Article:
McNeil, Alexander John orcid.org/0000-0002-6137-2890 and Gordy, Michael (2020) Spectral backtests of forecast distributions with application to risk management. Journal of Banking and Finance. 105817. ISSN: 1872-6372

https://doi.org/10.1016/j.jbankfin.2020.105817

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

March 30, 2020

Abstract

We study a class of backtests for forecast distributions in which the test statistic depends on a spectral transformation that weights exceedance events by a function of the modeled probability level. The weighting scheme is specified by a kernel measure which 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 further propose novel variants which are easily implemented, well-sized and have good power. 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 the editor, Carol Alexander, and two anonymous referees for their thoughtful suggestions. We thank Harrison Katz and Sathya Ramesh 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 Michael Gordy, Federal Reserve Board, Washington DC 20551, USA, +1-202-452-3705, michael.gordy@frb.gov.

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 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 weighting scheme is specified by a kernel measure which makes explicit the priorities for model performance. The backtest statistic and its asymptotic distribution are analytically tractable for a very large class of kernels.

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 is the case where the kernel is Dirac measure concentrated at the target VaR level. At the other extreme, the tests applied in Diebold et al. (1998) and in the related literature on conditional density estimation including Bai (2003), Hong and Li (2005) and Corradi and Swanson (2006a,b) represent a special case in which weights are uniform across all probability levels. The likelihood-ratio test of Berkowitz (2001), the expected shortfall tests of Du and Escanciano (2017) and spectral risk measure test of Costanzino and Curran (2015) represent intermediate cases of a kernel truncated to tail probabilities.

While these works are related to our own, we make a distinct threefold contribution: (i) we offer an overarching testing framework that embeds many existing tests and many new ones, including discrete spectral tests and multivariate spectral tests; (ii) we emphasize the

idea that choice of backtest should be guided by a user's preferences for model performance as expressed in kernel choice, rather than by the blind pursuit of power; and (iii) we propose a general form of conditional test which may be combined with any kernel and which nests the unconditional test as a special case.

Our application of a weighting function bears some similarity to the approach of Amisano and Giacomini (2007) and Diks et al. (2011) who apply weights to forecast scoring rules to obtain measures of forecast performance that accentuate the tails (or other regions) of the distribution. There is a particularly close connection to Gneiting and Ranjan (2011) who apply weights to the probability level of the quantile function of the forecast model to develop their quantile-weighted continuous ranked probability score (CRPS). The weighted testing approach is applied by these authors to the comparison of forecasting methodologies using tests in the style of Diebold and Mariano (1995). In contrast, we develop absolute tests for the weighted performance of a single forecast model. This is closer in spirit to the approach of Crnkovic and Drachman (1996) who apply a statistic based on a weighted Kuiper distance between the uniform distribution and the distribution of the estimated probabilities of realized sample values under the forecast model.

While the comparative testing approach is useful for the internal refinement of the forecasting method by the forecaster, the absolute testing approach in this paper facilitates external evaluation of the forecaster's results by another agent, such as a regulator. In this paper we adopt the perspective of such an agent who must make a judgment based on a predefined set of data supplied by the forecaster and who has very limited information about the forecaster's methodology. The kernel function must be chosen exogenously in accordance with the agent's own priorities for model performance.

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 profit-and-loss (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, then 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).

Under the reforms mandated by the Fundamental Review of the Trading Book (Basel Committee on Bank Supervision, 2013), 99%-VaR is replaced by 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 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 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.

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 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, notwithstanding the fact that (L_t) is typically non-stationary. 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 $t \in \mathbb{N}$. The regulator is assumed to have no direct knowledge of the sequence of models $\widehat{F}_t$, but can conduct tests and draw inferences based on a sample of the PIT-values. 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; tests of this property are tests that the sequence of models is *calibrated in probability*.

For any α in the unit interval, let $\widehat{\mathrm{VaR}}_{\alpha,t} := \widehat{F}_t^{\leftarrow}(\alpha)$ be an estimate of the α-VaR con-

structed at time $t-1$ by calculating the generalized inverse of $\widehat{F}_t$ at α. Since the VaR exceedance event $\{L_t \geqslant \widehat{\mathrm{VaR}}_{\alpha,t}\}$ is equal to the event $\{P_t \geqslant \alpha\}$, the PIT-value provides a sufficient statistic for the VaR exceedances at *all* possible levels.

Our tests, in common with the majority of tests based on PIT values (including VaR exception tests), make no assumptions about the procedures and models used by the bank in forecasting. This is desirable for preserving the objectivity and statistical integrity of the test regime, since it prevents the regulator exploiting knowledge of the design of the forecaster's risk model to bias the outcome of a test. 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. As HS fails to account for serial dependencies in returns due to time-varying volatility, 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 squared returns. Banks that do not use HS or FHS typically adopt a parametric model for the joint distribution of risk-factor changes.

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. Conversative forecasts have been documented as well for Canadian banks (Pérignon et al., 2008) and in a larger international sample (Pérignon and Smith, 2010). The sensitivity of such results to sample period is revealed by O'Brien and Szerszen (2017). In their sample of five large US banks from 2001–2014, tests of unconditional coverage reject VaR forecasts as excessively conservative for all banks in the periods of relative stability (2001–2006 and 2010–2014). In the crisis period of 2007–2009, however, O'Brien and Szerszen reject VaR forecasts as insufficiently conservative for all five banks, and serial 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. The transformations take the form

$$W_t = \int_{[0,1]} \mathbb{1}_{\{P_t \geqslant u\}} \mathrm{d}\nu(u) \tag{1}$$

where the *kernel measure* ν is a probability measure defined on $[0, 1]$ with distribution function G_ν. The kernel measure is designed to apply weight to the probability levels of greatest interest, typically (in practice) in the region of the standard VaR level $\alpha = 0.99$. Note that (1) implies $W_t = G_\nu(P_t)$ showing that W_t is increasing in P_t and that all moments of W_t are bounded in $[0, 1]$.

All theoretical results apply under the following condition on G_ν, which admits kernel measures corresponding to discrete distributions with finite sample space and continuous or mixed probability distributions.

Assumption 1. G_ν has at most a finite set of discontinuities, which may not be at 0 or 1, and is otherwise absolutely continuous.

In the discrete case, the measure places positive mass $\gamma_1, \ldots, \gamma_m$ satisfying $\sum_{i=1}^m \gamma_i = 1$ at the ordered values $0 < \alpha_1 < \cdots < \alpha_m < 1$ leading to

$$G_\nu(u) = \sum_{i=1}^{m} \gamma_i \mathbb{1}_{\{u \geqslant \alpha_i\}}. \tag{2}$$

For the continuous case, G_ν is defined in terms of a nonnegative density $g_\nu(u)$ on $[0, 1]$ which we refer to as the *kernel density*.

Figure 1 illustrates the transformation for a selection of very simple kernels. Under a discrete kernel transformation, G_ν is a step function. An example labeled *3-point* places mass of $\gamma_1 = 0.25$, $\gamma_2 = 0.5$, and $\gamma_3 = 0.25$ at PIT values $\alpha_1 = 0.985$, $\alpha_2 = 0.99$, and $\alpha_3 = 0.995$. This generalizes the single-point Heaviside function traditionally used in VaR backtesting. The three continuous examples place mass in the same interval $[0.985, 0.995]$. Within this window, the concave transformation G_ν associated with the *linear decreasing* density places most mass in the lower portion, whereas the convex transformation associated with the *linear increasing* density places most weight in the upper portion. The *uniform* density yields a linear transformation G_ν. Since $g_\nu(u) = 0$ outside of the interval, for all three kernels we have $G_\nu(u) = 0$ for $u < 0.985$ and $G_\nu(u) = 1$ for $u > 0.995$. For these kernel choices the resulting transformation $u \mapsto G_\nu(u)$ of the interval $[0, 1]$ is clearly not a bijection. The idea is that the tester sacrifices some information in the PIT values in order to prioritize the quantile levels at which the forecast model should perform.

