [CrossRef]

Costanzino, Nick, and Mike Curran. 2015. Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation* 9: 21–31. [CrossRef]

Danciulescu, Cristina. 2016. Backtesting aggregate risk. *Journal of Forecasting* 35: 285–307. [CrossRef]
Diks, Cees, Valentyn Panchenko, and Dick van Dijk. 2011. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* 163: 215–30. [CrossRef]
Du, Zaichao, and Juan Carlos Escanciano. 2017. Backtesting expected shortfall: accounting for tail risk. *Management Science* 63: 940–85. [CrossRef]
Dunn, Olive Jean. 1959. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics* 30: 192–97. [CrossRef]
Engle, Robert F., and Simone Manganelli. 2004. Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22: 367–81.
Gneiting, Tilmann, and Roopesh Ranjan. 2011. Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* 29: 411–22.
Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society* 69: 243–68. [CrossRef]
Gordy, Michael B., and Alexander J. McNeil. 2020. Spectral backtests of forecast distributions with application to risk management. *Journal of Banking & Finance* 116: 105817.
Hofert, Marius, and Martin Mächler. 2016. Parallel and other simulations in R made easy: An end-to-end study. *Journal of Statistical Software* 69: 1–44. [CrossRef]
Jorion, Philippe. 2007. *Value at Risk. The New Benchmark for Managing Financial Risk*, 3rd ed. New York: McGraw-Hill.
Kratz, Marie, Yen H. Lok, and Alexander J. McNeil. 2018. Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance* 88: 393–407.
Kupiec, Paul. 1995. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives* 3: 73–84. [CrossRef]
McNeil, Alexander J., Rüdiger Frey, and Paul Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools*, 2nd ed. Princeton: Princeton University Press.
Pérignon, Christophe, and Daniel R. Smith. 2008. A new approach to comparing VaR estimation methods. *The Journal of Derivatives* 16: 54–66. [CrossRef]
Rosenblatt, Murray. 1952. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* 23: 470–72. [CrossRef]
Van der Vaart, Aad W. 2000. *Asymptotic Statistics*. Cambridge: Cambridge University Press.
Wied, Dominik, Gregor N. F. Weiß, and Daniel Ziggel. 2016. Evaluating value-at-risk forecasts: A new set of multivariate backtests. *Journal of Banking & Finance* 72: 121–32.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.