Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Spectral backtests of forecast distributions with application to risk management

Michael B. Gordy and Alexander J. McNeil

2018-021

Please cite this paper as:
Gordy, Michael B., and Alexander J. McNeil (2018). "Spectral backtests of forecast distributions with application to risk management," Finance and Economics Discussion Series 2018-021. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2018.021.

NOTE: Staff working papers in the Finance and Economics Discussion Series (FEDS) are preliminary materials circulated to stimulate discussion and critical comment. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors. References in publications to the Finance and Economics Discussion Series (other than acknowledgement) should be cleared with the author(s) to protect the tentative character of these papers.

Spectral backtests of forecast distributions with application to risk management*

Michael B. Gordy

Federal Reserve Board, Washington DC

Alexander J. McNeil

The York Management School, University of York

February 21, 2018

Abstract

We study a class of backtests for forecast distributions in which the test statistic is a spectral transformation that weights exceedance events by a function of the modeled probability level. The choice of the kernel function makes explicit the user's priorities for model performance. The class of spectral backtests includes tests of unconditional coverage and tests of conditional coverage. We show how the class embeds a wide variety of backtests in the existing literature, and propose novel variants as well. In an empirical application, we backtest forecast distributions for the overnight P&L of ten bank trading portfolios. For some portfolios, test results depend materially on the choice of kernel.

JEL Codes: C52; G21; G28; G32
Keywords: Backtesting; Volatility; Risk management

*We thank Harrison Katz for excellent research assistance. We have benefitted from discussion with Mike Giles, Marie Kratz, Hsiao Yen Lok, David Lynch, David McArthur, Michael Milgram, and Johanna Ziegel. The opinions expressed here are our own, and do not reflect the views of the Board of Governors or its staff. Address correspondence to Alexander J. McNeil, The York Management School, University of York, Freboys Lane, York YO10 5GD, UK, +44 (0) 1904 325307, alexander.mcneil@york.ac.uk.

1 Introduction

In many forecasting exercises, fitting some range of quantiles of the forecast distribution may be prioritized in model design and calibration. In risk management applications, which will motivate this study, accuracy near the median of the distribution or in the "good tail" of high profits is generally much less important than accuracy in the "bad tail" of large losses. Even within the region of primary interest, preferences may be nonmonotonic in probabilities. For example, the modeller may care a great deal about assessing the magnitude of once-in-a-decade market disruptions, but care much less about quantiles in the extreme tail that are consequent to unsurvivable cataclysmic events. In this paper, we study a class of backtests for forecast distributions in which the test statistic weights exceedance events by a function of the modeled probability level. The choice of the kernel function makes explicit the priorities for model performance. The backtest statistic and its asymptotic distribution are analytically tractable for a very large family of kernel functions.

Our approach unifies a wide variety of existing approaches to backtesting. In the area of risk management, the time-honored test statistic (dating back to Kupiec, 1995) is simply a count of "VaR exceedances," i.e., indicator variables equal to one whenever the realized trading loss is in excess of the day-ahead value-at-risk (VaR) forecast. In our framework, this corresponds to a Dirac delta kernel function in which all weight is concentrated at exactly the target VaR level (e.g., at $\alpha = 0.99$). At the other extreme, the tests applied in Diebold et al. (1998) represent a special case in which weights are uniform across all probability levels. The likelihood-ratio test of Berkowitz (2001) represents an intermediate case of a kernel truncated to tail probabilities. The class of spectral backtests encompasses discrete kernels, which selectively weight forecasts at a discrete set of probability levels, as well as continuous kernels, which apply positive weight throughout an interval of levels. Perhaps of greater importance in practice, the class allows for both tests of unconditional coverage and tests of conditional coverage.

The application of a weighting function is this paper bears some similarity to the approach of Amisano and Giacomini (2007) and Gneiting and Ranjan (2011) in the literature on comparisons of density forecasts. In both of those papers, weights are applied to a forecast scoring rule to obtain measures of forecast performance that accentuate the tails (or other regions) of the distribution. However, the measure for any one forecasting method has no absolute meaning and is designed to facilitate comparison with other methods using the general comparative testing approach proposed by Diebold and Mariano (1995). In contrast, our tests are absolute tests of forecast quality in the spirit of Diebold et al. (1998). While the comparative testing approach is clearly useful for the *internal* refinement of the forecasting method by the forecaster, the absolute testing approach in this paper facilitates the *external* evaluation of the forecaster's results by another agent, such as a regulator.

Our investigation is motivated in part by a major expansion in the data available to regulators for the backtesting exercise. Prior to 2013, banks in the US reported to regulators VaR exceedances at the 99% level. The new Market Risk Rule mandates that banks report for each trading day the probability associated with the realized P&L in the prior day's forecast distribution, which is equivalent to providing the regulator with VaR exceedances *at every level* $\alpha \in [0, 1]$. The expanded reporting regime allows us to assess the tradeoff between power and specificity in backtesting. If a regulator is concerned narrowly with the validation of reported VaR at level α, than a count of VaR exceedances is a sufficient statistic for a test for unconditional coverage. However, if the regulator is willing to assign positive weight to probability levels in a *neighborhood* of α, we can construct more powerful backtests. Furthermore, our approach is consistent with a broader view of the risk manager's mandate to forecast probabilities over a range of large losses. The formal guidance of US regulators to banks on internal model validation explicitly requires "checking the distribution of losses against other estimated percentiles" (Board of Governors of the Federal Reserve System, 2011, p. 15).

The reforms mandated by the Fundamental Review of the Trading Book (Basel Committee on Bank Supervision, 2013) introduce a distinct set of challenges. Due to begin parallel run in 2018, the FRTB replaces 99%-VaR with 97.5%-Expected Shortfall (ES) as the determinant of capital requirements. While there has been a lot of debate around the question of whether or not ES is amenable to direct backtesting (Gneiting, 2011; Acerbi and Szekely, 2014; Fissler and Ziegel, 2015; Fissler et al., 2016), our contribution addresses a different issue. We devise tests of the *forecast distribution* from which risk measures are estimated and not tests of the *risk measure* estimates. When VaR if of primary interest it may be noted that some limiting special cases of our testing methodology are equivalent to VaR exceedance tests. When ES is of primary interest it may be argued that a satisfactory forecast of the tail of the loss distribution is of even greater importance, since the risk measure depends on the whole tail.

Two other aspects of FRTB are relevant to our contribution. First, although estimates of ES will be the cornerstone of the risk capital calculation, the model approval process will continue to be based on VaR estimates and VaR exceedances. Second, FRTB requires banks to go beyond the mandatory VaR backtesting regime to consider multiple levels or other features of the tail. Without being prescriptive, the Basel Committee explicitly mentions a number of possible directions for the extended model validation requirements including the use of probability integral transform values (Basel Committee on Bank Supervision, 2016, Appendix B), which also serve as the input in our class of backtests. For convenience in exposition, we mostly assume henceforth that the backtest is conducted by a regulator who is interested primarily in assessing the bank's 99%-VaR forecast, but our conclusions hinge little on the choice of risk measure, and furthermore apply as much to internal assessments of forecasting performance as

to external assessment by regulators.

In Section 2, we lay out the statistical setting for the risk manager's forecasting problem and the data to be collected for backtesting. The transformation that underpins the class of spectral backtests is introduced in Section 3. Spectral backtests of unconditional coverage are described in Section 4. In Section 5, we develop tests of conditional coverage based on the martingale difference property. As an application to real data, in Section 6 we backtest ten bank models for overnight P&L distributions for trading portfolios.

2 Theory and practice of risk measurement

We assume that a bank models profit and loss (P&L) on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{N}_0}, \mathbb{P})$ where $\mathcal{F}_t$ represents the information available to the risk manager at time t, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and $\mathbb{N}$ denotes the non-zero natural numbers. For any time $t \in \mathbb{N}$, L_t is an $\mathcal{F}_t$-measurable random variable representing portfolio loss (i.e., negative P&L) in currency units. We denote the conditional loss distribution given information to time $t-1$ by

$$F_t(x) = \mathbb{P}\left(L_t \leqslant x \mid \mathcal{F}_{t-1}\right).$$

The loss distribution cannot be assumed to be time-invariant. The distribution of returns on the underlying risk factors (e.g., equity prices, exchange rates) is time-varying, most notably due to stochastic volatility. Furthermore, F_t depends on the composition of the portfolio. Because the portfolio is rebalanced in each period, F_t can evolve over time even when factor returns are iid.

For $t \in \mathbb{N}$ we can define the process (U_t) by $U_t = F_t(L_t)$ using the probability integral transform (PIT). Under the assumption that the conditional loss distributions at each time point are continuous, the result of Rosenblatt (1952) implies that the process $(U_t)_{t\in\mathbb{N}}$ is a sequence of iid standard uniform variables. The risk manager builds a model $\widehat{F}_t$ of F_t based on information up to time $t-1$. *Reported PIT-values* are the corresponding rvs (P_t) obtained by setting $P_t = \widehat{F}_t(L_t)$ for $t \in \mathbb{N}$. If the models $\widehat{F}_t$ form a sequence of *ideal* probabilistic forecasts in the sense of Gneiting et al. (2007), i.e. coinciding with the conditional laws F_t of L_t for every t, then we expect the reported PIT-values to behave like an iid sample of standard uniform variates.[1]

Reported PIT-values contain information about VaR exceedances at any level α. To see this note that

$$P_t \geqslant \alpha \iff L_t \geqslant \widehat{\mathrm{VaR}}_{\alpha,t} \tag{1}$$

where $\widehat{\mathrm{VaR}}_{\alpha,t} := \widehat{F}_t^{\leftarrow}(\alpha)$ is an estimate of the α-VaR constructed at time $t-1$ by cal-

[1] In the statistical forecasting literature tests based on the uniformity and independence of PIT value are also referred to as tests that a sequence of models is *calibrated in probability* (Gneiting et al., 2007; Gneiting and Ranjan, 2011).

culating the generalized inverse of $\widehat{F}_t$ at α. Relationship (1) always holds for any model $\widehat{F}_t$, whether continuous or discrete.[2] Thus, we would expect well-designed tests that use reported PIT-values to be more powerful than VaR exceedance tests in detecting deficiencies in the models $\widehat{F}_t$.

Our tests are agnostic with respect to the procedures and models used by the bank in forecasting. In practice, there is considerable heterogeneity in methodology. For nearly two decades, most large banks have relied primarily on some variant of historical sampling (HS), which is a nonparametric method based on re-sampling of historical risk-factor changes or returns. A sufficient condition for the "plain-vanilla" HS estimator $\widehat{F}_t^{HS}$ to be a consistent estimator of F_t for all t is that the returns are iid; however the approach does not account for serial dependence in returns such as time-varying volatility. For this reason, some banks adopt *filtered* historical simulation (FHS) as suggested by Hull and White (1998) and Barone-Adesi et al. (1998). In this approach, the historical risk-factor returns are normalized by their estimated volatilities, which are typically obtained by taking an exponentially-weighted moving-average of past returns. Banks that do not use HS or FHS typically adopt a parametric model for the joint distribution of risk-factor changes.[3]

In our empirical application, testing for delayed response to changes in volatility is of special interest. Assuming a roughly symmetric loss distribution centered at zero, the frequent switching between positive and negative values will tend to cause PIT values to be serially uncorrelated, even when volatility is misspecified in the model. However, extreme PIT-values (i.e., near 0 or 1) will tend to beget extreme PIT-values in high volatility periods, and middling PIT-values (i.e., near ½) will tend to beget middling PIT-values in low volatility periods. This pattern can be inferred by examining autocorrelation in the transformed values $|2P_t - 1|$. We will exploit this transformation in implementing tests of conditional coverage in Section 6.

There are relatively few empirical studies of bank VaR forecasting. Berkowitz and O'Brien (2002) show that VaR estimates by US banks are conservative (i.e., there are fewer exceedances than expected) and that the forecasts underperform simple time-series models applied to daily P&L. In a sample of Canadian banks in 1999–2005, Pérignon et al. (2008) record only two 99%-VaR exceedances in 7354 observations. Pérignon and Smith (2010) report similar results for a larger international sample in 1996–2005. For the subsample of banks employing HS, they also show that reported VaR has little predictive power for subsequent volatility in P&L. Berkowitz et al. (2011) apply a suite of backtests to a proprietary sample of four business lines of a single bank in 2001–2004. While they find some evidence of excessive conservatism and/or clustering

[2] We can replace the weak inequalities with strict inequalities if the models $\widehat{F}_t$ are strictly increasing and continuous. Since it is somewhat more common to consider the event $\{L_t > \widehat{\text{VaR}}_{\alpha,t}\}$ to be a VaR exceedance, we will define a VaR exceedance in terms of the reported PIT-value as the event $\{P_t > u\}$.

[3] The classic RiskMetrics approach can be considered a progenitor of this class of models.

of VaR exceedances in three of the four business lines, the exercise also demonstrates the limited power of backtests in sample sizes of two to three years. The importance of sample size is evident in the contrasting results of O'Brien and Szerszen (2017). In a sample of five large US banks from 2001–2014, tests of unconditional coverage reject VaR forecasts as excessively conservative for all banks in the pre-crisis and post-crisis periods, for which the samples spanned at least 1000 trading days per bank. In the crisis period, tests of unconditional coverage reject VaR forecasts as insufficiently conservative for all five banks, and independence is rejected for four of the banks. This pattern is consistent with a failure to model stochastic volatility.

3 Spectral transformations of PIT exceedances

The tests in this paper are based on transformations of indicator variables for PIT exceedances.[4] The transformations take the form

$$W_t = \int_0^1 \mathbb{1}_{\{P_t > u\}} \mathrm{d}\nu(u) \tag{2}$$

where ν is a finite measure defined on $[0, 1]$ which is designed to apply weight to different levels in the interval $(0, 1]$, typically in the region of the standard VaR level $\alpha = 0.99$. We refer to ν as the *kernel measure* for the transform. From (2), we can easily derive the closed-form expression

$$W_t = \nu([0, P_t)) \tag{3}$$

which shows that W_t is increasing in P_t.

3.1 Weighting schemes

For the weighting scheme in (2) we consider three possibilities:

Discrete weighting in which the kernel measure takes the form $\nu = \sum_{i=1}^m \gamma_i \delta_{\alpha_i}$ for $m \geqslant 1$. This places positive mass $\gamma_1, \ldots, \gamma_m$ at the ordered values $\alpha_1 < \cdots < \alpha_m$ leading to

$$W_t = \sum_{i=1}^m \gamma_i \mathbb{1}_{\{P_t > \alpha_i\}}. \tag{4}$$

Continuous weighting in which the measure has density $\mathrm{d}\nu(u) = g(u)\mathrm{d}u$ on the interval $[\alpha_1, \alpha_2] \subset [0, 1]$, where the function g satisfies

Assumption 1. (i) $g(u) = 0, u \notin [\alpha_1, \alpha_2]$, (ii) g is continuous and (iii) $g(u) > 0, u \in (\alpha_1, \alpha_2)$.