The univariate transformation extends naturally to the multivariate case in which a set of distinct kernel measures with distribution functions $G_1, \ldots, G_m$ is applied to PIT-values to obtain the vector-valued variables $\boldsymbol{W}_1 \ldots, \boldsymbol{W}_n$ where

$$\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,m})', \quad W_{t,j} = G_j(P_t),\ j = 1, \ldots, m. \tag{3}$$

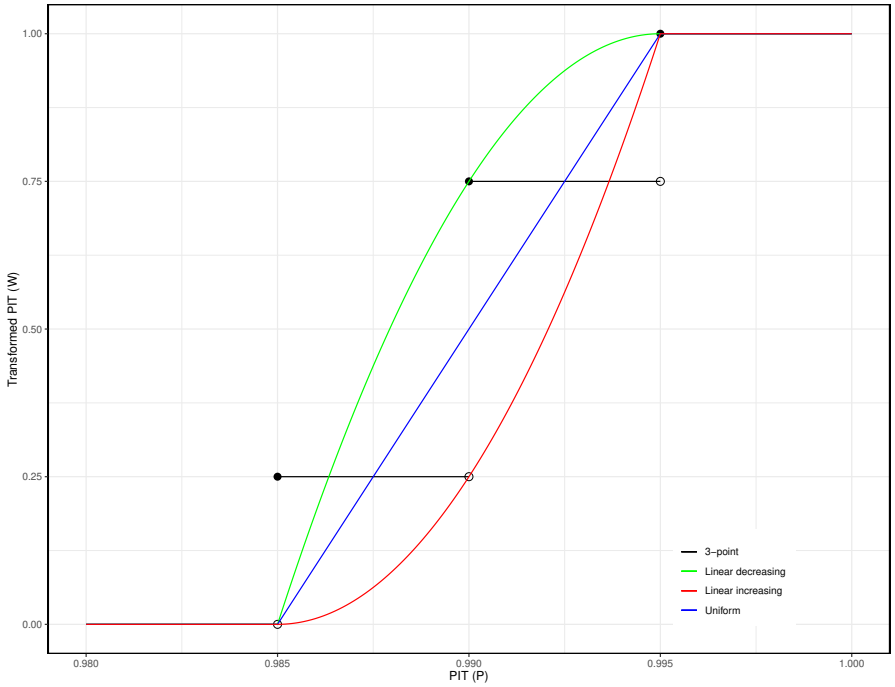

Figure 1: Examples of kernel transformation of PIT values.
We plot W_t as a function of P_t for various kernels. The *3-point* kernel places mass of $(0.25, 0.5, 0.25)$ at PIT values $(0.985, 0.99, 0.995)$. The remaining kernels, all continuous, share a window $[\alpha_1 = 0.985, \alpha_2 = 0.995]$. Within this window, the *linear decreasing* kernel has density $g_\nu(u) = (\alpha_2 - u)/(\alpha_2 - \alpha_1)$, the *uniform* kernel has density $g_\nu(u) = 1/(\alpha_2 - \alpha_1)$ and the *linear increasing* kernel has density $g_\nu(u) = (u - \alpha_1)/(\alpha_2 - \alpha_1)$. Outside the kernel window, the density is zero.

We will refer to any backtest based on spectrally transformed PIT exceedances as a *spectral backtest*. For the purposes of this paper, we assume that the regulator can utilize only present and past values of P_t in the backtest statistic. This restriction could be relaxed considerably.[1] What is essential to our contribution is that the regulator does not observe the entire distribution $\hat{F}_t$, but does observe more than the VaR exception indicator $\mathbb{1}_{\{L_t \geqslant \widehat{\mathrm{VaR}}_{\alpha,t}\}}$.

Let $(\mathcal{F}_t^*)$ be the regulator's filtration generated by the PIT values, i.e., $\mathcal{F}_t^* = \sigma(\{P_s : s \leqslant t\}) \subset \mathcal{F}_t$. Regardless of the form of the test, the null hypothesis is

$$H_0 : \quad \boldsymbol{W}_t \sim F_W^0 \text{ and } \boldsymbol{W}_t \perp\!\!\!\perp \mathcal{F}_{t-1}^*, \ \forall t, \tag{4}$$

[1]Our approach could easily be generalized to incorporate information in $(L_t, \widehat{\mathrm{VaR}}_{\alpha,t})$ and in publicly observed market variables (such as VIX). However, frequent change in portfolio composition implies that lagged VaR values are less reliably informative than lagged PIT values.

where F_W^0 denotes the distribution function of $\boldsymbol{W}_t$ when P_t is uniform. The null hypothesis (4) implies that the $(\boldsymbol{W}_t)$ are iid but is weaker than a null hypothesis that the (P_t) are iid Uniform. This is by intent. Since the regulator is free to choose the kernel measure in accordance with her priorities, she should not object to departures from uniformity and serial independence that arise outside the support of her chosen kernel.

Several recent papers propose to correct tests of forecasts for estimation error; see, e.g., Escanciano and Olmo (2010); Du and Escanciano (2017); Hurlin et al. (2017). Implementation of these corrections generally requires knowledge of the forecasting model and estimation scheme and is thus infeasible in the regulatory context we describe. Our null hypothesis imposes the high standard that the forecaster is an ideal forecaster working with a sequence of correctly specified, perfectly estimated models.

The spectral class encompasses a great variety of tests but we prioritize two general testing approaches: Z-tests and likelihood ratio (LR) tests. In the univariate case, the spectral Z-test is based on the asymptotic normality of $\overline{W}_n = n^{-1}\sum_{t=1}^n W_t$ under the null hypothesis (4). Writing $\mu_W = \mathbb{E}(W_t)$ and $\sigma_W^2 = \mathrm{var}(W_t)$ for the moments in the null model F_W^0, it follows from the central limit theorem that

$$Z_n = \frac{\sqrt{n}(\overline{W}_n - \mu_W)}{\sigma_W} \xrightarrow[n\to\infty]{d} N(0,1). \tag{5}$$

Since the transformed PIT W_t is bounded in the unit interval, the variance σ_W is guaranteed to be finite as required by the Central Limit Theorem (CLT).

The Z-test (5) is a simple test based on the mean of the spectrally transformed PIT values W_t. In the case where W_t is Bernoulli, the sample mean $\overline{W}_n$ is a sufficient statistic for the parameter of the Bernoulli distribution, but it is not sufficient for the parameters of W_t in general. A multispectral test can help address this shortcoming. In the multivariate case $(\dim \boldsymbol{W}_t = m)$ we have $\sqrt{n}\left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right) \xrightarrow[n\to\infty]{d} N_m(\mathbf{0}, \Sigma_W)$ 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

$$T_n = n\left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right)' \Sigma_W^{-1} \left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right) \sim \chi^2_m, \tag{6}$$

where we refer to T_n as an m-spectral Z-test statistic.

The first moment of the transformed PIT-values under the null hypothesis is easily obtained as

$$\mu_W = \int_0^1 (1-u)\mathrm{d}G_\nu(u) \tag{7}$$

The variance σ^2_W and the cross-moments in Σ_W are obtained using a simple product rule for spectrally transformed PIT values.

Theorem 3.1. *The set of spectrally transformed PIT values defined by $W_{t,j} = G_j(P_t)$ is closed under multiplication. The product $W^*_t = W_{t,1}W_{t,2}$ is given by $W^*_t = G^*(P_t)$ where G^* is a distribution function satisfying*

$$G^*(u) = \int_0^u \frac{1}{2}\Big(G_2(s) + G_2(s^-)\Big)\mathrm{d}G_1(s) + \int_0^u \frac{1}{2}\Big(G_1(s) + G_1(s^-)\Big)\mathrm{d}G_2(s).$$

It follows that $\sigma^2_W = \mu_{W^*} - \mu^2_W$, where μ_{W^*} is found by applying (7) using the distribution function G^* obtained when $G_1 = G_2 = G_\nu$. This yields

$$\mu_{W^*} = \int_0^1 (1-u)\left(G_\nu(u) + G_\nu(u^-)\right)\mathrm{d}G_\nu(u)\,. \tag{8}$$

Likelihood ratio tests are based on continuous parametric models $F_P(\cdot \mid \boldsymbol{\theta})$ for the PIT values P_t that nest uniformity as a special case corresponding to $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The implied model $F_W(\cdot \mid \boldsymbol{\theta})$ for the values $W_t = G_\nu(P_t)$ is used to test the null hypothesis (4) with $F^0_W = F_W(\cdot \mid \boldsymbol{\theta}_0)$; the alternative is that $\boldsymbol{W}_t \sim F_W(\cdot \mid \boldsymbol{\theta})$ with $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Writing $\mathcal{L}_W(\boldsymbol{\theta} \mid \boldsymbol{W})$ for the likelihood function, the test is based on the asymptotic chi-squared distribution of

the statistic

$$\mathrm{LR}_{W,n} = \frac{\mathcal{L}_W(\boldsymbol{\theta}_0 \mid \boldsymbol{W})}{\mathcal{L}_W(\hat{\boldsymbol{\theta}} \mid \boldsymbol{W})} \tag{9}$$

where $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate based on the transformed sample $(\boldsymbol{W}_t)$.

An important difference between the two classes of test is that the Z-test is sensitive to the kernel's functional form whereas the LR-test is sensitive only to the kernel's support. Considering the univariate case for simplicity, we show

Theorem 3.2. *Let G_1 and G_2 be distribution functions satisfying Assumption 1 and corresponding to probability measures ν_1 and ν_2. Let $W_{t,j} = G_j(P_t)$ for $j = 1, 2$ and $t = 1, \ldots, n$ be the respective samples of transformed PIT values. If $\mathrm{supp}(\nu_1) = \mathrm{supp}(\nu_2)$ then $\mathrm{LR}_{W_1,n} = \mathrm{LR}_{W_2,n}$ almost surely.*

This result can be viewed as being analogous to the invariance property of LR-tests under one-to-one transformations of the data.

4 Tests of unconditional coverage

It is common to divide backtesting methods into tests of unconditional coverage and tests of conditional coverage. In our setting, an unconditional test is a test for the distribution F_W^0 implied by the uniformity of the PIT-values while a conditional test is a test for both the correct distribution and the independence of $\boldsymbol{W}_t$ 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 and we show how our approach subsumes a number of important published tests or close relatives thereof. It is important to note that the convergence results on which our tests are based, although mostly stated under iid assumptions, do hold in situations where the independence assumption is relaxed, for example for stationary and ergodic martingale-difference processes (according to the martingale CLT of Billingsley, 1961). In the case of the univariate Z-test, the test will have no power to detect serial

dependence whenever $\lim_{n\to\infty} \text{var}(\sqrt{n}\overline{W}_n) \approx \sigma_W^2$. If, however, there is persistent positive serial correlation in (W_t) leading to $\lim_{n\to\infty} \text{var}(\sqrt{n}\overline{W}_n) > \sigma_W^2$ then the Z-test will have some power to detect dependencies; however, more targeted tests of the independence property are available and are the subject of Section 5.

4.1 Discrete weighting

Discrete tests are based on the univariate transformation $W_t = \sum_{i=1}^m \gamma_i \mathbb{1}_{\{P_t \geqslant \alpha_i\}}$ as defined in (2) and the multivariate transformation $\boldsymbol{W}_t = (\mathbb{1}_{\{P_t \geqslant \alpha_1\}}, \ldots, \mathbb{1}_{\{P_t \geqslant \alpha_m\}})'$ in (3) for the same set of ordered levels $\alpha_1 < \cdots < \alpha_m$. Obviously, when $m = 1$ (and $\gamma_1 = 1$) both transformations yield $W_t = \mathbb{1}_{\{P_t \geqslant \alpha\}}$, so that we obtain iid Bernoulli$(1-\alpha)$ variables under the null hypothesis (4). This is the basis for standard VaR exceedance testing based on the binomial distribution. The Z-test statistic (5) for $W_t = \mathbb{1}_{\{P_t \geqslant \alpha\}}$ coincides with the binomial score test statistic

$$Z_n = \frac{\sqrt{n}\left(\overline{W}_n - (1-\alpha)\right)}{\sqrt{\alpha(1-\alpha)}}. \tag{10}$$

The LR-test uses an implicit nesting model for P_t in which the W_t are iid Bernoulli(p) and tests $p = 1-\alpha$ against $p \neq 1-\alpha$ by comparing the statistic (9) to a χ_1^2 distribution; this is the approach taken by Kupiec (1995) and Christoffersen (1998).

When $m > 1$ the variables $W_t = \sum_{i=1}^m \gamma_i \mathbb{1}_{\{P_t \geqslant \alpha_i\}}$ take the ordered values $0 = \Gamma_0 < \Gamma_1 < \cdots < \Gamma_m = 1$, where $\Gamma_k = \sum_{i=1}^k \gamma_i$ for $k = 1, \ldots, m$. Under the null hypothesis (4) the distributions of W_t and $\boldsymbol{W}_t$ satisfy

$$\mathbb{P}(W_t = \Gamma_i) = \mathbb{P}(\mathbf{1}'\boldsymbol{W}_t = i) = \alpha_{i+1} - \alpha_i, \quad i \in \{0, 1, \ldots, m\}, \tag{11}$$

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

The univariate and multivariate tranformations result in different Z-tests which can be considered as alternative generalizations of the binomial score test (10). Application of Theorem 3.1 to the univariate case and use of (8) delivers moments under the null $\mu_W =$

$\sum_{i=1}^m \gamma_i(1-\alpha_i)$ and $\sigma_W^2 = \sum_{i=1}^m \gamma_i^*(1-\alpha_i) - \mu_W^2$ where $\gamma_i^* = (2\Gamma_i - \gamma_i)\gamma_i$. In constructing the test statistic Z_n in (5), we can vary the weights γ_i to emphasize different levels α_i and obtain a variety of new tests.

In the multivariate case, we construct an m-spectral Z-test as in (6) with $\boldsymbol{\mu}_W = (1-\alpha_1, \ldots, 1-\alpha_m)'$ and second moment matrix Σ_W with (i,j) element given by $\alpha_{i\wedge j}(1-\alpha_{i\vee j})$. We then 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* $i = 0, \ldots, m$.

To implement a multinomial (or multi-level) LR-test of (11) we use a nesting model for P_t in which $\mathbb{P}(W_t = \Gamma_i) = \mathbb{P}(\mathbf{1}'\boldsymbol{W}_t = i) = p_i$ and $\sum_{i=0}^m p_i = 1$. The likelihoods based on (W_t) and $(\boldsymbol{W}_t)$ yield the same sufficient statistics $O_i = \sum_{t=1}^n \mathbb{1}_{\{W_t=\Gamma_i\}} = \sum_{t=1}^n \mathbb{1}_{\{\mathbf{1}'\boldsymbol{W}_t=i\}}$ for the cell probabilities p_i. By the likelihood principle the univariate and multivariate LR-tests are identical and depend only on the levels $(\alpha_1, \ldots, \alpha_m)$ and not the weights γ_i in the univariate transformation. The invariance of the univariate LR-test under different choices for the weights is also a consequence of Theorem 3.2. The multinomial LR-test coincides with the test proposed in Pérignon and Smith (2008) which also underlies the work of Colletaz et al. (2013); see Kratz et al. (2018) for a comparison with Pearson test.