[4] We draw on the integral transform literature in describing our backtest as "spectral." Our approach is unconnected to the *spectral density* test of Durlauf (1991). The latter is a test of the martingale property that examines whether the spectrum (in the sense of the transformed autocovariance sequence) is flat.

In this case we have

$$W_t = \int_{\alpha_1}^{\alpha_2} g(u) \mathbb{1}_{\{P_t > u\}} \mathrm{d}u. \tag{5}$$

We refer to g as the *kernel density*. It plays the same role as the "kernel function" in the nonparametric statistics literature, but we use the term in the more general sense of the integral transform literature. When g satisfies the additional requirement that $\int_{\alpha_1}^{\alpha_2} g(u)\mathrm{d}u = 1$, it is a *normalized* kernel density. In nonparametric statistics, the kernel is often defined to be normalized and symmetric, but we do not impose either requirement here.

As in the nonparametric statistics literature, the interval $[\alpha_1, \alpha_2]$ is referred to as the *kernel window*. Note that g is strictly positive inside the kernel window, but may equal zero at the boundary points. This allows us to accommodate functions such as the Epanechnikov kernel that vanish at the boundaries. Writing G for the integral of g, (3) can be expressed as

$$W_t = G(\alpha_1 \vee (P_t \wedge \alpha_2)) \tag{6}$$

Since G is strictly increasing inside the kernel window, (6) implies that W_t is a strictly increasing function of the truncated PIT-value $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$.

Continuous weighting can be viewed as a way of building tests that incorporate information from reported PIT-values in a *neighborhood* of a particular VaR level α. Let g^* be a normalized kernel density on $[0, 1]$, and define a family of normalized kernel densities $g_{\alpha,\epsilon}$ on the intervals $[\alpha - \epsilon/2, \alpha + \epsilon/2]$ by

$$g_{\alpha,\epsilon}(u) = \frac{1}{\epsilon} g^* \left(\frac{u - \alpha + \epsilon/2}{\epsilon} \right). \tag{7}$$

Then we have that the measures $\nu_{\alpha,\epsilon}$ defined by $g_{\alpha,\epsilon}$ converge to Dirac measure δ_α as $\epsilon \to 0+$, and $\lim_{\epsilon \to 0} W_t = \mathbb{1}_{\{P_t > \alpha\}}$ almost surely. Thus, classic tests based on the exceedance indicator $\mathbb{1}_{\{P_t > \alpha\}}$ can be seen as limiting cases of more general continuous tests as the width ϵ of the kernel window vanishes to zero.

Combined discrete and continuous weighting. It is of course possible to consider a measure that is given by the sum of a discrete weighting and a continuous weighting scheme. We consider one test of this kind in Section 4.3. In this general case, the notion of the kernel window generalizes as the support of the kernel measure.

3.2 Univariate and multivariate transformations

We consider tests based on univariate and multivariate spectral transformations of the data. A univariate transformation applies a single kernel measure ν and yields spectrally

transformed PIT-values $W_1, \ldots, W_n$ according to (2). A multivariate transformation corresponds to a set of distinct kernel measures $\nu_1, \ldots, \nu_j$. The transformed PIT values are then vector-valued variables $\boldsymbol{W}_1 \ldots, \boldsymbol{W}_n$ where

$$\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,j})', \quad W_{t,i} = \int_0^1 \mathbb{1}_{\{P_t > u\}} \mathrm{d}\nu_i(u), \; j = 1, \ldots, j. \tag{8}$$

Spectrally transformed PIT values satisfy simple product rules that we will later exploit in calculating variances of the (W_t) and covariance matrices of the $(\boldsymbol{W}_t)$. Consider two discrete kernel measures ν_1 and ν_2 which share the same support. Then the product $W_{t,1}W_{t,2}$ is a spectral transformation of P_t on the same support, and the kernel weights are easily calculated as summarized in the following result.

Proposition 3.1. *Fix a set of distinct levels $0 < \alpha_1 < \cdots < \alpha_m < 1$, and let $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \ldots, \gamma_{i,m})'$ be a set of positive weights. The set of spectrally transformed PIT values defined by $W_{t,i} = \sum_{\ell=1}^m \gamma_{i,\ell} \mathbb{1}_{\{P_t > \alpha_\ell\}}$ is closed under multiplication and $W_{t,1}W_{t,2} = \sum_{\ell=1}^m \gamma^*_\ell \mathbb{1}_{\{P_t > \alpha_\ell\}}$ where γ^*_ℓ are positive weights satisfying*

$$\gamma^*_\ell = \gamma_{1,\ell} \sum_{\ell'=1}^{\ell} \gamma_{2,\ell'} + \gamma_{2,\ell} \sum_{\ell'=1}^{\ell} \gamma_{1,\ell'} - \gamma_{1,\ell}\gamma_{2,\ell}.$$

If $\sum_{\ell=1}^m \gamma_{1,\ell} = \sum_{\ell=1}^m \gamma_{2,\ell} = 1$, then $\sum_{\ell=1}^m \gamma^*_\ell = 1$.

An analogous product rule holds for the set of spectral transformations with continuous kernels on the same kernel window.

Proposition 3.2. *Fix a kernel window $[\alpha_1, \alpha_2] \subset [0, 1]$, and let g_i be a kernel density on $[\alpha_1, \alpha_2]$ satisfying Assumption 1. The set of spectrally transformed PIT values defined by $W_{t,i} = \int_{\alpha_1}^{\alpha_2} g_i(u) \mathbb{1}_{\{P_t > u\}} \mathrm{d}u$ is closed under multiplication and $W_{t,1}W_{t,2} = \int_{\alpha_1}^{\alpha_2} g^*(u) \mathbb{1}_{\{P_t > u\}} \mathrm{d}u$ where*

$$g^*(u) = g_1(u)G_2(u) + g_2(u)G_1(u).$$

If g_1 and g_2 are normalized kernel densities on $[\alpha_1, \alpha_2]$, then so is g^*.

Proofs for these proposition and other mathematical results are found in Appendix A.

3.3 Spectral backtests

We will refer to any backtest based on spectrally transformed PIT exceedances as a spectral backtest. This encompasses a great variety of tests but two general testing approaches will feature prominently in our presentation: Z-tests and likelihood ratio tests (LRTs).

To formulate these tests we state the null hypothesis in this paper to be

$$H_0: \quad \boldsymbol{W}_t \sim F_W^0 \text{ and } \boldsymbol{W}_t \perp\!\!\!\perp \mathcal{F}_{t-1},\ \forall t, \tag{9}$$

where F_W^0 denotes the distribution function of $\boldsymbol{W}_t$ in (8) when P_t is uniform; obviously this subsumes the univariate case where we will simply write W_t for the spectrally-transformed variables. The null hypothesis (9) implies that $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ are iid random variables but also requires that $\boldsymbol{W}_t$ is independent of all information in the time $t-1$ information set $\mathcal{F}_{t-1}$, such as the values P_{t-j} for $j > 0$. Observe that our null hypothesis is strictly weaker than a null hypothesis that the (P_t) are iid Uniform. This is by intent. Since the regulator is free to choose ν in accordance with her priorities, she should not object to departures from uniformity and serial independence that arise outside her chosen kernel window.

Z-tests. In the univariate case these are based on the asymptotic normality of $\overline{W}_n = n^{-1}\sum_{t=1}^n W_t$ under the null hypothesis (9). Using Propositions 3.1 and 3.2, we calculate $\mu_W = \mathbb{E}(W_t)$ and $\sigma_W^2 = \text{var}(W_t)$ in the null model F_W^0. It then follows trivially from the central limit theorem (CLT) that, under the null hypothesis (9),

$$Z_n = \frac{\sqrt{n}(\overline{W}_n - \mu_W)}{\sigma_W} \xrightarrow[n\to\infty]{d} N(0,1). \tag{10}$$

In the multivariate case ($\dim \boldsymbol{W}_t = j$) we have

$$\sqrt{n}\left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right) \xrightarrow[n\to\infty]{d} N_j(\mathbf{0}, \Sigma_W)$$

where $\overline{\boldsymbol{W}}_n = n^{-1}\sum_{t=1}^n \boldsymbol{W}_t$ and $\boldsymbol{\mu}_W$ and Σ_W are the mean vector and covariance matrix of the null distribution F_W^0. Hence a test can be based on assuming for large enough n that

$$T_n = n\left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right)' \Sigma_W^{-1} \left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right) \sim \chi_j^2, \tag{11}$$

where we refer to T_n as a j-spectral Z-test statistic.

Likelihood ratio tests. These are based on parametric models $F_W(\cdot \mid \theta)$ that nest the model in the null hypothesis (9). In other words $F_W^0 = F_W(\cdot, \theta_0)$ for some value θ_0. Writing $\mathcal{L}_W(\theta \mid \boldsymbol{W})$ for the likelihood function, the test is based on the asymptotic distribution of the statistic

$$\text{LR}_{W,n} = \frac{\mathcal{L}_W(\theta_0 \mid \boldsymbol{W})}{\mathcal{L}_W(\hat{\theta} \mid \boldsymbol{W})} \tag{12}$$

where $\hat{\theta}$ denotes the maximum likelihood estimate.

An important difference between the two classes of test is that the Z-tests are sensitive to the choice of weighting scheme whereas the likelihood ratio tests are not. Consider the univariate case for simplicity. The only aspect of the kernel measure ν that determines the likelihood test statistic $\mathrm{LR}_{W,n}$ is its support; the actual weighting scheme applied on the support plays no role. For example, in the case of continuous weighting, it is the kernel window $[\alpha_1, \alpha_2]$ that determines the test statistic and not the kernel density g. Apart from the choice of the support of the measure the only discretion we have over the likelihood ratio test is the choice of nesting family $F_W(\cdot \mid \theta)$.

This is a consequence of the well-known invariance of the likelihood ratio test under strictly increasing tranformations. To make this assertion clearer we will now give a version of the invariance result in the case of univariate continuous weighting, which will facilitate some of our later arguments.

Theorem 3.3. *Let $F_P(p \mid \theta)$ be a parametric model for the reported PIT values $P_1 \ldots, P_n$ that nests the uniform model as a special case corresponding to $\theta = \theta_0$. Let $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$ denote the corresponding truncated PIT values and $W_t = T(P_t^*)$ the values that are obtained under any transformation T which is strictly increasing and continuous on $[\alpha_1, \alpha_2]$ such as* (6)*.*

Let $\mathcal{L}_P(\theta \mid \boldsymbol{P}^*)$ denote the likelihood for the truncated PIT values under $F_P(p \mid \theta)$ and let $\mathcal{L}_W(\theta \mid \boldsymbol{W})$ denote the likelihood for the (W_t) values under the distribution $F_W(w \mid \theta)$ implied by $F_P(p \mid \theta)$. Then the maximizing values of $\mathcal{L}_P(\theta \mid \boldsymbol{P}^*)$ and $\mathcal{L}_W(\theta \mid \boldsymbol{W})$ are the same and the corresponding likelihood ratio test statistics of the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_0 : \theta \neq \theta_0$ coincide regardless of the choice of the transformation T.

4 Tests of unconditional coverage

It is common to divide backtesting methods into tests of unconditional calibration and tests of conditional calibration. In the context of VaR backtesting, an unconditional test is a test that exceedances are Bernoulli events with the correct probability of occurrence while a conditional test is a test that exceedances have the correct conditional probability of occurrence, which is equivalent to requiring that they are also independent events. For spectrally transformed PIT-values, an unconditional test would test for the distribution F_W^0 implied by the uniformity of the PIT-values while a conditional test would explicitly test for both the correct distribution and the independence of W_t and $\mathcal{F}_{t-1}$ for all t.

In this section we present a number of unconditional tests based on the Z-test and LR-test ideas discussed in Section 3. It is important to note that the convergence results on which these tests are based, although mostly stated under iid assumptions, do hold in situations where the independence assumption is relaxed. Consider the Z-test

convergence result in (10) and recall the martingale CLT of Billingsley (1961): if (X_t) is a stationary and ergodic process adapted to a filtration $(\mathcal{F}_t)$ satisfying the martingale-difference property $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = 0$, then $\sqrt{n}\overline{X} \xrightarrow[n\to\infty]{d} N(0, \sigma_X^2)$ where σ_X^2 denotes the variance of X_t. Thus, the same convergence in (10) would be obtained if $(W_t - \mu_W)$ is a stationary and ergodic martingale difference sequence, which would entail that (W_t) is an uncorrelated sequence. More generally, provided that $\lim_{n\to\infty} \mathrm{var}(\sqrt{n}\overline{W}_n) \approx \sigma_W^2$ the test statistic Z in (10) will have no power to detect serial dependence. If, however, there is persistent positive serial correlation in (W_t) leading to $\lim_{n\to\infty} \mathrm{var}(\sqrt{n}\overline{W}_n) > \sigma_W^2$ then the test statistic Z will have some power to detect dependencies; however, more targeted tests of the independence property are available and are the subject of Section 5.

An early paper on backtesting in a risk-management setting is Kupiec (1995), who proposed a binomial likelihood ratio test for the number of VaR exceedances. Ziggel et al. (2014) offer a refinement of this count-based test. Campbell (2006) recommended testing exceedances at multiple levels, and introduced the Pearson chi-squared test in this context. Pérignon and Smith (2008) proposed a multilevel likelihood ratio test generalizing the binomial test of Kupiec (1995). A multinomial LRT also underlies the work of Colletaz et al. (2013) on the concept of a "risk map" to describe VaR exceedances at two different levels. Kratz et al. (2016) provide a comparison of unconditional multilevel tests (including Pearson and LRT) in a typical set-up for backtesting trading book models and advocate the use of Nass's variant on the Pearson test for control of size and power.

Crnkovic and Drachman (1996) appear to have been first to advocate the use of PIT-values for backtesting risk management models. They also allow for a weighting function that plays the role of our kernel density, but the distribution for the resulting test statistic must be simulated.[5] The seminal paper of Diebold et al. (1998) described a number of tests for the uniformity and independence of PIT values. Berkowitz (2001) advocated a likelihood-ratio test based on fitting a truncated normal distribution to probit-transformed PIT-values for regulatory application.

Most closely related to our work, Du and Escanciano (2017) and Costanzino and Curran (2015) have proposed test statistics for spectral risk measures which can be viewed as special cases of our univariate spectral Z-test approach. Both papers consider a mathematical framework that permits a variety of kernels but focus on the case of a uniform kernel and interpret the tests in terms of backtesting expected shortfall. In contrast, we provide a general methodology that allows a bespoke choice of one or more kernels according to testing priorities, show how this embeds many existing tests and new tests and show how the framework may be easily generalized to the conditional

[5]The test of Crnkovic and Drachman (1996) is based on a weighted Kuiper distance between the distribution of PIT values and the uniform. They refer to their weighting scheme as a "worry" function, and propose that it should place higher weight on extreme PIT values.

case.[6] Other contributions using PIT-values include Kerkhof and Melenberg (2004), who derive VaR and expected shortfall backtesting statistics by applying a functional delta method to the empirical distribution function of PIT-values and Zumbach (2006), who refers to PIT-values as probtiles.