4.2 Continuous weighting

The continuous tests we consider are based on distribution functions G_ν with densities g_ν satisfying $g_\nu(u) > 0$ for $\alpha_1 < u < \alpha_2$ and $g_\nu(u) = 0$ for $u < \alpha_1$ and $u > \alpha_2$. We refer to the interval $[\alpha_1, \alpha_2]$ as the *kernel window*.

Consider spectral transformations $W_{t,1} = G_1(P_t)$ and $W_{t,2} = G_2(P_t)$ corresponding to kernel densities g_1 and g_2 with common kernel window $[\alpha_1, \alpha_2]$. Theorem 3.1 implies that the spectral transformation $W_t^* = G^*(P_t) = W_{t,1}W_{t,2}$ has kernel window $[\alpha_1, \alpha_2]$ and kernel

density given by $g^*(u) = G_1(u)g_2(u) + G_2(u)g_1(u)$. Hence moments and cross-moments of the $W_{t,i}$ can be obtained analytically for a wide variety of kernel densities, e.g., based on polynomials, exponential functions, or on beta-type densities of the form $(u-\alpha_1)^{a-1}(\alpha_2 - u)^{b-1}$ for $a, b > 0$; see Section 4.3 for examples of new tests based on this idea. Thus, our compact presentation of the continuous spectral Z-test subsumes a very large class of possible tests. In particular, the tests proposed by Du and Escanciano (2017) and Costanzino and Curran (2015) are special cases of our univariate spectral Z-test.

For the LR-test, recall that we require a family of distributions $F_P(\cdot \mid \boldsymbol{\theta})$ for the PIT values that nests uniformity as a special case corresponding to $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Since the kernels of this section have support $[\alpha_1, \alpha_2]$, Theorem 3.2 implies that they all give rise to identical LR-tests, depending only on α_1, α_2 and the nesting model $F_P(\cdot \mid \boldsymbol{\theta})$. The form taken by G_ν on $[\alpha_1, \alpha_2]$ is immaterial.

Following Berkowitz (2001), we draw upon the probitnormal as nesting model. We assume that the PIT values $P_1, \ldots, P_n$ have a 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

$$F_P(p \mid \boldsymbol{\theta}) = \Phi\left(\frac{\Phi^{-1}(p) - \mu}{\sigma}\right), \quad f_P(p \mid \boldsymbol{\theta}) = \frac{\phi\left(\frac{\Phi^{-1}(p)-\mu}{\sigma}\right)}{\phi(\Phi^{-1}(p))\sigma}, \quad p \in [0,1], \tag{12}$$

and the uniform distribution corresponds to $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (0,1)'$.

The Berkowitz test is an LR-test that the data $P_t^* = \max(\alpha_1, P_t)$ have a uniform distribution truncated to $[\alpha_1, 1]$ against the alternative that they have a probitnormal distribution truncated to the same interval. Having restricted ν to the set of probability measures defined on $[0,1]$, there is no ν such that $P_t^* = G_\nu(P_t)$. However, it is easily seen that taking ν as the uniform kernel on $[\alpha_1, 1]$ yields the spectral transformation

$$W_t = G_\nu(P_t) = \max\left(0, \frac{P_t - \alpha_1}{1-\alpha_1}\right) = \frac{P_t^* - \alpha_1}{1-\alpha_1}.$$

Since W_t and P_t^* are related by a simple one-to-one transformation we obtain the same LR-test as Berkowitz by arguments similar to Theorem 3.2. The equivalence holds as well when we generalize the tests to allow the kernel window to have upper bound $\alpha_2 \in (\alpha_1, 1]$.

4.3 Size and power

We have performed extensive Monte Carlo analyses to explore how the size and power of unconditional spectral backtests depend on the kernel and test. Here we offer representative examples. Details of all simulation experiments may be found in the Online Supplement to this paper.

We consider kernels of discrete, continuous and mixed form. 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 2-level kernels at points (α_1, α_2) and 3-level kernels at 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 is a wide variety of plausible candidates for the kernel. Table 1 lists the kernels that we discuss below; each may be thought of as describing a family of kernel densities for different windows $[\alpha_1, \alpha_2]$. For parsimony, all are special cases of the beta kernel. The uniform and hump-shaped Epanechnikov kernels are commonly used in the nonparametric statistics literature. The kernels are all quite natural for weighting quantile probabilities and are similar to choices made by Gneiting and Ranjan (2011, Table 2). In the Online Supplement, we provide analytical solutions for the moments of transformed PIT values for the general beta(a, b) case.

Discrete monospectral Z-tests: two-sided binomial score test at level α^* (BIN); and three-point discrete uniform kernel (ZU3) test;

Kernel family	Mnemonic	Density $g(\tilde{u})$	Beta representation
Uniform	ZU	1	1,1
Arcsin	ZA	$1/\sqrt{\tilde{u}(1-\tilde{u})}$	½,½
Epanechnikov	ZE	$\tilde{u}(1-\tilde{u})$	2,2
Linear increasing	ZL_+	$\tilde{u}$	2,1
Linear decreasing	ZL_-	$1-\tilde{u}$	1,2

Table 1: Kernel density functions on $[\alpha_1, \alpha_2]$.
$\tilde{u}$ denotes the rescaled value $\tilde{u} = (u - \alpha_1)/(\alpha_2 - \alpha_1)$. Density functions are not scaled to integrate to 1.

Continuous monospectral Z-tests: tests based on the uniform kernel (ZU); the arcsin kernel (ZA); Epanechnikov kernel (ZE); and increasing (ZL_+) and decreasing (ZL_-) linear kernels;

Discrete multispectral Z-tests: two-point and three-point Pearson tests (PE2 and PE3);

Continuous multispectral Z-tests: forming symmetric pairs of beta kernels with parameters $(p, 1)$ and $(1, p)$, for $p = 1$ we have the bi-linear kernel (ZLL), and for $p = 25$ the sharply concave/convex bi-power kernel (ZPP); finally, we insert a uniform kernel into the bi-power kernel to create a tri-power kernel (ZPUP).

We consider three different choices for the cdf F of the *true model* of L_t: the standard normal, the scaled t_5 and scaled t_3. The Student t distributions are scaled to have variance one so differences stem from different tail shapes rather than different variances. We take the *risk manager's model* $\widehat{F}$ to be the standard normal, i.e., we transform the sampled L_t to PIT-values as $P_t = \Phi(L_t)$. Therefore, when the samples of L_t are drawn from the standard normal, the PIT-values are uniformly distributed and are used to evaluate the size of the tests. The PIT samples arising from the Student t distributions show the kind of departures from uniformity that are observed when the risk manager's model is too thin-tailed.

We fix a sample size $n = 750$ corresponding approximately to the three-year samples of bank data studied in Section 6.[2] In Table 2, we report the percentage of rejections of the

[2]All findings in this section hold qualitatively for $n = 250$ and $n = 500$; see our Online Supplement for details.

null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. All reported p-values are based on two-sided tests, though one-sided versions of some tests are of course available.

		Monospectral							Bispectral			Trispectral	
window	F | kernel	BIN	ZU3	ZU	ZA	ZE	ZL_+	ZL_-	PE2	ZLL	ZPP	PE3	ZPUP
narrow	Normal	6.1	4.9	4.7	4.7	4.7	4.6	4.8	4.8	4.8	4.8	5.3	5.2
	Scaled t5	33.9	35.0	33.8	34.4	33.0	40.3	27.1	44.0	40.0	45.3	40.3	39.3
	Scaled t3	24.0	24.8	23.9	24.3	23.3	32.7	16.5	50.7	43.3	50.9	43.4	42.7
wide	Normal	6.1	5.0	4.9	4.9	4.9	4.9	4.9	4.8	5.0	4.9	5.1	5.0
	Scaled t5	33.9	10.7	6.4	6.6	6.1	11.9	5.8	60.7	45.1	59.2	55.5	51.8
	Scaled t3	24.0	13.5	17.7	20.4	15.4	7.4	31.9	94.0	85.8	93.0	90.6	88.4

Table 2: Estimated size and power of unconditional Z-tests.
We report the percentage of rejections of the null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. The number of days in each backtest sample is $n = 750$. The narrow window is [0.985, 0.995] and the wide window is [0.95, 0.995].