In Section 4.1 we describe unconditional coverage tests based on discrete kernels. Continuous kernels are considered in Section 4.2. Mixed kernels emerge in Section 4.3 through the study of tests based on a truncated probitnormal distribution.

4.1 Discrete weighting

Discrete tests are based on the univariate transformation $W_t = \sum_{i=1}^{m} \gamma_i \mathbb{1}_{\{P_t > \alpha_i\}}$ as defined in (4) and the multivariate transformation $\boldsymbol{W}_t = (\mathbb{1}_{\{P_t > \alpha_1\}}, \ldots, \mathbb{1}_{\{P_t > \alpha_m\}})'$ in (8) for the same set of ordered levels $\alpha_1 < \cdots < \alpha_m$. Obviously, when $m = 1$ (and $\gamma_1 = 1$) both transformations yield $W_t = \mathbb{1}_{\{P_t > \alpha\}}$, so that we obtain iid Bernoulli$(1-\alpha)$ variables under the null hypothesis (9). This is the basis for standard VaR exceedance testing based on the binomial distribution. The case $m > 1$ yields multinomial tests. We consider first the binomial case followed by the multinomial case, in each case treating the LRT followed by the Z-test.

A two-sided binomial LRT of the null $p = 1 - \alpha$ against the alternative $p \neq 1 - \alpha$ can be based on the asymptotic chi-squared distribution of the LR statistic under the null in (12); this is the approach taken in Kupiec (1995) and Christoffersen (1998). Note that the traffic-light system and model approval rules under Basel (see, e.g., Basel Committee on Bank Supervision, 2016, Appendix B) are actually based on a one-sided LRT of the null hypothesis against the simple alternative $p = p_1$ for $p_1 > 1 - \alpha$; this amounts to comparing the exception count $\sum_{t=1}^{n} W_t$ to a critical value defined by the binomial distribution.

The Z-test statistic (10) for $W_t = \mathbb{1}_{\{P_t > \alpha\}}$ coincides with the binomial score test statistic

$$Z_n = \frac{\sqrt{n}\left(\overline{W}_n - (1-\alpha)\right)}{\sqrt{\alpha(1-\alpha)}}. \tag{13}$$

Kratz et al. (2016) give a comparison of different binomial tests and find that the binomial score test perfoms best for the probability levels and sample sizes that are of typical regulatory interest.

When $m > 1$ the variables $W_t = \sum_{i=1}^{m} \gamma_i \mathbb{1}_{\{P_t > \alpha_i\}}$ take the ordered values $\Gamma_0 < \Gamma_1 < \cdots < \Gamma_m$ where $\Gamma_0 = 0$ and $\Gamma_k = \sum_{i=1}^{k} \gamma_i$ for $k = 1, \ldots, m$. Under the null

[6] Du and Escanciano (2017) also show how the asymptotic distribution of the test can be adapted to account for estimation error. We view this as less relevant in our setting since a regulator will tend to take the strict line that backtests should penalize a failure to estimate models accurately even when the models used are essentially correct in form.

hypothesis (9) the distributions of W_t and $\boldsymbol{W}_t$ satisfy

$$\mathbb{P}(W_t = \Gamma_i) = \mathbb{P}(\mathbf{1}'\boldsymbol{W}_t = i) = \alpha_{i+1} - \alpha_i, \quad i \in \{0, 1, \ldots, m\}, \tag{14}$$

where $\alpha_0 = 0$ and $\alpha_{m+1} = 1$. In both cases this describes a multinomial distribution.

The multinomial generalization of the binomial LRT of Kupiec (1995) as proposed by Pérignon and Smith (2008) is nested in our framework. The test depends on the spectrally transformed PIT values through the observed cell counts $O_i = \sum_{t=1}^{n} \mathbb{1}_{\{W_t=\Gamma_i\}}$ (univariate transformation) or $O_i = \sum_{t=1}^{n} \mathbb{1}_{\{\mathbf{1}'\boldsymbol{W}_t=i\}}$ (multivariate transformation). Note in the former case that the cumulative weights Γ_i play no role in the resulting test statistic, a consequence of the invariance property of the LRT noted in Section 3.3.

The univariate and multivariate tranformations do however result in different Z-tests which can be considered as alternative generalizations of the binomial score test. In the univariate case we can apply Proposition 3.1 to obtain

$$W_t^2 = \sum_{i=1}^{m} \gamma_i^* \mathbb{1}_{\{P_t > \alpha_i\}} \quad \text{where} \quad \gamma_i^* = 2\gamma_i \sum_{j=1}^{i} \gamma_j - \gamma_i^2 = 2\gamma_i \Gamma_i \gamma_i^2,$$

from which it is straightforward to calculate that the first two moments of W_t are given by

$$\mu_W = \sum_{i=1}^{m} \gamma_i (1 - \alpha_i), \quad \sigma_W^2 = \sum_{i=1}^{m} \gamma_i^* (1 - \alpha_i) - \mu_W^2.$$

Hence we can construct a Z-test based on the statistic Z_n in (10) and vary the weights γ_i to emphasise different levels α_i.

In the multivariate case, if we construct an m-spectral Z-test as in (11), then we obtain the classical Pearson chi-squared statistic as proposed by Campbell (2006).

Theorem 4.1.

$$n(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W)' \Sigma_W^{-1} (\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W) = \sum_{i=0}^{m} \frac{(O_i - n\theta_i)^2}{n\theta_i}$$

where $O_i = \sum_{t=1}^{n} \mathbb{1}_{\{\mathbf{1}'\boldsymbol{W}_t=i\}}$ *and* $\theta_i = \alpha_{i+1} - \alpha_i$ *for* $i = 0, \ldots, m$.

The Pearson test statistic $S_m = \sum_{i=0}^{m} (O_i - n\theta_i)^2 / (n\theta_i)$ is usually compared with a chi-squared distribution with m degrees of freedom; Theorem 4.1 in fact provides a proof of the asymptotic law of the Pearson test by showing that it can be written as an m-spectral Z-test.[7]

[7] Pearson's test is known to perform poorly when cell counts are small, which is typically the case in our tail-focussed applications. Nass's variant on the test (Nass, 1959), which is based on an improved approximation to the distribution of S_m gives improved results; see Cai and Krishnamoorthy (2006) and Kratz et al. (2016) for more details of the approximation.

4.2 Continuous weighting

In this section, W_t takes the form of (5) for a kernel density g satisfying Assumption 1; we also consider a bispectral test where $\boldsymbol{W}_t = (W_{t,1}, W_{t,2})'$ is constructed from two different kernel densities on the same kernel window.

In the univariate case, we apply the Z-test approach described in (10). It follows from the application of Proposition 3.2 in the case where $W_{t,1} = W_{t,2} = W_t$ that, under the null hypothesis (9),

$$\mathbb{E}(W_t) = \int_{\alpha_1}^{\alpha_2} g(u)(1-u)\mathrm{d}u \quad \text{and} \quad \mathbb{E}(W_t^2) = \int_{\alpha_1}^{\alpha_2} 2g(u)G(u)(1-u)\mathrm{d}u.$$

These moments are straightforward to calculate analytically for a wide variety of kernel densities, e.g., based on linear, quadratic, or exponential functions, or on beta-type densities of the form $(u-\alpha_1)^{a-1}(\alpha_2-u)^{b-1}$ for $a, b > 0$. Thus, our compact presentation of the continuous spectral Z-test subsumes a very large family of possible tests.

The bispectral generalization is a new test that extends the idea of the continuous spectral Z-test. For a bivariate spectral transformation $\boldsymbol{W}_t = (W_{t,1}, W_{t,2})'$ based on two distinct kernel densities g_1 and g_2 with the same kernel window it is straightforward to calculate $\boldsymbol{\mu}_W = \mathbb{E}(\boldsymbol{W}_t)$ and $\Sigma_W = \mathrm{cov}(\boldsymbol{W}_t)$. The off-diagonal element of the matrix Σ_W requires the calculation of $E(W_{t,1}W_{t,2})$ which can be achieved using Proposition 3.2. The test is based on assuming for large enough n the statistic T_n of (11) is distributed χ^2_2 under H_0.

The intuition for the bispectral test is that by considering two different spectral transformations we can test for two different features of the distribution of reported PIT values in the tail. Obviously, we could consider higher dimensional generalizations but the empirical results of Section 6 and the simulation results in our companion paper show that the bivariate test works well.

4.3 Tests based on truncated probitnormal distribution

The tests in this section nest the null hypothesis (9) in a model where the underlying reported PIT values $P_1, \ldots, P_n$ have a probitnormal distribution satisfying $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$. Writing $\boldsymbol{\theta} = (\mu, \sigma)'$, the distribution function and density of P_t are respectively

$$F_P(p \mid \boldsymbol{\theta}) = \Phi\left(\frac{\Phi^{-1}(p) - \mu}{\sigma}\right), \quad f_P(p \mid \boldsymbol{\theta}) = \frac{\phi\left(\frac{\Phi^{-1}(p)-\mu}{\sigma}\right)}{\phi(\Phi^{-1}(p))\sigma}, \quad p \in [0,1], \tag{15}$$

which gives a flexible family containing the uniform distribution, which corresponds to $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (0,1)'$. Other choices of nesting model are possible, for example a beta distribution.

The test statistics are based on the PIT values truncated to the interval $[\alpha_1, \alpha_2]$, that is, the values $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$. The likelihood contribution $\mathcal{L}(\boldsymbol{\theta} \mid P_t^*)$ of an observation P_t^* in the truncated model can be written as

$$\mathcal{L}(\boldsymbol{\theta} \mid P_t^*) = \begin{cases} F_P(\alpha_1 \mid \boldsymbol{\theta}) & P_t^* = \alpha_1, \\ f_P(P_t^* \mid \boldsymbol{\theta}) & \alpha_1 < P_t^* < \alpha_2, \\ \bar{F}_P(\alpha_2 \mid \boldsymbol{\theta}) & P_t^* = \alpha_2. \end{cases} \tag{16}$$

See (A.1) for the explicit likelihood of the sample $P_1^*, \ldots, P_n^*$.

We first consider an LRT that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Recall that (6) shows that spectrally transformed PIT values W_t are given by continuous, strictly increasing transformations of the P_t^*. Theorem 3.3 implies that the LR test of the null hypothesis that the truncated PIT values P_t^* have a truncated uniform distribution, against the alternative that they do not, is equivalent to a whole family of LR tests for the spectrally transformed PIT values under continuous weighting. In the case where $\alpha_2 = 1$, this test is also equivalent to the test proposed by Berkowitz (2001); in the case where $\alpha_2 < 1$ we obtain a generalization of the Berkowitz test–a Berkowitz interval test.[8]

An alternative to the LRT is the classical score test, which has the advantage that no maximization of the likelihood is required. It will turn out that this test is also a bispectral Z-test. Denote the observed score vector for P_t^* by

$$\boldsymbol{S}_t(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \mu} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*), \frac{\partial}{\partial \sigma} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*) \right)' \tag{17}$$

and let $\overline{\boldsymbol{S}}_n(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{t=1}^n \boldsymbol{S}_t(\boldsymbol{\theta}_0)$ be the mean of the observed score vectors under the null. The score test follows from the asymptotic distribution

$$\sqrt{n}\overline{\boldsymbol{S}}_n(\boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{d} N_2\big(\mathbf{0}, I(\boldsymbol{\theta}_0)\big),$$

where $I(\boldsymbol{\theta})$ denotes the expected Fisher information matrix. Consequently, for large n we have approximately that

$$n\overline{\boldsymbol{S}}_n(\boldsymbol{\theta}_0)' I(\boldsymbol{\theta}_0)^{-1} \overline{\boldsymbol{S}}_n(\boldsymbol{\theta}_0) \sim \chi_2^2$$

An analytical expression for $I(\boldsymbol{\theta}_0)$ is provided in Appendix B.

The following result shows that this is a bispectral test with the structure (11) under a generalization that allows some additional point mass at the endpoints of the interval

[8] Berkowitz (2001) models the data $\Phi^{-1}(P_t^*)$ with a normal $N(\mu, \sigma^2)$ distribution truncated to $[\Phi^{-1}(\alpha_1), \infty)$. This coincides with our approach because Φ^{-1} is a continuous and strictly increasing transformation and Theorem 3.3 again applies.

$[\alpha_1, \alpha_2]$.

Theorem 4.2. *$\boldsymbol{S}_t(\boldsymbol{\theta}_0) = \boldsymbol{W}_t - \boldsymbol{\mu}_W$, almost surely, where $W_{t,i}$ can be expressed as*

$$W_{t,i} = \gamma_{i,1}\mathbb{1}_{\{P_t>\alpha_1\}} + \gamma_{i,2}\mathbb{1}_{\{P_t>\alpha_2\}} + \int_{\alpha_1}^{\alpha_2} g_i(u)\mathbb{1}_{\{P_t>u\}}\mathrm{d}u$$

for $\gamma_{i,1}$, $\gamma_{i,2}$ and $g_i(u)$ with analytical solution.

5 Tests of conditional coverage

Whereas unconditional tests are focused on testing for the hypothesized distribution F_W^0 of the spectrally transformed PIT-values, conditional backtests are joint tests of the correct distribution and the independence of W_t and $\mathcal{F}_{t-1}$ for all t, as asserted by the null hypothesis (9). We have noted in Section 4 that the Z-tests presented there may have some limited power to detect the presence of serial dependencies. The aim in this section is to propose conditional extensions of our spectral tests that *explicitly* address the independence of W_t and $\mathcal{F}_{t-1}$. These tests should have more power to detect departures from the null hypothesis resulting from a failure to use all the information in $\mathcal{F}_{t-1}$ when building the predictive model $\widehat{F}_t$. In the context of risk management, where models often fail to address time-varying volatility in adequate fashion, there is a particular need for tests of this kind.

In his early paper on backtesting, Kupiec (1995) proposed a test for independence of VaR exceedances based on the fact that the spacings between them should be geometrically distributed. This latter property follows from the fact that a series of VaR exceedances should behave like a Bernoulli trials process, that is iid Bernoulli events with independent geometric waiting times.[9]

The tests that we develop below follow an alternative regression-based approach to testing conditional coverage. Christoffersen (1998) proposed an early test in this vein in which the iid Bernoulli hypothesis for VaR exceedances is tested against the alternative hypothesis that VaR exceedances show first-order Markov dependence; this has been generalized to a multilevel test by Leccadito et al. (2014). The Christofferson test can be viewed as a likelihood-ratio test that the parameters in a simple linear regression model are zero. An especially influential regression-based test is the dynamic quantile (DQ) test of Engle and Manganelli (2004), in which exceedance indicators are regressed on lagged exceedance indicators and lagged estimates of VaR to assess the null hypothesis of independent exceedances occurring at the desired rate. Our martingale difference framework generalizes the DQ test and includes a variant on the Christoffersen (1998) test.