In both narrow and wide windows, we observe that the size of the Z-tests is very close to the nominal size of 5%, except in the case of the binomial score test, which is slightly oversized. The power of the tests, in contrast, is sensitive to the choice of kernel. We summarize the results as follows:

1. Differences across tests in power are more pronounced on the wide window than on the narrow window. The monospectral tests are broadly similar in power to the binomial score test on the narrow window.

2. The monospectral tests offer more power against the scaled t_5 model than the more fat-tailed scaled t_3 model on the narrow window, but the opposite is true in most cases on the wide window.

3. Increasing the window width *reduces* the power of most of the monospectral tests.

4. For the wide window, the increasing linear kernel ZL_+ offers more power than the decreasing linear kernel $ZL_-$when the true model is the scaled t_5, but the opposite holds when the true model is the scaled t_3.

5. The bispectral tests offer more power than the monospectral tests, but the differences are relatively small in the case of the narrow window when the true model is the scaled t_5.

6. Among the bispectral tests, the PE2 and ZPP, which weight heavily at or near the boundaries of the support, markedly outperform the ZLL, which weights heavily in the interior. The trispectral tests PE3 and ZPUP offer *less* power than the corresponding bispectral tests PE2 and ZPP.

The first finding is easily understood. For the families of kernel densities in Table 1, the associated function G_ν converges to a step function as the window narrows. Put another way, all kernel families degenerate to the binomial score kernel as the window shrinks around α^*. To illuminate the second finding, we plot in Figure 2 the distribution function for reported PIT-values under each of the true models, i.e., $\Pr(P_t \leqslant u) = \Pr\left(L_t \leqslant \Phi^{-1}(u)\right) = F\left(\Phi^{-1}(u)\right)$. The cdf is simply the identity line ($y = x$) when the null hypothesis is true. Within the narrow window of $[0.985, 0.995]$, the cdf for the scaled t_3 lies closer to the identity line on average than does the cdf for the scaled t_5. The monospectral tests, being tests of the first moment of $W_t = G_\nu(P_t)$, are sensitive to this distance, so have greater power against the scaled t_5 than the scaled t_3. On the wide window, however, the cdf for the scaled t_5 lies closer to the identity line on average, so the tests have greater power against the scaled t_3.

The figure also illuminutes the third and fourth findings. Both of the scaled Student t cdfs cross the identity line outside the boundaries of the narrow window, but near the middle of the wide window. Crossings within the window reduce the average distance, so pose a particular challenge for the monospectral tests. On the wide window, the scaled t_5 cdf crosses the identity line near 0.971, which is slightly below the midpoint. This slight asymmetry favors the ZL_+ kernel, which puts heavier weight on the upper side of the window. The scaled t_3 cdf crosses the identity line above the midpoint (near 0.982), and furthermore the distance between the cdf and identity line is much larger at the lower end of the window. This asymmetry favors the ZL_- kernel.

With regard to the fifth finding, the greater power of the bispectral tests is most apparent when the kernel window contains a crossing of the type just described. While the crossing reduces the average distance between the cdf of the reported PIT-values and the identity line, the cdf will be too steep or too shallow. Cross-moments of the kernels in a bispectral test can effectively detect such a *slope violation*. In contrast, when the cdf of the reported PIT-values lies roughly parallel to the identity line throughout the kernel window (as is the case for the scaled t_5 in the narrow window), the advantage of the bispectral test is expected to be less pronounced.

Finally, the power of bispectral tests to discern slope violations is greatest when the component kernels emphasize opposite ends of the kernel support. Put another way, the lower the correlation between $W_{t,1}$ and $W_{t,2}$, the greater the additional information gain in introducing the second kernel. We tend to lose power when introducing a third kernel because the average correlation across the components must rise. The marginal gain in information is too small to offset the additional degree of freedom in the χ^2 test.

In Table 3, we compare the size and power of LR-tests against Z-test counterparts. The classic Kupiec (1995) LR-test (LR1) is matched to the binomial score Z-test. Two- and three-point multinomial LR-tests (LR2 and LR3) are matched to the two- and three-point Pearson tests. We have some discretion in how we choose a bispectral Z-test counterpart to the LR-test of Berkowitz (2001) (LRB); we match it against the bi-power kernel (ZPP) as indicative of what can be achieved when the kernels have low correlation.

Overall, we find the Z-tests outperform their LR-test counterparts in size and power. The 3-point multinomial LR-test is notably oversized on both narrow and wide support, whereas the worst of the Z-tests (BIN) is only slightly oversized. When the true model F is the scaled t_5, the Z-tests in each case offer greater power than the corresponding LR-test. When the true model F is the scaled t_3, the LR-tests and Z-tests perform similarly overall. Since the Z-tests avoid specification and estimation of a nesting model, they are also much simpler to implement and faster to execute. Therefore, we will henceforth limit our attention to the

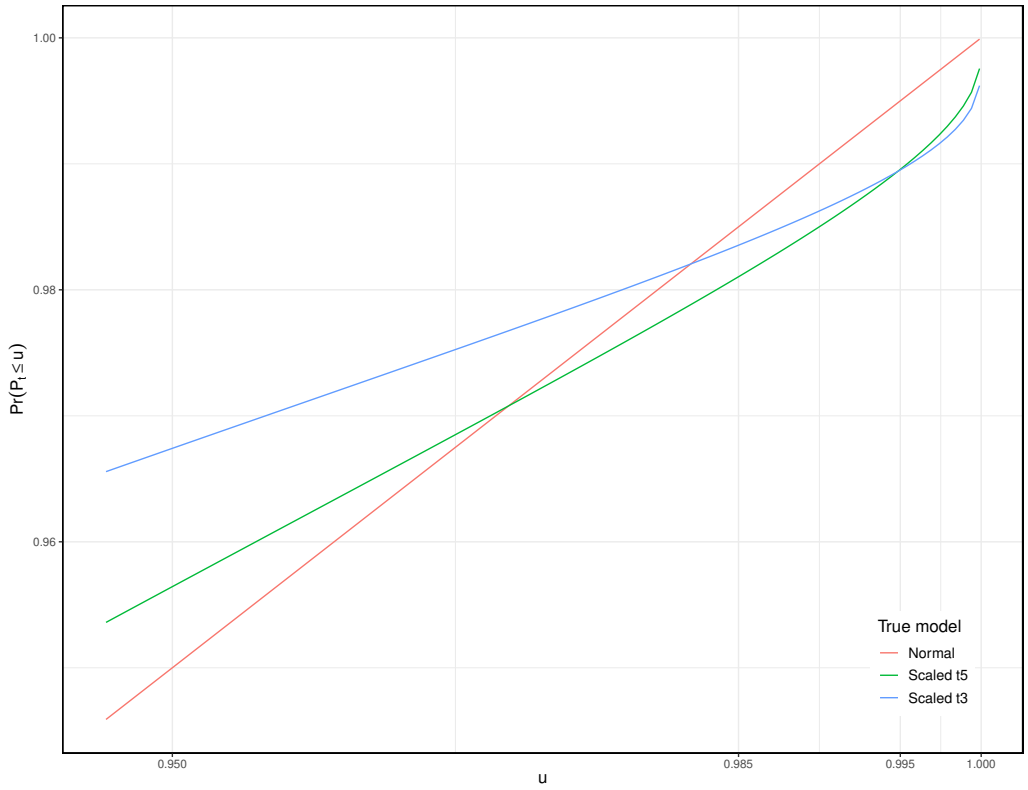

Figure 2: Distribution functions for reported PIT-values.
CDFs for the reported PIT-values when the risk manager assumes standard normal losses ($\widehat{F} = \Phi$) but the true loss model F is standard normal (red line), scaled t_5 (green line) or scaled t_3 (blue line).

class of Z-tests.