[9] Christoffersen and Pelletier (2004) further developed the idea of testing the spacings between VaR exceedances using the fact that a discrete geometric distribution can be approximated by a continuous exponential distribution. See McNeil et al. (2015) for more details of the theory.

There are a number of other tests that are related to, but not directly subsumed by the regression-based testing approach we develop below. Berkowitz et al. (2011) suggest adapting the DQ test to use a standard link function for modelling binary response data resulting in a generalized linear regression model. Dumitrescu et al. (2012) build on this idea by considering the application to backtesting of the dynamic binary model of Kauppi and Saikkonen (2008). Hurlin and Topkavi (2007) propose a multivariate portmanteau test based on the autocorrelations of VaR exceedances at different lags and different confidence levels. Leccadito et al. (2014) propose a generalization of the Pearson multilevel test to test for independence of numbers of level exceedances across time periods. Du and Escanciano (2017) develop a Box-Pierce-type test based on a backtest statistic for expected shortfall that takes PIT values as input. Berkowitz et al. (2011) provide a comprehensive overview of tests of conditional coverage and advocate the DQ and geometric spacing tests in particular.

In the following subsections, we consider testing for the independence of transformed reported PIT-values within a regression or conditional framework. We introduce the notation $(\widetilde{W}_t)$ for the sequence of transformed reported PIT-values $\widetilde{W}_t = W_t - \mu_W$ centered at their theoretical mean μ_W under the null hypothesis (9). Recall from Section 2 that the filtration $(\mathcal{F}_t)$ represents the information available to the risk manager and that P_t is $\mathcal{F}_t$-measurable. We test that $(\widetilde{W}_t)$ has the martingale difference (MD) property with respect to $(\mathcal{F}_t)$:

$$E(\widetilde{W}_t \mid \mathcal{F}_{t-1}) = 0 \tag{18}$$

which is necessary for (9) to hold.

5.1 Conditional spectral Z-test

When MD property (18) holds, we must have $E(h_{t-1}\widetilde{W}_t) = 0$ for any $\mathcal{F}_{t-1}$-measurable random variable h_{t-1}. We form the $k+1$-dimensional lagged vector

$$\boldsymbol{h}_{t-1} = (1, h(P_{t-1}), \ldots, h(P_{t-k}))'$$

for a function h, to which we refer as a *conditioning variable transformation*. To guarantee the existence of the second moment of $\boldsymbol{h}_{t-1}$, we assume that (P_t) is covariance-stationary and that h is bounded.[10] Particular examples that we will use in our empirical analysis are $h(p) = \mathbb{1}_{\{p>\alpha\}}$ for some α and $h(p) = |2p-1|^c$ for $c > 0$.

We base our test on the vector-valued process $\boldsymbol{Y}_t = \boldsymbol{h}_{t-1}\widetilde{W}_t$ for $t = k+1, \ldots, n$. Under the null hypothesis (9), $(\boldsymbol{Y}_t)$ is a MD sequence satisfying $\mathbb{E}(\boldsymbol{Y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{0}$. We want to test that $\boldsymbol{Y}_{k+1}, \ldots, \boldsymbol{Y}_n$ are close to the zero vector on average. The conditional

[10] The restriction on h can be relaxed considerably, but in practice we find that bounded functions lead to more stable tests.

predictive test of Giacomini and White (2006) which was developed for comparing forecasting methods can be applied in this context. Let $\overline{\boldsymbol{Y}}_{n,k} = (n-k)^{-1}\sum_{t=k+1}^{n}\boldsymbol{Y}_t$ and let $\hat{\Sigma}_Y$ denote a consistent estimator of $\Sigma_Y := \mathrm{cov}(\boldsymbol{Y}_t)$. Giacomini and White show that under very weak assumptions, for large enough n and fixed k,

$$(n-k)\,\overline{\boldsymbol{Y}}'_{n,k}\,\hat{\Sigma}_Y^{-1}\,\overline{\boldsymbol{Y}}_{n,k} \sim \chi^2_{k+1}. \tag{19}$$

Giacomini and White (2006) use the estimator $\hat{\Sigma}_Y^{GW} = (n-k)^{-1}\sum_{t=k+1}^{n}\boldsymbol{Y}_t\boldsymbol{Y}_t'$ but we can use the fact that $\mathbb{E}(\widetilde{W}_t^2 \mid \mathcal{F}_{t-1}) = \sigma_W^2$ for all t under the null hypothesis (9) to form an alternative estimator. We compute that

$$\begin{aligned}\Sigma_Y = \mathbb{E}(\mathrm{cov}(\boldsymbol{Y}_t \mid \mathcal{F}_{t-1})) &= \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{Y}_t\boldsymbol{Y}_t' \mid \mathcal{F}_{t-1}\right)\right)\\ &= \mathbb{E}\left(\boldsymbol{h}_{t-1}\boldsymbol{h}_{t-1}'\mathbb{E}\left(\widetilde{W}_t^2 \mid \mathcal{F}_{t-1}\right)\right) = \sigma_W^2 H\end{aligned} \tag{20}$$

where $H = \mathbb{E}\left(\boldsymbol{h}_{t-1}\boldsymbol{h}_{t-1}'\right)$ which suggests the estimator $\hat{\Sigma}_Y = \sigma_W^2\hat{H}$ where[11]

$$\hat{H} = (n-k)^{-1}\sum_{t=k+1}^{n}\boldsymbol{h}_{t-1}\boldsymbol{h}_{t-1}'. \tag{21}$$

The decomposition in (20) has the advantage that it generalizes our unconditional spectral Z-test, which may be thought of as the case $k = 0$. The case $k = 1$ may be viewed as a Z-test version of the first-order Markov chain test of Christoffersen (1998). Moreover, as we now show, our conditional test contains as a special case the dynamic quantile (DQ) test statistic proposed by Engle and Manganelli (2004). Let X be the $(n-k)\times(k+1)$ matrix whose rows are given by $\boldsymbol{h}_{t-1}$ for $t = k+1,\ldots,n$. Let $\widetilde{\boldsymbol{W}} = (\widetilde{W}_{k+1},\ldots,\widetilde{W}_n)'$. It follows that

$$\hat{\Sigma}_Y = \sigma_W^2(n-k)^{-1}\sum_{t=k+1}^{n}\boldsymbol{h}_{t-1}\boldsymbol{h}_{t-1}' = \sigma_W^2(n-k)^{-1}X'X$$

and $\overline{\boldsymbol{Y}}_{n,k} = (n-k)^{-1}X'\widetilde{\boldsymbol{W}}$ so that (19) may be rewritten as

$$\sigma_W^{-2}\widetilde{\boldsymbol{W}}'X(X'X)^{-1}X'\widetilde{\boldsymbol{W}} \sim \chi^2_{k+1}. \tag{22}$$

The DQ test statistic of Engle and Manganelli (2004) corresponds to the binomial score case, i.e., the case where $W_t = \mathbb{1}_{\{P_t>\alpha\}}$ and the CVT is $h(p) = \mathbb{1}_{\{p>\alpha\}}$.[12]

[11] We have also experimented with the test obtained under the stronger hypothesis that the P_t are uniform, which allows us to calculate $H = \mathrm{diag}(1, \mathbb{E}(h(P_t)^2), \ldots, \mathbb{E}(h(P_t)^2))$ analytically. The resulting test has poorer size and is somewhat in conflict with our general philosophy that we should focus tests for uniformity in the region where we require the risk model to perform.

[12] Engle and Manganelli (2004) allow as well for lagged VaR values to be included as regressors, but change in portfolio composition implies that lagged VaR values are less informative than lagged PIT values.

For an alternative interpretation of our test, consider the time series regression model

$$\widetilde{W}_t = \beta_0 + \sum_{i=1}^{k} \beta_i h(P_{t-i}) + \epsilon_t, \quad t = k+1, \ldots, n \tag{23}$$

for which X is the design matrix. Under the standard assumptions for time series regression and assuming homoscedastic errors with known variance σ_W^2, the least squares estimator of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ is $(X'X)^{-1}X'\widetilde{\boldsymbol{W}}$ and this is asymptotically normal with covariance matrix $\sigma_W^2(X'X)^{-1}$. Thus expression (22) describes the natural chi-squared test that $\boldsymbol{\beta} = \mathbf{0}$.

5.2 Conditional bispectral Z-test

The conditional spectral Z-test generalizes to a conditional bispectral Z-test. We construct two sets of transformed reported PIT-values $(W_{t,1}, W_{t,2})$ for $t = 1, \ldots, n$, and form the vector $\boldsymbol{Y}_t$ of length $k_1 + k_2 + 2$ given by

$$\boldsymbol{Y}_t = \left(\boldsymbol{h}'_{t-1,1}\widetilde{W}_{t,1}, \boldsymbol{h}'_{t-1,2}\widetilde{W}_{t,2} \right)', \tag{24}$$

where $\widetilde{W}_{t,i} = W_{t,i} - \mu_{W,i}$ and $\boldsymbol{h}_{t-1,i} = (1, h_i(P_{t-1}), \ldots, h_i(P_{t-k_i}))'$. Parallel to the previous section, let $\overline{\boldsymbol{Y}}_{n,k} = (n-k)^{-1} \sum_{t=k+1}^{n} \boldsymbol{Y}_t$ for $k = k_1 \vee k_2$, and let $\hat{\Sigma}_Y$ denote a consistent estimator of $\Sigma_Y := \mathrm{cov}(\boldsymbol{Y}_t)$. By the theory of Giacomini and White (2006), for n large and (k_1, k_2) fixed,

$$(n-k)\, \overline{\boldsymbol{Y}}'_{n,k}\, \hat{\Sigma}_Y^{-1}\, \overline{\boldsymbol{Y}}_{n,k} \sim \chi^2_{k_1+k_2+2}. \tag{25}$$

Working under the null hypothesis, we can generalize (20) to $\Sigma_Y = A_W \circ H$, where $\circ$ denotes element-by-element multiplication (Hadamard product). The matrices are

$$H = \begin{pmatrix} \mathbb{E}\left(\boldsymbol{h}_{t-1,1}\boldsymbol{h}'_{t-1,1}\right) & \mathbb{E}\left(\boldsymbol{h}_{t-1,1}\boldsymbol{h}'_{t-1,2}\right) \\ \mathbb{E}\left(\boldsymbol{h}_{t-1,2}\boldsymbol{h}'_{t-1,1}\right) & \mathbb{E}\left(\boldsymbol{h}_{t-1,2}\boldsymbol{h}'_{t-1,2}\right) \end{pmatrix}$$

and

$$A_W = \begin{pmatrix} \sigma^2_{W,1} J_{k_1+1,k_1+1} & \sigma_{W,12} J_{k_1+1,k_2+1} \\ \sigma_{W,12} J_{k_2+1,k_1+1} & \sigma^2_{W,2} J_{k_2+1,k_2+1} \end{pmatrix} \tag{26}$$

where $J_{m,n}$ denotes the $m \times n$ matrix of ones and $\sigma_{W,12} = \mathbb{E}\left(\widetilde{W}_{t,1}\widetilde{W}_{t,2}\right)$. Our tests use the estimator $\hat{\Sigma}_Y = A_W \circ \hat{H}$, where $\hat{H}$ generalizes (21) as

$$\hat{H} = (n - (k_1 \vee k_2))^{-1} \sum_{t=(k_1 \vee k_2)+1}^{n} (\boldsymbol{h}'_{t-1,1}, \boldsymbol{h}'_{t-1,2})'(\boldsymbol{h}'_{t-1,1}, \boldsymbol{h}'_{t-1,2}). \tag{27}$$

5.3 Conditional probitnormal score test

The theory of the conditional bispectral test carries over to the probitnormal case. Letting $\boldsymbol{\theta} = (\mu, \beta_1, \ldots, \beta_k, \sigma)'$, consider a regression extension of (15) in which

$$F_{P_t|P_{t-1},\ldots,P_{t-k}}(p \mid \boldsymbol{\theta}, p_{t-1}, \ldots, p_{t-k}) = \Phi\left(\frac{\Phi^{-1}(p) - \mu - \sum_{i=1}^{k} \beta_k h(p_{t-i})}{\sigma}\right) \tag{28}$$

and write $f_{P_t|P_{t-1},\ldots,P_{t-k}}$ for the corresponding conditional density. This gives a dynamic model in which we can test for $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (0, \ldots, 0, 1)'$.

As in Section 4.3, we model truncated PIT values $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$, but here we condition on information about past PIT values. The likelihood contribution of an observation P_t^* in the truncated model can be written as

$$\mathcal{L}(\boldsymbol{\theta} \mid P_t^*, P_{t-1}, \ldots, P_{t-k}) = \begin{cases} F_{P_t|P_{t-1},\ldots,P_{t-k}}(\alpha_1 \mid \boldsymbol{\theta}, P_{t-1}, \ldots, P_{t-k}) & P_t^* = \alpha_1, \\ f_{P_t|P_{t-1},\ldots,P_{t-k}}(P_t^* \mid \boldsymbol{\theta}, P_{t-1}, \ldots, P_{t-k}) & \alpha_1 < P_t^* < \alpha_2, \\ \bar{F}_{P_t|P_{t-1},\ldots,P_{t-k}}(\alpha_2 \mid \boldsymbol{\theta}, P_{t-1}, \ldots, P_{t-k}) & P_t^* = \alpha_2. \end{cases} \tag{29}$$

The following result shows that the score test of the null hypothesis (9) in the regression model described by (29) takes precisely the form (24) for a conditional bispectral test.

Proposition 5.1. *The score statistic $\tilde{\boldsymbol{S}}_t(\boldsymbol{\theta})$ for the model described by* (29) *satisfies $\tilde{\boldsymbol{S}}_t(\boldsymbol{\theta}_0) = \left(\boldsymbol{h}'_{t-1,1}\widetilde{W}_{t,1}, \widetilde{W}_{t,2}\right)'$ where $\boldsymbol{h}'_{t-1,1} = (1, h(P_{t-1}), \ldots, h(P_{t-k}))'$, $\widetilde{W}_{t,i} = S_{t,i}(\boldsymbol{\theta}_0)$ and $S_{t,i}(\boldsymbol{\theta}_0)$ denotes a component of the score vector in (17).*

6 Application to bank-reported PIT values

We apply our spectral backtests to a set of ten samples of PIT values reported by US banks to the Federal Reserve Board. Due to the generality of our framework, design of such an empirical exercise involves choices along several dimensions, most notably with respect to test type (Z-test vs LRT), kernel function, and kernel window. To guide these choices, we have conducted an extensive set of simulation analyses, which are available from the authors in a companion paper. For the tests of unconditional coverage, we summarize our key findings as follows.

First, power typically increases with the width of the kernel window, but counterexamples abound. Intuitively, a test is most powerful in rejecting a false model when the kernel function weights heavily on probability levels for which the inverse cdf of the risk manager's model diverges from the true model inverse cdf. If widening the window leads to increased weight in the neighborhood of a crossing between the two cdfs, power may diminish. As historical simulation in particular tends to understate the tails of the distribution, in practice we expect that the most powerful tests will weight heavily on

extreme probability levels. However, this can come at the expense of the stability of the test, in the sense that the outcome can be determined by the presence or absence of one or two very large reported PIT-values. Furthermore, testing at extreme tail values of α runs counter to the primary regulatory motivation for the backtest, which is to verify the bank's 99% VaR.