As a general caveat, we do not advocate that power alone should dictate the choice of kernel. A general intuition from our simulation studies is that a test is most powerful in rejecting a false model when the kernel weights heavily on probability levels for which the quantiles of the risk manager's model diverge from the true quantiles. 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 very extreme tail values of α runs counter to a primary regulatory motivation for the backtest, which is to verify the bank's 99% VaR.

window	F | test	BIN	LR1	PE2	LR2	PE3	LR3	ZPP	LRB
narrow	Normal	6.1	4.1	4.8	6.3	5.3	8.2	4.8	5.5
	Scaled t5	33.9	24.0	44.0	36.5	40.3	34.3	45.3	37.6
	Scaled t3	24.0	16.1	50.7	47.7	43.4	46.5	50.9	49.2
wide	Normal	6.1	4.1	4.8	5.9	5.1	7.3	4.9	5.1
	Scaled t5	33.9	24.0	60.7	57.1	55.5	53.0	59.2	57.7
	Scaled t3	24.0	16.1	94.0	94.8	90.6	93.0	93.0	95.0

Table 3: Estimated size and power of unconditional Z-tests and LR-tests. We report the percentage of rejections of the null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. The number of days in each backtest sample is $n = 750$. The narrow window is [0.985, 0.995] and the wide window is [0.95, 0.995].

5 Tests of conditional coverage

While the unconditional tests of Section 4 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 $\mathcal{F}^*_{t-1}$ as well as the correctness of the distribution of W_t. These tests should have more power to detect departures from the null hypothesis resulting from a bank's failure to use all the information in $\mathcal{F}_{t-1}$ when building the predictive model $\widehat{F}_t$, such as a failure to address time-varying volatility in adequate fashion.

5.1 Tests of the martingale difference property

A necessary condition for null hypothesis (4) to hold is the martingale difference (MD) property with respect to the regulator's filtration:

$$E(W_t - \mu_W \mid \mathcal{F}^*_{t-1}) = 0 \tag{13}$$

where we recall that $\mathcal{F}^*_t = \sigma(\{P_s : s \leqslant t\})$. When the MD property (13) holds, we must have $E(h_{t-1}(W_t - \mu_W)) = 0$ for any $\mathcal{F}^*_{t-1}$-measurable random variable h_{t-1}. Using a function h, which we refer to as a *conditioning variable transformation* (CVT), we form the $k+1$-dimensional lagged vector $\boldsymbol{h}_{t-1} = (1, h(P_{t-1}), \ldots, h(P_{t-k}))'$. 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. Particular examples that we will use in our empirical analysis are $h(p) = \mathbb{1}_{\{p \geqslant \alpha\}}$ for some α and $h(p) = |2p-1|^c$ for $c > 0$.

We base our test on the vector-valued process $\boldsymbol{Y}_t = \boldsymbol{h}_{t-1}(W_t - \mu_W)$. Under the null hypothesis, $\boldsymbol{Y}_{k+1}, \ldots, \boldsymbol{Y}_n$ should be close to the zero vector on average. We apply the conditional predictive test of Giacomini and White (2006) which was developed for comparing forecasting methods, and which has also been used by Nolde and Ziegel (2017) in the backtesting 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 $\Sigma_Y := \operatorname{cov}(\boldsymbol{Y}_t)$. Giacomini and White show that under very weak assumptions, for large enough n and fixed k,

$$T_{n,k} = (n-k)\, \overline{\boldsymbol{Y}}'_{n,k}\, \hat{\Sigma}_Y^{-1}\, \overline{\boldsymbol{Y}}_{n,k} \sim \chi^2_{k+1}. \tag{14}$$

Since Σ_Y decomposes as $\sigma_W^2 \mathbb{E}\left(\boldsymbol{h}_{t-1}\boldsymbol{h}'_{t-1}\right)$ under the null hypothesis, we form the estimator

$$\hat{\Sigma}_Y = \sigma_W^2 (n-k)^{-1} \sum_{t=k+1}^{n} \boldsymbol{h}_{t-1}\boldsymbol{h}'_{t-1}. \tag{15}$$

Alternative estimators may be used, but the decomposition in (15) has the advantage that it nests our unconditional spectral Z-test in (6), which corresponds to the case $k = 0$. The case $k = 1$ may be viewed as a Z-test analog of the first-order Markov chain test of Christoffersen (1998).

In the Online Supplement, we show that $T_{n,k}$ can also be interpreted as the chi-squared statistic for a regression of $W_t - \mu_W$ on $\boldsymbol{h}_{t-1}$. Thus, our conditional test embeds the regression-based DQ test statistic proposed by Engle and Manganelli (2004), which corresponds to the binomial score case, i.e., the case where $W_t = \mathbb{1}_{\{P_t \geqslant \alpha\}}$ and the CVT is $h(p) = \mathbb{1}_{\{p \geqslant \alpha\}}$.[3]

The extension to a conditional bispectral Z-test is based on generalizing the statistic (14) to accommodate vectors $\boldsymbol{Y}_t = \left(\boldsymbol{h}'_{t-1,1}(W_{t,1} - \mu_{W,1}), \boldsymbol{h}'_{t-1,2}(W_{t,2} - \mu_{W,2})\right)'$ formed from two different series of spectrally transformed PITs $(W_{t,i})$ and two different series of conditioning

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

variable vectors ($\boldsymbol{h}'_{t-1,i}$). Details are given in Appendix B. The approach clearly extends to conditional multispectral tests which offer more general Z-test analogs of the conditional multilevel tests in Leccadito et al. (2014).

5.2 Size and power

We build on the Monte Carlo exercises of Section 4.3 to study the size and power of the conditional tests of coverage. Here we present a representative extract of simulation studies documented in our Online Supplement. The data are generated from three different "true" models: iid standard normal; a time series model known as VT-ARMA(1,1) in which the squares of the data have the serial dependence of an ARMA(1,1) process but the data have a standard normal marginal distribution; and a VT-ARMA(1,1) model in which the marginal distribution is scaled t_5. Our calibration of the ARMA parameters (AR = 0.95, MA = -0.85), which is described in the Online Supplement, is designed to mimic the serial dependence in PIT values when stochastic volatility is neglected. The resulting process gives similar behavior to a GARCH(1,1) model but allows the marginal distribution to be freely specified. As in Section 4.3, we assume the risk manager reports PIT-values based on the standard normal model $\widehat{F} = \Phi$.

small change

In addition to a choice of kernel, the MD test requires 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. As listed in Table 4, we consider four candidates for the CVT. Whereas the DQ requires only a time-series of traditional exceedance indicators, the three CVT based on the $V(u)$ transformation require that the regulator observe PIT values.

Table 5 gives a flavor of the main findings for the example of the uniform kernel (ZU) and the narrow kernel window of $[0.985, 0.995]$. We report the percentage of rejections of the null hypothesis at the 5% confidence level based on 2^{16} = 65,536 replications. In the first column (CVT="None"), we set $k = 0$ to obtain the unconditional Z-test. Each of the

Mnemonic	$h(P)$	Description
DQ	$\mathbb{1}_{\{P \geqslant 0.99\}}$	Flags upper-tail PIT values, as in Engle and Manganelli (2004).
V.BIN	$\mathbb{1}_{\{V(P) \geqslant 0.98\}}$	Two-tailed version of DQ, flags PIT values near zero or one.
V.4	$V(P)^4$	Places heavier weight on tail PIT values in the recent past.
V.½	$\sqrt{V(P)}$	Dampens sensitivity to tail PIT values relative to V.4.

Table 4: Conditioning variable transformations. $V(u) \equiv |2u - 1|$

remaining columns corresponds to a CVT with $k = 4$. As seen in the first row (iid standard normal model), size is more difficult to control in the conditional tests. However, the CVT choices V.4 and V.½ are only slightly oversized whereas V.BIN and, in particular, DQ are very oversized.

The model depicted in the second row gives uniformly distributed PIT values with a serial dependence structure that is typical when stochastic volatility is ignored. The power of the unconditional test in this situation is very limited (10.8%), while the MD tests show power ranging from 21.7% to 32.6%. There is a further increase in power when the simulated PIT data are both non-uniform and serially dependent (third row).

F	Serial Dependence | CVT	None	DQ	V.BIN	V.4	V.½
Normal	None	4.8	14.4	9.0	6.7	6.7
Normal	VT-ARMA(1, 1)	10.8	31.5	30.9	32.6	21.7
Scaled t5	VT-ARMA(1, 1)	36.2	54.9	52.7	60.7	54.5

Table 5: Estimated size and power of conditional tests.
MD tests using the ZU kernel on the narrow window [0.985, 0.995]. We report the percentage of rejections of the null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. The number of days in each backtest sample is $n = 750$. VT-ARMA parameters are AR = 0.95, MA = -0.85.

Check small changes in this section to introduce VT-ARMA terminology, including in Table 5. Readers (and referees) tend to take ARMA(1,1) literally, if they don't read carefully. I think it helps to add the VT prefix. I also think it is reasonale to have the citation of my paper in the OS only, as it is currently still a preprint.

6 Application to bank-reported PIT values

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 (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, which can include derivative as well as cash positions. Our samples are taken from the three-year period from 2014–2016.

Summary statistics for the unconditional distributions are found in Table 6. As is often the case with new regulatory reporting requirements, the data are not uniform in quality. Two of the samples (coded $Pf104$ and $Pf110$) have missing values (0.9% and 3.2% of trading days, respectively). Furthermore, close inspection reveals that most of the samples contain a small number of observations that are potentially spurious. In a few extreme cases, a PIT value of 1 is matched to a realized loss smaller than the forecast VaR. We apply 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; details are provided in the Online Supplement. In test results reported below, we treat spurious values as missing to make the tests less sensitive to reporting error. Our conclusions are robust to taking all non-missing observations as valid.

Remaining columns of the table provide a histogram of PIT values. For some portfolios, tail PIT values are underrepresented (e.g., $Pf107$) or overrepresented (e.g., $Pf105$) in the sample. For some other portfolios, the histograms appear to be close to uniform, e.g., for $Pf110$, 85.9% of PIT values lie in $[0.05, 0.95)$ and remaining mass is distributed roughly

symmetrically.

6.2 Tests of unconditional coverage

Due to the generality of our framework, application of spectral backtests to data involves choices along several dimensions. As in Section 4.3, we fix $\alpha^* = 0.99$ as the conventional VaR level, define a *narrow* window as $[0.985, 0.995]$ and a *wide* window as $[0.95, 0.995]$. Kernels are drawn from Table 1. Guided by our simulation results and the need for brevity, we exclusively employ two-sided Z-tests in our empirical analysis.

Table 7 presents p-values for the tests of unconditional coverage.[4] We find that the forecast models for portfolios $Pf105$, $Pf106$ and $Pf107$ are rejected at the 1% level for all kernels and on both the narrow and wide kernel windows. In view of the histograms observed in Table 6, this is unsurprising. When an empirical distribution function (edf) lies above the uniform cdf within the kernel window (as observed for $Pf107$), 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 $Pf105$ and $Pf106$), large PIT values are overrepresented in the sample, which suggests that the forecast model understates the upper quantiles. By contrast, there are no rejections at all for $Pf110$, for which the edf is reasonably close to the theoretical cdf throughout the upper tail.

For the remaining six portfolios, test results are sensitive to the choice of kernel. This is to be expected and desirable, as the different tests prioritize different quantiles of the unconditional distribution. To shed light on the differences, in Figure 3 we plot the edf for three portfolios on the narrow window (upper panel) and wide window (lower panel). This plot is the empirical counterpart to Figure 2 in Section 4.3. For $Pf104$, we observe that the edf intersects with the theoretical cdf at the common upper window boundary $\alpha_2 = 0.995$,

[4] All p-values in the tables below should be interpreted in the context of a single test of the null hypothesis. If multiple tests are conducted, inferences would have to be based on a standard correction method such as that of Bonferroni; see Shaffer (1995) for a review.

ID	Trading days	of which:		Frequencies						
		Missing	Spurious	[0, .005)	[.005, .015)	[.015, .05)	[.05, .95)	[.95, .985)	[.985, .995)	[.995, 1]
101	756	0	0	0.0132	0.0172	0.0357	0.8796	0.0317	0.0106	0.0119
102	751	0	7	0.0027	0.0108	0.0215	0.9113	0.0323	0.0121	0.0094
103	750	0	8	0.0040	0.0040	0.0189	0.9474	0.0162	0.0081	0.0013
104	774	7	0	0.0000	0.0000	0.0026	0.9804	0.0104	0.0013	0.0052
105	750	0	2	0.0174	0.0214	0.0294	0.8596	0.0388	0.0187	0.0147
106	629	0	1	0.0111	0.0191	0.0510	0.8328	0.0557	0.0175	0.0127
107	750	0	1	0.0000	0.0000	0.0013	0.9960	0.0027	0.0000	0.0000
108	756	0	8	0.0000	0.0053	0.0267	0.9278	0.0174	0.0120	0.0107
109	734	0	6	0.0082	0.0151	0.0495	0.8654	0.0371	0.0206	0.0041
110	774	25	0	0.0134	0.0200	0.0507	0.8585	0.0427	0.0107	0.0040

Table 6: Sample statistics.
Missing and spurious observations excluded from the reported frequencies. Sample period is 2014-01-01 to 2016-12-31.

		Monospectral					Bispectral			Trispectral
ID	window	BIN	ZU3	ZU	ZL_+	ZL_-	PE2	ZLL	ZPP	PE3
101	narrow	0.1046	0.0340	0.0356	0.0230	0.0572	0.0263	0.0595	0.0371	0.0502
	wide	0.1046	0.1407	0.2050	0.0610	0.4243	0.0254	0.0128	0.0318	0.0593
102	narrow	0.0016	0.0200	0.0800	0.1346	0.0558	0.1968	0.1085	0.2255	0.0027
	wide	0.0016	0.0063	0.2456	0.2114	0.2898	0.0722	0.4477	0.4258	0.0121
103	narrow	0.1029	0.1158	0.0987	0.1094	0.0995	0.3207	0.2545	0.3854	0.4212
	wide	0.1029	0.0035	0.0058	0.0156	0.0038	0.0085	0.0126	0.0092	0.0229
104	narrow	0.1829	0.1691	0.1651	0.3063	0.0987	0.0533	0.0767	0.0636	0.1119
	wide	0.1829	0.0010	0.0004	0.0031	0.0002	0.0001	0.0003	0.0001	0.0002
105	narrow	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	wide	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0001	0.0000	0.0002
106	narrow	0.0005	0.0004	0.0010	0.0019	0.0007	0.0037	0.0033	0.0035	0.0055
	wide	0.0005	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001
107	narrow	0.0059	0.0018	0.0026	0.0062	0.0016	0.0033	0.0053	0.0036	0.0097
	wide	0.0059	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
108	narrow	0.0166	0.0213	0.0311	0.0191	0.0520	0.0740	0.0482	0.0414	0.0933
	wide	0.0166	0.6971	0.8422	0.5358	0.4445	0.0113	0.0034	0.0033	0.0052
109	narrow	0.5217	0.2499	0.1358	0.3171	0.0634	0.0152	0.0165	0.0147	0.0222
	wide	0.5217	0.2584	0.0563	0.0503	0.0706	0.2435	0.1471	0.3192	0.3373
110	narrow	0.8514	0.9479	0.6660	0.6321	0.7032	0.9126	0.8737	0.9647	0.8692
	wide	0.8514	0.5407	0.3017	0.4527	0.2325	0.5048	0.3407	0.5109	0.6916

Table 7: Tests of unconditional coverage.
We report p-values by portfolio, kernel window, and kernel family. Narrow kernel window is [0.985,0.995] and wide kernel window is [0.95,0.995]. Sample period is 2014-01-01 to 2016-12-31.

but lies above the theoretical cdf at lower PIT values. Over the wide window, the average distance between edf and theoretical cdf is large, so any test that assigns significant weight to PIT values near the center of the window will reject. When restricted to the narrow window, the average distance is reduced, so the tests fail to reject. The edf for portfolio $Pf103$ (not shown) is qualitatively close to that of $Pf104$, which explains the similarity in test results.