Second, multinomial and truncated probitnormal LR tests are outperformed by the corresponding score tests. They are similar in power, but the LRT tends to be oversized. Overall, the Pearson and truncated probitnormal score tests are among the most powerful in our study, so in the exercises below we include these tests and exclude the corresponding LR tests.

Third, for the discrete tests, we find that 3-level tests perform as well as 5-level tests. Therefore, we focus on the 3-level case in the multinomial tests below.

Fourth, bispectral tests tend to be more powerful than (single-kernel) spectral tests. However, when the two kernels are too similar in shape, the gain in information from combining these kernels is insufficient to compensate for the increased degrees of freedom in the χ^2 test.

6.1 Data

Our data consist of ten confidential backtesting samples provided by US banks to the Federal Reserve Board at the subportfolio level. Mandatory reporting to bank regulators pursuant to the Market Risk Rule took effect on January 1, 2013. For each significant subportfolio and each business day, the bank is required to report the overnight VaR at the 99% level, the realized clean P&L, and the associated PIT-value (see Federal Register, 2012, p. 53105). While the first two fields have been available to regulators for a long time (at least at an aggregate trading book level), access to PIT values is new.

Each of our ten samples represents returns on an equity or foreign exchange subportfolio. We have data on both subportfolios for four banks, and for two banks we have data on only one subportfolio each. Banks have some discretion in defining subportfolios, but in general these are broader than what might be associated with a "trading desk." The equity subportfolio, for example, is likely to contain equity derivatives (vanilla and exotic) as well as cash positions. All of the samples lie within the three-year period from 2013–2015, inclusive.

Summary statistics for the unconditional distributions are found in Table 1. Six series span the entire period, and the shortest sample is about one year in length. As is often the case with new regulatory reporting requirements, data quality are not uniform. Two of the samples (coded $Pf104$ and $Pf110$) have a significant number of missing values (3.4% and 6.7%, respectively). Furthermore, close inspection reveals that most of the samples contain a small number of observations that are clearly or

very likely to be spurious, e.g., a PIT value of 1 matched to a realized loss that was smaller than the forecast VaR. We developed a heuristic procedure to identify spurious values based on the distance between the reported PIT-value and an imputed value. The latter is constructed using a portfolio-specific model that fits PIT to the ratio of realized loss to VaR; see Appendix C for details. In test results reported below, we treat spurious values as missing to make the tests less sensitive to reporting error. Our conclusions are qualitatively robust to taking all non-missing observations as valid.

Remaining columns of the table provide a histogram of PIT values. For some portfolios, the histograms appear to be unconditionally close to uniform. For example, for $Pf109$, 87.9% of PIT values lie in $[0.05, 0.95)$ and remaining mass appears to be symmetrically distributed. For some other portfolios, tail PIT values are underrepresented (e.g., $Pf104$, $Pf107$) or overrepresented (e.g., $Pf110$) in the sample.

6.2 A menagerie of tests and kernel functions

We consider kernels of discrete, continuous, and mixed form. All the backtests described below fall within our spectral Z-test class. All reported p-values are based on two-sided tests, though one-sided versions of some tests are of course available.

Parameters α_1 and α_2 control the kernel window. For the continuous tests, α_1 and α_2 are the infimum and supremum of the kernel support. For the discrete case, we consider 3-level kernels at the set of points $(\alpha_1, \alpha^*, \alpha_2)$, where $\alpha^* = 0.99$ is the conventional VaR level. We define a *narrow* window for which $\alpha_1 = 0.985$ and $\alpha_2 = 0.995$, and a *wide* window for which $\alpha_1 = 0.95$ and $\alpha_2 = 0.995$. Observe that the narrow window is symmetric around α^*, whereas the wide window is asymmetric.

For the continuous case, there are a wide variety of plausible candidates for the kernel density. Table 2 lists the kernel density functions on $[\alpha_1, \alpha_2]$ that we discuss below. The uniform and hump-shaped Epanechnikov kernels are borrowed from the nonparametric statistics literature. The exponential kernel allows for weights that are either increasing ($\zeta > 0$) or decreasing ($\zeta < 0$) in u. All but the exponential kernel are special cases of the beta kernel. In view of the flexibility of the beta kernel class, in Appendix D we provide analytical solutions for the moments of the transformed PIT values for the general beta(a, b) case.

We next list the backtests to be implemented. For use in tables later, we assign each test a mnenomic.

Binomial score test: the two-sided binomial score test at level α^* (BIN).

3-level multinomial tests: we apply the Pearson test (Pearson3) and the Z-test with discrete uniform kernel (ZU3).

Continuous spectral tests: we apply tests based on the uniform kernel (ZU); the arcsin kernel (ZA); Epanechnikov kernel (ZE); increasing (ZL_+) and decreasing

ID	Trading days	of which:		Frequencies						
		Missing	Spurious	$[0, .005)$	$[.005, .015)$	$[.015, .05)$	$[.05, .95)$	$[.95, .985)$	$[.985, .995)$	$[.995, 1]$
101	758	0	0	0.0119	0.0132	0.0290	0.8865	0.0396	0.0119	0.0079
102	751	0	8	0.0013	0.0081	0.0135	0.9341	0.0310	0.0081	0.0040
103	750	0	7	0.0121	0.0054	0.0081	0.9489	0.0175	0.0054	0.0027
104	646	22	8	0.0000	0.0000	0.0016	0.9951	0.0032	0.0000	0.0000
105	624	0	3	0.0145	0.0177	0.0290	0.8841	0.0290	0.0177	0.0081
106	252	0	0	0.0119	0.0278	0.0556	0.8333	0.0397	0.0238	0.0079
107	750	0	2	0.0000	0.0000	0.0013	0.9973	0.0013	0.0000	0.0000
108	758	0	3	0.0000	0.0053	0.0331	0.9179	0.0225	0.0119	0.0093
109	748	4	6	0.0095	0.0122	0.0420	0.8794	0.0352	0.0176	0.0041
110	646	43	7	0.0218	0.0252	0.0453	0.8389	0.0352	0.0201	0.0134

Table 1: Sample statistics.
Missing and spurious observations excluded from the reported frequencies. Trading dates for all portfolios fall between 2012-12-31 and 2015-12-31.

Kernel	Mnemonic	Density $g(u)$	Beta representation
Uniform	ZU	1	1,1
Arcsin	ZA	$1/\sqrt{u^*(1-u^*)}$	½,½
Epanechnikov	ZE	$1-(2u^*-1)^2$	2,2
Linear increasing	ZL_+	u^*	2,1
Linear decreasing	ZL_-	$1-u^*$	1,2
Exponential	ZX_ζ	$\exp(\zeta u^*)$ for some $\zeta \in \mathbb{R}$	–

Table 2: Kernel density functions on $[\alpha_1, \alpha_2]$.
u^* denotes the rescaled value $u^* = (u-\alpha_1)/(\alpha_2-\alpha_1)$. Density functions are not scaled to integrate to 1. The exponential kernel is outside the class of beta kernels, so has no beta representation.

(ZL_-) linear kernels; and increasing and decreasing exponential kernels (ZX_ζ) with parameter ζ of 2 and -2, respectively.

Continuous bispectral tests: we apply combinations of the increasing and decreasing linear kernels (ZLL), of exponential kernels with $\zeta = \pm 2$ (ZXX), and of the arcsin and Epanechnikov kernels (ZAE); we also apply the truncated probitnormal score test (PNS).

6.3 Tests of unconditional coverage

Table 3 presents p-values for the tests of unconditional coverage. When we adopt a narrow kernel window, we find that all of the tests reject at the 5% level the forecast model for portfolio $Pf104$ and at the 1% level for $Pf107$ and $Pf110$. In view of the histograms observed in Table 1, this is unsurprising. When an empirical distribution function (edf) lies above the uniform cdf within the kernel window (as observed for $Pf104$ and $Pf107$), large PIT values are underepresented in the sample, which suggests that the forecast model overstates the upper quantiles of the loss distribution. When an edf lies below the uniform cdf (as observed for $Pf110$), large PIT values are overrepresented in the sample, which suggests that the forecast model understates the upper quantiles.

For four of the portfolios ($Pf101$, $Pf102$, $Pf103$ and $Pf106$), none of the tests reject. For the remaining three portfolios ($Pf105$, $Pf108$, $Pf109$), the test p-values vary considerably across the kernel functions. This is to be expected and desirable, as the kernel functions prioritize different quantiles of the unconditional distribution.

In the upper panel of Figure 1, we plot the edf for five of the portfolios ($Pf101$, $Pf103$, $Pf104$, $Pf108$ and $Pf109$) to illustrate the differences in test performance. We see that the edf for $Pf101$ is relatively close to the theoretical uniform cdf (dot-dash line) throughout the kernel window. The edf for $Pf103$ lies well above the theoretical cdf, but still is much closer to uniform than the edf for $Pf104$. This indicates that departures from uniformity must be fairly large to generate a test rejection in backtest

ID	window	BIN	Pearson3	ZU3	ZU	ZA	ZE	ZL_+	ZL_-	ZLL	ZAE	PNS
101	narrow	0.6042	0.4302	0.3158	0.3884	0.3289	0.4476	0.4610	0.3460	0.5832	0.2427	0.5356
	wide	0.6042	0.4597	0.2196	0.5062	0.4569	0.5352	0.4257	0.5828	0.6583	0.6528	0.5047
102	narrow	0.2060	0.4623	0.3800	0.6204	0.5922	0.6524	0.5244	0.7149	0.6463	0.7543	0.8120
	wide	0.2060	0.3970	0.1554	0.3990	0.3263	0.4557	0.4672	0.3707	0.6522	0.3435	0.3861
103	narrow	0.1024	0.3994	0.1151	0.1283	0.1369	0.1197	0.1877	0.0995	0.2048	0.2816	0.2973
	wide	0.1024	0.0225	0.0050	0.0151	0.0099	0.0225	0.0329	0.0107	0.0337	0.0166	0.0096
104	narrow	0.0126	0.0246	0.0046	0.0062	0.0052	0.0075	0.0130	0.0041	0.0135	0.0166	0.0092
	wide	0.0126	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
105	narrow	0.1264	0.1657	0.0590	0.0521	0.0524	0.0542	0.0790	0.0411	0.1107	0.1521	0.0860
	wide	0.1264	0.5018	0.2750	0.0993	0.1408	0.0775	0.0509	0.1683	0.0881	0.0918	0.1222
106	narrow	0.1164	0.1503	0.0746	0.1059	0.0900	0.1269	0.2038	0.0631	0.0671	0.1444	0.0661
	wide	0.1164	0.2833	0.0872	0.0114	0.0197	0.0078	0.0105	0.0154	0.0372	0.0113	0.1040
107	narrow	0.0060	0.0098	0.0018	0.0026	0.0021	0.0032	0.0062	0.0016	0.0054	0.0069	0.0034
	wide	0.0060	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
108	narrow	0.0183	0.0948	0.0470	0.0554	0.0601	0.0477	0.0433	0.0756	0.1218	0.1233	0.1591
	wide	0.0183	0.0213	0.5723	0.7733	0.8121	0.7366	0.6979	0.4525	0.0208	0.8496	0.0925
109	narrow	0.3324	0.2055	0.3367	0.1866	0.2208	0.1646	0.3923	0.0953	0.0280	0.2318	0.1715
	wide	0.3324	0.3416	0.4150	0.2654	0.2968	0.2605	0.2061	0.3324	0.4027	0.5082	0.6326
110	narrow	0.0002	0.0015	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	wide	0.0002	0.0025	0.0011	0.0006	0.0007	0.0009	0.0001	0.0029	0.0002	0.0031	0.0002

Table 3: Tests of unconditional coverage.
We report test p-values by portfolio, kernel window, and kernel function. Narrow kernel window is $[.985, .995]$ and Wide kernel window is $[.95, .995]$.

samples of 2–3 years.

With the exception of portfolio $Pf108$, the continuous spectral and bispectral Z-tests tend to deliver lower p-values than the binomial score test. As seen in Figure 1, the edf for $Pf108$ is nearly flat in the lower half of the narrow window, and then rises sharply in the upper half. A step function at the center point $\alpha^* = 0.99$ is especially sensitive to this particular form of departure from uniformity, but its performance would not be robust to relatively small changes in a handful of observations.

In the case of $Pf109$, the forecast model is rejected (at the 5% level) only by the bispectral ZLL test. Figure 1 reveals a crossing within the narrow kernel window between the edf and the uniform cdf, which implies that the forecast model underestimates quantiles at one boundary of the kernel window and overestimates quantiles at the other boundary. We refer to this as a *slope deviation* from the uniform cdf. The overall proximity of the edf to the uniform cdf presents a challenge for single-kernel spectral tests in general. In a bispectral test, by contract, when the two kernels differ markedly in how they weight the lower and upper ends of the kernel window, the test can effectively identify slope deviations.

Backtests for portfolios $Pf103$ and $Pf106$ are most sensitive to the choice of kernel window. The associated forecast models are never rejected under the narrow window, but rejected by most of the tests for the wider window. (Of course, the binomial score test is invariant to the choice of kernel window.) For $Pf105$ and $Pf109$, however, the few rejections under the narrow window vanish under the wider window. For $Pf108$, we find that widening the window increases test sensitivity to the choice of kernel function.

EDFs for these portfolios are depicted in the lower panel of Figure 1. For portfolios $Pf103$ and $Pf106$, the edf departs most markedly from uniformity on the expanded portion $[.95, .985]$ of the wide window, whereas the edfs for $Pf105$ and $Pf109$ are relatively close to the uniform cdf within this region. Similar to what was observed for $Pf109$ within the narrow window, the ZLL test for $Pf108$ appears to be picking up the slope deviation associated with the single crossing between the edf and uniform cdf within the wide window.

For brevity, the tables omit results for the increasing and decreasing exponential kernels (ZX_{+2} and ZX_{-2}, respectively) and the bispectral test that combines them (ZXX). These exponential kernel functions coincide closely with the linear kernel functions, so we find for all portfolios that p-values are very similar when we substitute ZX_{+2} for ZL_{+}, ZE_{-2} for ZL_{-}, and ZXX for ZLL.

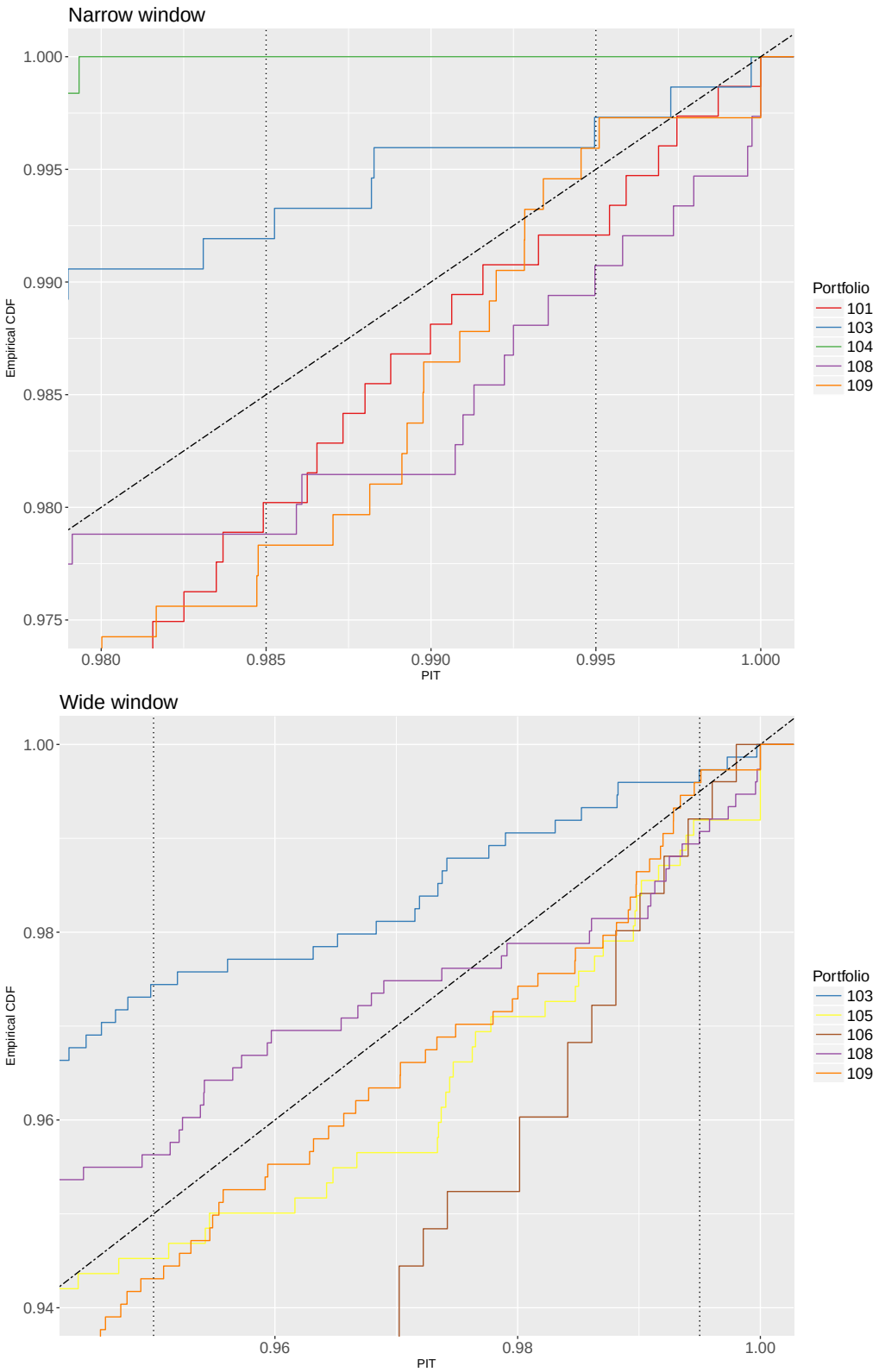

Figure 1: Empirical distribution functions for select portfolios.
EDFs for narrow window (upper panel) and wide window (lower panel). Note that the set of illustrated portfolios differs between the two panels.

6.4 Tests of conditional coverage

Tests of conditional coverage involve all the design choices of the unconditional tests, and further require the choice of the number (k) of lagged PIT values and the conditioning variable transformation $h(P)$. Define $V(u) = |2u - 1|$; this V-shaped transformation of PIT values is well-suited to uncover dependence arising from stochastic volatility. We consider four candidates for the conditioning variable transformation (CVT):

EM: $h(P) = \mathbb{1}_{\{P>0.99\}}$. This test regresses the spectrally transformed PIT-values on indicator variables for previous exceedances of the 99% VaR as in Engle and Manganelli (2004).

V.BIN: $h(P) = \mathbb{1}_{\{V(P)>0.98\}}$. This two-tailed version of EM flags PIT values near zero or one. Note that this small change requires that the regulator observe PIT values, and not only the traditional exceedance indicators.

V.4: $h(P) = V(P)^4$. Raising $V(P)$ to the fourth power places heavier weight on tail PIT values in the recent past.

V.½: $h(P) = \sqrt{V(P)}$. Relative to V.4, this transformation dampens sensitivity to tail PIT values.

Drawing guidance from simulation analyses in our companion paper, we fix $k = 4$ lags in the monospectral tests. In the context of daily backtesting, this corresponds to looking at dependencies over a time horizon of one trading week. To facilitate comparison to the monospectral tests, we fix $(k_1 = 4, k_2 = 0)$ for the bispectral tests. For parsimony, we consider only the narrow kernel window $[0.985, 0.995)$, and a subset of the kernel functions included in the previous section.

Missing or spurious values may be especially troublesome in a test of conditional coverage because a PIT value missing at time t introduces missing regressors at $t + 1, \ldots, t + k$. To avoid losing the subsequent k observations, we replace missing or spurious $P_{t-\ell}$ with an inputed value when computing the lagged vector $\boldsymbol{h}_{t-1}$. (As in the tests of unconditional coverage, we do not impute missing P_t to backfill the dependent variables W_t, but simply drop these observations.) Details of our imputation algorithm are provided in Appendix C.

Table 4 presents p-values for the tests of conditional coverage. For portfolios $Pf108$ and $Pf110$, forecast models are strongly rejected (0.01% level) regardless of the choice of CVT or kernel function; for brevity we drop these portfolios from the table. For only a single portfolio ($Pf109$), the forecast model is never rejected. In the other seven cases, the choice of CVT and kernel function matter. We find:

- For portfolios $Pf102$, $Pf103$ and $Pf105$, the V.4 CVT generally leads to rejection at the 5% level, but tests using the EM CVT never reject. The V.BIN and V.½

ID	CVT	BIN	ZU	ZL_+	ZL_-	ZLL	PNS
101	EM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	V.BIN	0.1450	0.0158	0.0145	0.0209	0.0361	0.0601
	V.4	0.0599	0.0183	0.0102	0.0305	0.0512	0.0529
	V.½	0.4504	0.3084	0.3396	0.2721	0.3633	0.2928
102	EM	0.8987	0.9960	0.9926	0.9977	0.9838	0.9970
	V.BIN	0.8785	0.9045	0.9726	0.7709	0.7721	0.8222
	V.4	0.3313	0.0418	0.1087	0.0185	0.0261	0.0393
	V.½	0.4683	0.1628	0.3167	0.0819	0.1042	0.1472
103	EM	0.7530	0.8042	0.8838	0.7445	0.7877	0.8754
	V.BIN	0.0226	0.0124	0.0061	0.0275	0.0423	0.0149
	V.4	0.0788	0.0256	0.0277	0.0305	0.0466	0.0157
	V.½	0.3834	0.2512	0.3210	0.2233	0.2837	0.2326
104	EM	NA	NA	NA	NA	NA	NA
	V.BIN	NA	NA	NA	NA	NA	NA
	V.4	0.2889	0.1903	0.2935	0.1471	0.2005	0.1564
	V.½	0.2889	0.1903	0.2935	0.1471	0.2005	0.1564
105	EM	0.6178	0.3689	0.4902	0.3095	0.4010	0.3265
	V.BIN	0.4124	0.0637	0.2813	0.0079	0.0144	0.0133
	V.4	0.2355	0.0078	0.0862	0.0006	0.0013	0.0002
	V.½	0.3196	0.0214	0.0935	0.0049	0.0092	0.0009
106	EM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	V.BIN	0.0098	0.0001	0.0003	0.0000	0.0000	0.0000
	V.4	0.0088	0.0019	0.0103	0.0005	0.0005	0.0002
	V.½	0.0485	0.0418	0.1425	0.0155	0.0137	0.0073
107	EM	NA	NA	NA	NA	NA	NA
	V.BIN	NA	NA	NA	NA	NA	NA
	V.4	0.1850	0.1076	0.1889	0.0772	0.1090	0.0787
	V.½	0.1851	0.1076	0.1889	0.0772	0.1090	0.0787
109	EM	0.8836	0.6208	0.9293	0.2894	0.1021	0.2545
	V.BIN	0.4884	0.3959	0.6654	0.1910	0.0658	0.1797
	V.4	0.8716	0.7150	0.9099	0.4371	0.1606	0.4444
	V.½	0.3425	0.2560	0.3632	0.1638	0.0561	0.2181

Table 4: Tests of conditional coverage.
We report test p-values by portfolio, conditioning variable transformation, and kernel function. The monospectral tests utilize $k = 4$ lags, and for the bispectral tests we set $(k_1 = 4, k_2 = 0)$. We fix a narrow kernel window of $[.985, .995]$. Forecast models for $Pf108$ and $Pf110$ (not tabulated) are rejected at the 0.01% level for all choices of CVT and kernel.

CVT are less robust in performance than V.4. This reflects the greater sensitivity of the V.4 transformation to local spikes in market volatility.

- Only in the case of $Pf101$ does the Engle-Manganelli CVT pick up serial dependence more effectively than the CVT based on $V(P)$, though here too the V.BIN and V.4 CVT lead to rejection at the 5% for uniform and linear kernel functions.

- In two cases ($Pf104$, $Pf107$), the test statistic is undefined for the EM CVT and its two-tailed counterparty (V.BIN). As there were no observed violations in either tail ($P_t < .01$ or $P_t > .99$), in both cases the matrix $\hat{H}$ of (21) is singular, so $\hat{\Sigma}_Y$ in the test statistic cannot be inverted. This demonstrates a practical limitation of a binary CVT, as short samples may often contain no tail values.

- Despite the adoption of a narrow kernel window in these tests, the spectral backtests often give improvements in power over the traditional binomial score test. In particular, for portfolios $Pf102$, $Pf103$ and $Pf105$, p-values for tests using the continuous kernel functions are often much lower than p-values for corresponding test using the BIN kernel.

7 Conclusion

The class of spectral backtests embeds many of the most widely used tests of unconditional coverage and tests of conditional coverage, including the binomial likelihood ratio test of Kupiec (1995), the interval likelihood ratio test of Berkowitz (2001), and the dynamic quantile test of Engle and Manganelli (2004). As we demonstrate with many examples, viewing these tests in terms of the associated kernel functions facilitates the construction of new tests. From the perspective of the practice of risk management, making explicit the choice of kernel function may help to discipline the backtesting process because the kernel function directly expresses the user's priorities for model performance.

Our results illustrate the value to regulators of access to bank-reported PIT-values. Until recently, regulators effectively observed only a sequence of VaR exceedance event indicators at a single level α, and therefore backtests were designed to take such data as input. In some jurisdictions, including the United States, PIT-values have been collected for some time. Besides opening the possibility of forming spectral test statistics, we have demonstrated that lagged PIT-values are especially effective as conditioning variables in regression-based tests of conditional coverage.

There is a growing literature on multivariate or multi-desk backtesting including Wied et al. (2016) and Berkowitz et al. (2011) (see §4.4 and the CavMult test in Table 7, specifically). The new standard for capital requirements for market risk (Basel Committee on Bank Supervision, 2016) calls for backtesting at individual desk level

and typical investment banks may have in excess of 50 desks. The spectral and bispectral tests that we propose in this paper admit multi-desk generalizations that allow the simultaneous evaulation of backtest results across multiple desks. We leave this as a topic for future work.

A Proofs

A.1 Proofs of Propositions 3.1 and 3.2

The logic of these two proofs is identical and we give the proof of Proposition 3.2 only.

$$\begin{aligned}
W_{t,1}W_{t,2} &= \left(\int_{\alpha_1}^{\alpha_2} g_1(u)\mathbb{1}_{\{P_t>u\}}\mathrm{d}u\right)\left(\int_{\alpha_1}^{\alpha_2} g_2(v)\mathbb{1}_{\{P_t>v\}}\mathrm{d}v\right)\\
&= \int_{u=\alpha_1}^{\alpha_2}\int_{v=\alpha_1}^{\alpha_2} g_1(u)g_2(v)\mathbb{1}_{\{P_t>u\}}\mathbb{1}_{\{P_t>v\}}\mathrm{d}v\mathrm{d}u\\
&= \int_{u=\alpha_1}^{\alpha_2}\int_{v=\alpha_1}^{\alpha_2} g_1(u)g_2(v)\mathbb{1}_{\{P_t>\max\{u,v\}\}}\mathrm{d}v\mathrm{d}u\\
&= \int_{u=\alpha_1}^{\alpha_2}\int_{v=\alpha_1}^{u} g_1(u)g_2(v)\mathbb{1}_{\{P_t>u\}}\mathrm{d}v\mathrm{d}u + \int_{u=\alpha_1}^{\alpha_2}\int_{v=u}^{\alpha_2} g_1(u)g_2(v)\mathbb{1}_{\{P_t>v\}}\mathrm{d}v\mathrm{d}u\\
&= \int_{u=\alpha_1}^{\alpha_2} g_1(u)\left(\int_{v=\alpha_1}^{u} g_2(v)\mathrm{d}v\right)\mathbb{1}_{\{P_t>u\}}\mathrm{d}u + \int_{v=\alpha_1}^{\alpha_2} g_2(v)\left(\int_{u=\alpha_1}^{v} g_1(u)\mathrm{d}u\right)\mathbb{1}_{\{P_t>v\}}\mathrm{d}v\\
&= \int_{u=\alpha_1}^{\alpha_2} g_1(u)G_2(u)\mathrm{d}u + \int_{v=\alpha_1}^{\alpha_2} g_2(v)G_1(v)\mathrm{d}v
\end{aligned}$$

Note that $g^*(u)$ clearly satisfies Assumption 1. If g_1 and g_2 are normalized kernel densities on $[\alpha_1, \alpha_2]$ then it follows that

$$\int_{\alpha_1}^{\alpha_2} g^*(u)\mathrm{d}u = \Big[G_1(u)G_2(u)\Big]_{\alpha_1}^{\alpha_2} = 1.$$

A.2 Proof of Theorem 3.3

The likelihood $\mathcal{L}_P(\theta \mid \boldsymbol{P}^*)$ takes the form

$$\mathcal{L}_P(\theta \mid \boldsymbol{P}^*) = \prod_{t\,:\,P_t^*=\alpha_1} F_P(\alpha_1 \mid \theta) \prod_{t\,:\,\alpha_1<P_t^*<\alpha_2} f_P(P_t^* \mid \theta) \prod_{t\,:\,P_t^*=\alpha_2} \bar{F}_P(\alpha_2 \mid \theta) \qquad \text{(A.1)}$$

where $\bar{F}(u)$ denotes the tail probability $1 - F(u)$. Since T is strictly increasing and continuous on $[\alpha_1, \alpha_2]$, the distribution $F_W(w \mid \theta)$ implied by $F_P(p \mid \theta)$ satisfies

$$\begin{aligned}
\mathbb{P}(W = T(\alpha_1) \mid \theta) &= F_P(\alpha_1 \mid \theta),\\
f_W(w \mid \theta) &= \frac{f_P(T^{-1}(w) \mid \theta)}{T'(T^{-1}(w))}, \quad w \in (T(\alpha_1), T(\alpha_2)),\\
\mathbb{P}(W = T(\alpha_2) \mid \theta) &= \bar{F}_P(\alpha_2 \mid \theta).
\end{aligned}$$

It follows that the likelihood $\mathcal{L}_W(\theta \mid \boldsymbol{W})$ is given by

$$\begin{aligned}\mathcal{L}_W(\theta \mid \boldsymbol{W}) &= \prod_{t\,:\,W_t=T(\alpha_1)} F_P(\alpha_1 \mid \theta) \prod_{t\,:\,T(\alpha_1)<W_t<T(\alpha_2)} \frac{f_P(T^{-1}(W_t) \mid \theta)}{T'(T^{-1}(W_t))} \prod_{t\,:\,W_t=T(\alpha_2)} \bar{F}_P(\alpha_2 \mid \theta) \\ &= \frac{\mathcal{L}_P(\theta \mid \boldsymbol{P}^*)}{\prod_{t\,:\,\alpha_1<P_t^*<\alpha_2} T'(P_t^*)}.\end{aligned}$$

It is clear that the same value $\hat{\theta}$ must maximize both these likelihoods and that the likelihood ratio statistics must satisfy

$$\mathrm{LR}_{W,n} = \frac{\mathcal{L}_W(\theta_0 \mid \boldsymbol{W})}{\mathcal{L}_W(\hat{\theta} \mid \boldsymbol{W})} = \frac{\mathcal{L}_P(\theta_0 \mid \boldsymbol{P}^*)}{\mathcal{L}_P(\hat{\theta} \mid \boldsymbol{P}^*)} = \mathrm{LR}_{P,n}.$$

A.3 Sketch of proof of Theorem 4.1

The Pearson test is one of the best known tests in statistics. The result can be proved by adapting an approach that is used to derive the asymptotic distribution of the Pearson test statistic.

Let $\boldsymbol{X}_t = (X_{t,0}, \ldots, X_{t,m})'$ be the $(m+1)$-dimensional random vector with $X_{t,i} = \mathbb{1}_{\{\mathbf{1}'\boldsymbol{W}_t=i\}}$ for $i = 0, \ldots, m$. Under (9) $\boldsymbol{X}_t$ has a multinomial distribution satisfying $\mathbb{E}(X_{t,i}) = \theta_i$, $\mathrm{var}(X_{t,i}) = \theta_i(1-\theta_i)$ and $\mathrm{cov}(X_{t,i}, X_{t,j}) = -\theta_i\theta_j$ for $i \neq j$.

Suppose we define $\boldsymbol{Y}_t$ to be the m-dimensional random vector obtained from $\boldsymbol{X}_t$ by omitting the first component. Then $\mathbb{E}(\boldsymbol{Y}_t) = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ and Σ_Y is the $m \times m$ submatrix of $\mathrm{cov}(\boldsymbol{X}_t)$ resulting from deletion of the first row and column. A standard approach to the asymptotics of the Pearson test is to show that

$$S_m = \sum_{i=0}^{m} \frac{(O_i - n\theta_i)^2}{n\theta_i} = \sum_{i=0}^{m} \frac{(\sum_{t=1}^{n} X_{t,i} - n\theta_i)^2}{n\theta_i} = n(\overline{\boldsymbol{Y}} - \boldsymbol{\theta})'\Sigma_Y^{-1}(\overline{\boldsymbol{Y}} - \boldsymbol{\theta}),$$

where $\overline{\boldsymbol{Y}} = n^{-1}\sum_{t=1}^{n} \boldsymbol{Y}_t$. The central limit theorem is then applied to $\overline{\boldsymbol{Y}}$ to argue that $S_m \sim \chi^2_m$ in the limit.

Let A be the $m \times m$ matrix with rows given by $(\boldsymbol{e}_1 - \boldsymbol{e}_2, \boldsymbol{e}_2 - \boldsymbol{e}_3, \ldots, \boldsymbol{e}_m)$ where $\boldsymbol{e}_i$ denotes the ith unit vector. The inverse of this matrix is the upper triangular matrix of one's. It may be verified that $\boldsymbol{Y}_t = A\boldsymbol{W}_t$, $\boldsymbol{\theta} = A\boldsymbol{\mu}_W$ and $\Sigma_W = A^{-1}\Sigma_Y(A')^{-1}$. We note that $\boldsymbol{\mu}_W = (1-\alpha_1, \ldots, 1-\alpha_m)'$ and that Σ_W is a matrix with diagonal entries $\mathrm{var}(W_{t,i}) = \alpha_i(1-\alpha_i)$ and off-diagonal entries $\mathrm{cov}(W_{t,i}, W_{t,j}) = \min(\alpha_i, \alpha_j)(1 - \max(\alpha_i, \alpha_j))$ for $i, j \in \{1, \ldots, m\}$. It follows that

$$S_m = n(\overline{\boldsymbol{Y}}-\boldsymbol{\theta})'\Sigma_Y^{-1}(\overline{\boldsymbol{Y}}-\boldsymbol{\theta}) = n(\overline{\boldsymbol{W}}-\boldsymbol{\mu}_W)'A'\Sigma_Y^{-1}A(\overline{\boldsymbol{W}}-\boldsymbol{\mu}_W) = n(\overline{\boldsymbol{W}}-\boldsymbol{\mu}_W)'\Sigma_W^{-1}(\overline{\boldsymbol{W}}-\boldsymbol{\mu}_W).$$

A.4 Proof of Theorem 4.2

Computing the score statistic and evaluating it at $\boldsymbol{\theta}_0 = (0,1)'$ yields

$$\boldsymbol{S}_t(\boldsymbol{\theta}_0) = \begin{cases} \boldsymbol{\psi}_1(\alpha_1) & P_t^* = \alpha_1, \\ \boldsymbol{\psi}_*(P_t^*) & \alpha_1 < P_t^* < \alpha_2, \\ \boldsymbol{\psi}_2(\alpha_2) & P_t^* = \alpha_2. \end{cases} \tag{A.2}$$

where

$$\begin{aligned} \boldsymbol{\psi}_1(u) &= \begin{pmatrix} -\phi(\Phi^{-1}(u))/u \\ -\phi(\Phi^{-1}(u))\Phi^{-1}(u)/u \end{pmatrix} \\ \boldsymbol{\psi}_*(u) &= \begin{pmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(u)^2 - 1 \end{pmatrix} \\ \boldsymbol{\psi}_2(u) &= \begin{pmatrix} \phi(\Phi^{-1}(u))/(1-u) \\ \phi(\Phi^{-1}(u))\Phi^{-1}(u)/(1-u) \end{pmatrix} \end{aligned}$$

The jumps at α_1 and α_2 are given by

$$(\gamma_{1,1}, \gamma_{2,1})' = \boldsymbol{\psi}_*(\alpha_1) - \boldsymbol{\psi}_1(\alpha_1), \qquad (\gamma_{1,2}, \gamma_{2,2})' = \boldsymbol{\psi}_2(\alpha_2) - \boldsymbol{\psi}_*(\alpha_2)$$

The weighting functions can be obtained by differentiating $\boldsymbol{\psi}_*(u)$ with respect to u on (α_1, α_2) and are thus

$$g_1(u) = \frac{1}{\phi(\Phi^{-1}(u))}, \qquad g_2(u) = \frac{2\Phi^{-1}(u)}{\phi(\Phi^{-1}(u))}.$$

Finally, since $\boldsymbol{\mu}_{\boldsymbol{W}} = \boldsymbol{W}_t - \boldsymbol{S}_t(\boldsymbol{\theta}_0)$, we must have that $\boldsymbol{\mu}_{\boldsymbol{W}} = -\boldsymbol{\psi}_1(\alpha_1)$.

A.5 Sketch of proof of Proposition 5.1

It may be verified that the partial derivatives $\frac{\partial}{\partial \mu} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*, P_{t-1}, \ldots, P_{t-k})$ and $\frac{\partial}{\partial \sigma} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*, P_{t-1}, \ldots, P_{t-k})$ take the same essential form as the partial derivatives of (16), from which it follows that $\tilde{S}_{t,1}(\boldsymbol{\theta}_0)$ and $\tilde{S}_{t,2+k}(\boldsymbol{\theta}_0)$ coincide with $S_{t,1}(\boldsymbol{\theta}_0)$ and $S_{t,2}(\boldsymbol{\theta}_0)$ respectively. Moreover,

$$\frac{\partial}{\partial \beta_i} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*, P_{t-1}, \ldots, P_{t-k}) = h(P_{t-i}) \frac{\partial}{\partial \mu} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*, P_{t-1}, \ldots, P_{t-k}),$$

hence $\tilde{S}_{t,1+i}(\boldsymbol{\theta}_0) = h(P_{t-i}) S_{t,1}(\boldsymbol{\theta}_0)$ for $i = 1, \ldots, k$.

B Probitnormal score test

The following identities are useful for dealing with the probitnormal distribution:

$$\int_{\alpha_1}^{\alpha_2} \Phi^{-1}(u)\mathrm{d}u = \phi(\Phi^{-1}(\alpha_1)) - \phi(\Phi^{-1}(\alpha_2)) \tag{B.1}$$

$$\int_{\alpha_1}^{\alpha_2} \left(\Phi^{-1}(u)^2 - 1\right) \mathrm{d}u = \Phi^{-1}(\alpha_1)\phi(\Phi^{-1}(\alpha_1)) - \Phi^{-1}(\alpha_2)\phi(\Phi^{-1}(\alpha_2)). \tag{B.2}$$

Let $\xi(p \mid \boldsymbol{\theta}) = (\Phi^{-1}(p) - \mu)/\sigma$. The first derivatives of the log-likelihood of the truncated probitnormal distribution are

$$\frac{\partial}{\partial \mu} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*) = \begin{cases} -\frac{\phi\left(\xi(\alpha_1|\boldsymbol{\theta})\right)}{\sigma\Phi\left(\xi(\alpha_1|\boldsymbol{\theta})\right)} & P_t^* = \alpha_1, \\ -\frac{\xi\left(P_t^*|\boldsymbol{\theta}\right)}{\sigma} & \alpha_1 < P_t^* < \alpha_2, \\ \frac{\phi\left(\xi(\alpha_2|\boldsymbol{\theta})\right)}{\sigma\overline{\Phi}\left(\xi(\alpha_2|\boldsymbol{\theta})\right)} & P_t^* = \alpha_2, \end{cases} \tag{B.3}$$

and

$$\frac{\partial}{\partial \sigma} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*) = \begin{cases} -\frac{\phi\left(\xi(\alpha_1|\boldsymbol{\theta})\right)\xi(\alpha_1|\boldsymbol{\theta})}{\sigma\Phi\left(\xi(\alpha_1|\boldsymbol{\theta})\right)} & P_t^* = \alpha_1, \\ -\frac{\xi\left(P_t^*|\boldsymbol{\theta}\right)^2+1}{\sigma} & \alpha_1 < P_t^* < \alpha_2, \\ \frac{\phi\left(\xi(\alpha_2|\boldsymbol{\theta})\right)\xi(\alpha_2|\boldsymbol{\theta})}{\sigma\overline{\Phi}\left(\xi(\alpha_2|\boldsymbol{\theta})\right)} & P_t^* = \alpha_2. \end{cases} \tag{B.4}$$

Recall that the expected Fisher information matrix is defined as

$$I(\boldsymbol{\theta})_{ij} = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*)\right).$$

The conditional second derivatives of the log-likelihood are

$$-\frac{\partial^2}{\partial \mu^2} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*) = \begin{cases} \frac{\phi(\xi(\alpha_1|\boldsymbol{\theta}))\left(\phi(\xi(\alpha_1|\boldsymbol{\theta}))+\xi(\alpha_1|\boldsymbol{\theta})\Phi(\xi(\alpha_1|\boldsymbol{\theta}))\right)}{\sigma^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))^2} & P_t^* = \alpha_1, \\ \frac{1}{\sigma^2} & \alpha_1 < P_t^* < \alpha_2, \\ \frac{\phi(\xi(\alpha_2|\boldsymbol{\theta}))\left(\phi(\xi(\alpha_2|\boldsymbol{\theta}))-\xi(\alpha_2|\boldsymbol{\theta})\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))\right)}{\sigma^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))^2} & P_t^* = \alpha_2, \end{cases} \tag{B.5}$$

$$-\frac{\partial^2}{\partial \sigma^2} \ln \mathcal{L}(\boldsymbol{\theta} \mid P_t^*) = \begin{cases} \frac{\phi(\xi(\alpha_1|\boldsymbol{\theta}))\left(\xi(\alpha_1|\boldsymbol{\theta})^2\phi(\xi(\alpha_1|\boldsymbol{\theta}))+\xi(\alpha_1|\boldsymbol{\theta})^3\Phi(\xi(\alpha_1|\boldsymbol{\theta}))-2\xi(\alpha_1|\boldsymbol{\theta})\Phi(\xi(\alpha_1|\boldsymbol{\theta}))\right)}{\sigma^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))^2} & P_t^* = \alpha_1, \\ \frac{3\xi(P_t^*|\boldsymbol{\theta})^2-1}{\sigma^2} & \alpha_1 < P_t^* < \alpha_2, \\ \frac{\phi(\xi(\alpha_2|\boldsymbol{\theta}))\left(\xi(\alpha_2|\boldsymbol{\theta})^2\phi(\xi(\alpha_2|\boldsymbol{\theta}))-\xi(\alpha_2|\boldsymbol{\theta})^3\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))+2\xi(\alpha_2|\boldsymbol{\theta})\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))\right)}{\sigma^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))^2} & P_t^* = \alpha_2, \end{cases} \tag{B.6}$$

$$-\frac{\partial^2}{\partial\mu\partial\sigma}\ln\mathcal{L}(\boldsymbol{\theta}\mid P_t^*)=\begin{cases}\frac{\phi(\xi(\alpha_1|\boldsymbol{\theta}))\Big(\phi(\xi(\alpha_1|\boldsymbol{\theta}))\xi(\alpha_1|\boldsymbol{\theta})-\Phi(\xi(\alpha_1|\boldsymbol{\theta}))+\xi(\alpha_1|\boldsymbol{\theta})^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))\Big)}{\sigma^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))^2} & P_t^*=\alpha_1,\\ \frac{2\xi(P_t^*|\boldsymbol{\theta})}{\sigma^2} & \alpha_1<P_t^*<\alpha_2,\\ \frac{\phi(\xi(\alpha_2|\boldsymbol{\theta}))\Big(\phi(\xi(\alpha_2|\boldsymbol{\theta}))\xi(\alpha_2|\boldsymbol{\theta})+\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))-\xi(\alpha_2|\boldsymbol{\theta})^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))\Big)}{\sigma^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))^2} & P_t^*=\alpha_2.\end{cases}\tag{B.7}$$

By taking expectations using (B.1) and (B.2) and evaluating at $\boldsymbol{\theta}_0=(0,1)'$ we obtain the elements of $I(\boldsymbol{\theta}_0)$:

$$\begin{aligned}I(\boldsymbol{\theta}_0)_{1,1}&=\phi(\Phi^{-1}(\alpha_1))^2/\alpha_1+\phi(\Phi^{-1}(\alpha_2))^2/(1-\alpha_2)\\&\quad+\phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)-\phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2)+(\alpha_2-\alpha_1),\end{aligned}\tag{B.8}$$

$$\begin{aligned}I(\boldsymbol{\theta}_0)_{2,2}&=\phi(\Phi^{-1}(\alpha_1))^2\Phi^{-1}(\alpha_1)^2/\alpha_1+\phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)^3\\&\quad+\phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)+\phi(\Phi^{-1}(\alpha_2))^2\Phi^{-1}(\alpha_2)^2/(1-\alpha_2)\\&\quad-\phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2)^3-\phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2)+2(\alpha_2-\alpha_1),\end{aligned}\tag{B.9}$$

$$\begin{aligned}I(\boldsymbol{\theta}_0)_{1,2}&=\phi(\Phi^{-1}(\alpha_1))^2\Phi^{-1}(\alpha_1)/\alpha_1+\phi(\Phi^{-1}(\alpha_1))\big(1+\Phi^{-1}(\alpha_1)^2\big)\\&\quad+\phi(\Phi^{-1}(\alpha_2))^2\Phi^{-1}(\alpha_2)/(1-\alpha_2)-\phi(\Phi^{-1}(\alpha_2))\big(1+\Phi^{-1}(\alpha_2)^2\big).\end{aligned}\tag{B.10}$$

C Identification of spurious PIT values

Consider a stylized Gaussian model in which loss is given by

$$L_t=\sigma_{t-1}Z_t\tag{C.1}$$

where (Z_t) is an iid sequence of standard normal random variables and volatility σ_{t-1} is $\mathcal{F}_{t-1}$-measurable. Time variation in σ_t may arise from stochastic volatility or from changes over time in portfolio composition. Suppose that the risk-manager knows the true underlying distribution and the volatility. The risk-manager's ideal value-at-risk forecast at $\alpha=0.99$ is then

$$\widehat{\mathrm{VaR}}_t=\Phi^{-1}(0.99)\sigma_{t-1}$$

where Φ is the standard normal cdf. We do not observe σ_{t-1}, but from observing L_t and $\widehat{\mathrm{VaR}}_t$, we can back out the realized value of Z_t as

$$Z_t=\Phi^{-1}(0.99)\times L_t/\widehat{\mathrm{VaR}}_t.\tag{C.2}$$

Furthermore, the PIT values can be expressed as

$$P_t = \widehat{F}_{t-1}(L_t) = \Phi(L_t/\sigma_{t-1}) = \Phi(Z_t). \tag{C.3}$$

In general, we would not expect the Z_t to be Gaussian, so (C.3) will not hold. However, so long as (Z_t) is iid, there will still be a monotonic relationship between Z_t (as defined by (C.2)) and P_t. We find that the predicted relationship holds qualitatively for all bank-reported portfolios, but with more noise in some portfolios than in others. This suggests that we can use violations of monotonicity to identify spurious PIT values, but the threshold for identification must vary across portfolios.

Let $H(z;\theta_i) : \mathbb{R} \to [0,1]$ be a family of fitting functions with parameter θ_i for portfolio i, and replace (C.3) by

$$P_{i,t} = H(Z_{i,t};\theta_i) + \epsilon_{i,t} \tag{C.4}$$

where the $\epsilon_{i,t}$ are white-noise residuals. Since the H function should be increasing, it is convenient to take H to be a cdf, even though it does not have a statistical interpretation in our context. For convenience, we take H to be the normal cdf with unrestricted (μ_i, σ_i) as θ_i.

For each portfolio i, we proceed as follows:

1. Fit θ_i by nonlinear least squares, and construct residuals $\epsilon_{it} = P_{it} - H(Z_{it};\hat{\theta}_i)$.

2. The (ϵ_{it}) are bounded in the open interval $(-1,1)$, because $H(Z_{it})$ does not produce boundary values. We model ϵ_{it} as drawn from a rescaled beta distribution on $(-1,1)$ with parameters $(a = \tau_i/2, b = \tau_i/2)$. This distribution has mean zero and variance $1/(\tau_i+1)$, so we simply fit τ_i to the variance of the regression residuals.

3. Let $B(\epsilon;\hat{\tau}_i)$ be the fitted beta distribution. We flag an observation P_{it} as spurious whenever $B(\epsilon_{it};\hat{\tau}_i) < q/2$ or $B(\epsilon_{it};\hat{\tau}_i) > 1-q/2$, where q is a tolerance parameter.

4. We reestimate τ_i as in step 3 on a sample that excludes the spurious observations. Repeat step 4 with the updated $\hat{\tau}_i$. An observation is flagged as spurious if it is rejected in *either* round of estimation.

In our baseline procedure, we set the tolerance parameter to $q = 10^{-5}$, which is intended to flag only the most egregious inconsistencies between P_{it} and the pair $(L_{it}, \widehat{\mathrm{VaR}}_{it})$. A typical case involves a PIT value very close to zero or one associated with a modest P&L such that $|L_{it}| < \widehat{\mathrm{VaR}}_{it}$. Setting $q = 0$ is equivalent to shutting down the identification of spurious values.

The procedure yields *imputed* PIT values as $\hat{P}_{it} = H(Z_{it};\hat{\theta}_i)$. As noted in Section 6.4, we use the imputed values to fill in for spurious values in forming regressors in the tests of conditional coverage.

D Moments for the beta kernel

We provide a general solution to the moments and cross-moments of the transformed PIT values when the kernel densities take the form

$$g(u) = \frac{(u-\alpha_1)^{a-1}(\alpha_2-u)^{b-1}}{(\alpha_2-\alpha_1)^{a+b-1}B(a,b)}$$

for parameters $(a > 0, b > 0)$ and $\alpha_1 \leqslant u \leqslant \alpha_2$. The normalization guarantees that $G(\alpha_2) = 1$, and helps align the solution with standard beta distribution functions provided by statistical packages. In `R` notation, the kernel function is simply

$$G(u) = \text{pbeta}\left(\frac{\max\{\alpha_1, \min\{u, \alpha_2\}\} - \alpha_1}{\alpha_2 - \alpha_1}, a, b\right).$$

Solving for moments and cross-moments of kernels $(g_1(P), g_2(P))$ for uniform P involves the following integral:

$$\begin{aligned} M(a_1,b_1,a_2,b_2) &= \int_{\alpha_1}^{\alpha_2} (1-u)g_1(u)G_2(u)du \\ &= \frac{B(a_1+a_2, 1+b_1)}{a_2 B(a_1,b_1)B(a_2,b_2)} {}_3F_2(a_2, a_1+a_2, 1-b_2; 1+a_2, 1+a_1+a_2+b_1; 1) \\ &= \frac{B(a_1+a_2, 1+b_1+b_2)}{a_2 B(a_1,b_1)B(a_2,b_2)} {}_3F_2(1, a_1+a_2, a_2+b_2; 1+a_2, 1+a_1+a_2+b_1+b_2; 1) \end{aligned} \tag{D.1}$$

where ${}_3F_2(c_1, c_2, c_3; d_1, d_2; 1)$ denotes a hypergeometric function of order $(3, 2)$ and argument unity. The final line follows from the Thomae transformation T7 in Milgram (2010, Appendix A). Due to the normalization of the kernels, M does not depend on the choice of kernel window.

When its parameters are all positive, as in the final expression for M, computing ${}_3F_2(c_1, c_2, c_3; d_1, d_2; 1)$ is straightforward via the standard hypergeometric series expansion. In practice, we are most often interested in integer-valued cases for which M has a simple closed-form solution.

For given kernel window and PIT value, let $W_{a,b}$ be the transformed PIT value under a beta kernel with parameters (a, b). A recurrence rule for the incomplete beta function (Abramowitz and Stegun, 1965, eq. 6.6.7) leads to a linear relationship among "neighboring" transformations:

$$(a+b)W_{a,b} = aW_{a+1,b} + bW_{a,b+1} \tag{D.2}$$

An immediate implication is that the uniform, linear increasing and linear decreasing transformations (parameter sets (1,1), (2,1) and (1,2), respectively) are linearly de-

pendent. Any pair of these kernels would yield an equivalent bispectral test, and a trispectral test using all three kernels would be undefined due to a singular covariance matrix Σ_W. By iterating the recurrence relationship, we can derive linear relationships among sets of kernels with integer-valued parameter differences $a_i - a_\ell$ and $b_i - b_\ell$, which would lead to redundancies among the corresponding j-spectral tests.

References

Abramowitz, M., and I. A. Stegun, eds., 1965, *Handbook of Mathematical Functions* (Dover Publications, New York).

Acerbi, C., and B. Szekely, 2014, Back-testing expected shortfall, *Risk* 1–6.

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

Barone-Adesi, G., F. Bourgoin, and K. Giannopoulos, 1998, Don't look back, *Risk* 11, 100–103.

Basel Committee on Bank Supervision, 2013, Fundamental review of the trading book: A revised market risk framework, Publication No. 265, Bank for International Settlements.

Basel Committee on Bank Supervision, 2016, Minimum capital requirements for market risk, Publication No. 352, Bank for International Settlements.

Berkowitz, J., 2001, Testing the accuracy of density forecasts, applications to risk management, *Journal of Business & Economic Statistics* 19, 465–474.

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

Berkowitz, J., and J. O'Brien, 2002, How accurate are Value-at-Risk models at commercial banks?, *The Journal of Finance* 57, 1093–1112.

Billingsley, P., 1961, The Lindeberg–Lévy theorem for martingales, *Proceedings of the American Mathematical Society* 12, 788–792.

Board of Governors of the Federal Reserve System, 2011, Supervisory guidance on model risk management, SR Letter 11-7.

Cai, Y., and K. Krishnamoorthy, 2006, Exact size and power properties of five tests for multinomial proportions, *Communications in Statistics - Simulation and Computation* 35, 149–160.

Campbell, S.D., 2006, A review of backtesting and backtesting procedures, *Journal of Risk* 9, 1–17.

Christoffersen, P., 1998, Evaluating interval forecasts, *International Economic Review* 39.

Christoffersen, P. F., and D. Pelletier, 2004, Backtesting Value-at-Risk: A duration-based approach, *Journal of Econometrics* 2, 84–108.

Colletaz, Gilbert, Christophe Hurlin, and Christophe Pérignon, 2013, The risk map: A new tool for validating risk models, *Journal of Banking and Finance* 37, 3843–3854.

Costanzino, N., and M. Curran, 2015, Backtesting general spectral risk measures with application to expected shortfall, *The Journal of Risk Model Validation* 9, 21–31.

Crnkovic, C., and J. Drachman, 1996, Quality control, *Risk* 9, 139–143.

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

Diebold, F.X., and R.S. Mariano, 1995, Comparing predictive accuracy, *Journal of Business & Economic Statistics* 13, 253–265.

Du, Z., and J.C. Escanciano, 2017, Backtesting expected shortfall: accounting for tail risk, *Management Science* 63, 940–958.

Dumitrescu, E., C. Hurlin, and V. Pham, 2012, Backtesting Value-at-Risk: From dynamic quantile to dynamic binary tests, *Finance* 33, 79–112.

Durlauf, S., 1991, Spectral based testing of the martingale hypothesis, *Journal of Econometrics* 50, 355–376.

Engle, R.F., and S. Manganelli, 2004, CAViaR: conditional autoregressive value at risk by regression quantiles, *Journal of Business & Economic Statistics* 22, 367–381.

Federal Register, 2012, Risk-based capital guidelines: Market risk.

Fissler, T., and J. Ziegel, 2015, Higher order elicitability and Osband's principle, Working paper.

Fissler, T., J.F. Ziegel, and T. Gneiting, 2016, Expected shortfall is jointly elicitable with value-at-risk: implications for backtesting, *Risk* 58–61.

Giacomini, R., and H. White, 2006, Tests of conditional predictive ability, *Econometrica* 74, 1545–1578.

Gneiting, T., 2011, Making and evaluating point forecasts, *Journal of the American Statistical Association* 106, 746–762.

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

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

Hull, J. C., and A. White, 1998, Incorporating volatility updating into the historical simulation method for Value-at-Risk, *Journal of Risk* 1, 5–19.

Hurlin, C., and S. Topkavi, 2007, Backtesting value-at-risk accuracy: a simple new test, *Journal of Risk* 9, 19–37.

Kauppi, H., and P. Saikkonen, 2008, Predicting U.S. recessions with dynamic binary response models, *The Review of Economics and Statistics* 90, 777–791.

Kerkhof, J., and B. Melenberg, 2004, Backtesting for risk-based regulatory capital, *Journal of Banking and Finance* 28, 1845–1865.

Kratz, M., Y.H. Lok, and A.J. McNeil, 2016, Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall, to appear in the Journal of Banking and Finance.

Kupiec, P. H., 1995, Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* 3, 73–84.

Leccadito, Arturo, Simona Boffelli, and Giovanni Urga, 2014, Evaluating the accuracy of Value-at-Risk forecasts: New multilevel tests, *International Journal of Forecasting* 30, 206–216.

McNeil, A. J., R. Frey, and P. Embrechts, 2015, *Quantitative Risk Management: Concepts, Techniques and Tools*, second edition (Princeton University Press, Princeton).

Milgram, Michael S., 2010, On hypergeometric 3F2(1) - a review, Working Paper 1011.4546, arXiv.

Nass, C.A.G., 1959, A χ^2-test for small expectations in contingency tables, with special reference to accidents and absenteeism, *Biometrika* 46, 365–385.

O'Brien, J., and P.J. Szerszen, 2017, An evaluation of bank measures for market risk before, during and after the financial crisis, *Journal of Banking and Finance* 80, 215–234.

Pérignon, C., Z.Y. Deng, and Z.J. Wang, 2008, Diversification and Value-at-Risk, *Journal of Banking and Finance* 32, 783–794.

Pérignon, C., and D. R. Smith, 2010, The level and quality of Value-at-Risk disclosure by commercial banks, *Journal of Banking and Finance* 34, 362–377.

Pérignon, C., and D.R. Smith, 2008, A new approach to comparing VaR estimation methods, *Journal of Derivatives* 16, 54–66.

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

Wied, D., G.N.F. Weiß, and D. Ziggel, 2016, Evaluating Value-at-Risk forecasts: a new set of multivariate backtests, *Journal of Banking and Finance* 72, 121–132.

Ziggel, D., T. Berens, G.N.F. Weiss, and D. Wied, 2014, A new set of improved Value-at-Risk backtests, *Journal of Banking and Finance* 48, 29–41.

Zumbach, G., 2006, Backtesting risk methodologies from one day to one year, *Journal of Risk* 9, 55–91.