The edf for portfolio $Pf108$ lies somewhat below and roughly parallel to the theoretical cdf throughout the narrow window. Tests reject at the 5% level for some of the kernels and fail to reject for others, but the p-values all lie between 1.5% and 10%. When we consider the wide window, we find that the edf lies above the theoretical cdf in the lower half of the window and below in the upper half, which implies that the forecast model underestimates quantiles at one boundary of the kernel window and overestimates quantiles at the other boundary, i.e., a slope deviation from the uniform cdf. As shown in Section 4.3, bispectral tests generally outperform monospectral tests in this situation. We find that the tests based on the bivariate and trivariate kernels all reject at the 1% level. The edf for portfolio $Pf101$ (not shown) also displays a slope violation on the wide window, and again we find the bivariate kernels most effective.

In the case of portfolio $Pf109$, the edf displays a slope violation within the narrow window. As before, we find that the tests based on the bivariate and trivariate kernels reject at the 5% level, whereas the monospectral tests all fail to reject. On the wide window, the edf lies uniformly below the theoretical cdf, so the slope violation loses salience. The bispectral and trispectral tests now fail to reject, whereas several of the monospectral tests reject at a 10% level.

6.3 Tests of conditional coverage

In this section, we emphasize the role of the conditioning variable transformation $h(P)$ in revealing serial dependence in PIT-values. For parsimony, we consider only a subset of the kernels used in the previous section. We include the binomial score kernel (BIN)

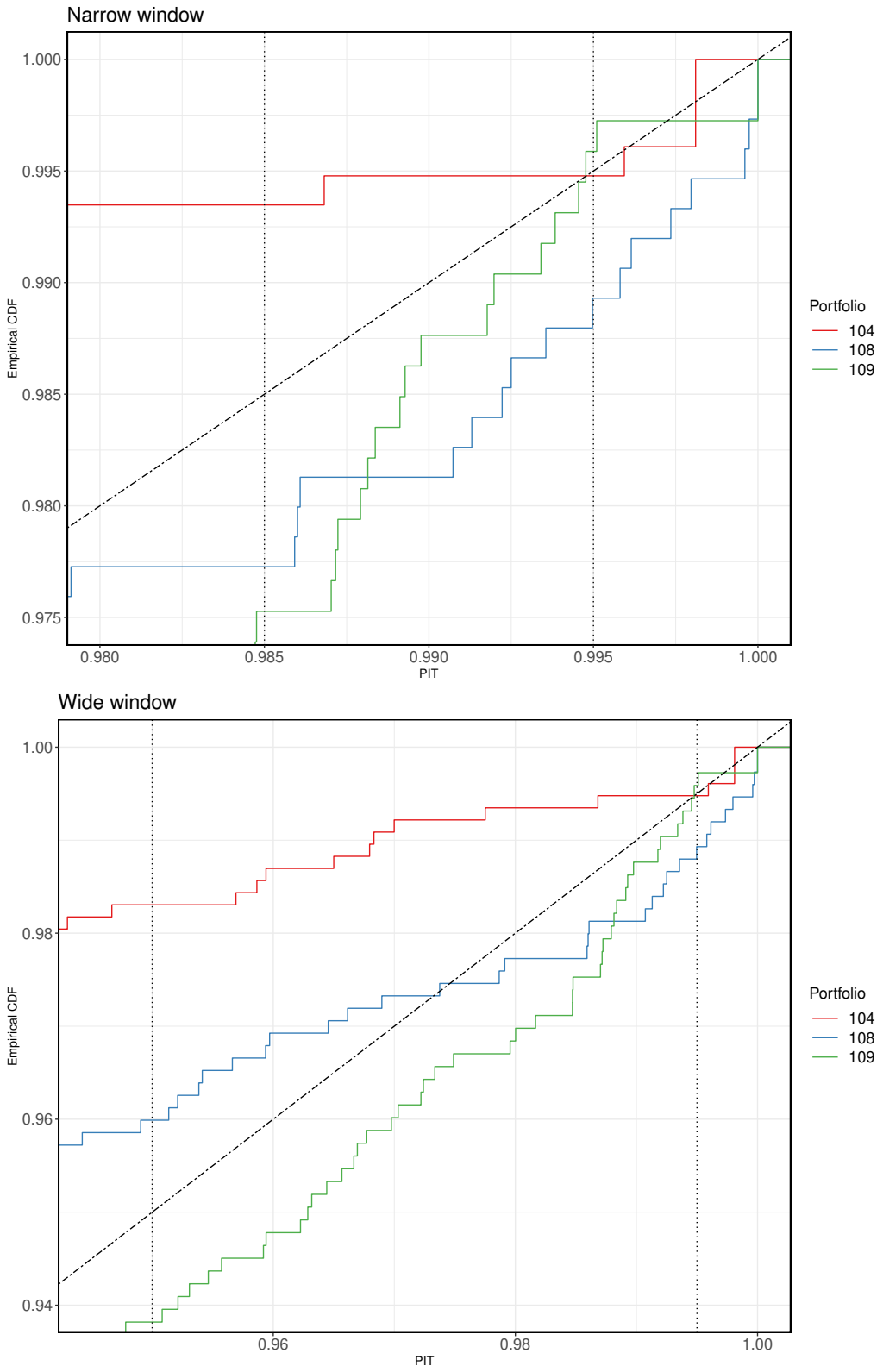

Figure 3: Empirical distribution functions for select portfolios.
EDFs for narrow kernel window (upper panel) and wide kernel window (lower panel). The uniform cdf is plotted as a dashed black line.

as representative of the traditional test, the uniform kernel (ZU) as representative of the continuous monospectral tests, and the (ZLL) as representative of the multispectral tests. We fix $k = 4$ lags in the monospectral tests which 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 ZLL test.

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 found in the Online Supplement.

Table 8 presents p-values for the tests of conditional coverage. For portfolios $Pf105$, $Pf106$ and $Pf108$, forecast models are strongly rejected (always at the 1% level, and nearly always at the 0.01% level) regardless of the choice of CVT or kernel; for brevity we drop these portfolios from the table. For only a single portfolio ($Pf109$), the forecast model is never rejected. In the other six cases, choice of CVT and kernel matters.

For portfolios $Pf102$ and $Pf110$, the V.4 CVT generally leads to rejection at the 5% level, whereas tests using the DQ CVT generally do not. The V.BIN and V.½ CVT are effective in many cases, but appear less robust than V.4. This reflects the greater sensitivity of the V.4 transformation to local spikes in market volatility. In the case of portfolio $Pf103$, the tests reject on the wide window except when using the DQ CVT.

For portfolios $Pf101$ and $Pf104$, variation in p-value across tests is driven primarily by kernel choice, and in a manner consistent with the tests of unconditional coverage in Table 7. Thus, for these two portfolios, serial dependence in the PIT-values does not appear to be the salient shortcoming in the forecast model.

In the case of $Pf107$, the test statistic is undefined for the DQ CVT and its two-tailed counterpart (V.BIN). As there were no observed violations in either tail ($P_t < .01$ or $P_t >$

			narrow window		wide window	
ID	CVT	BIN	ZU	ZLL	ZU	ZLL
101	DQ	0.0017	0.0002	0.0020	0.0342	0.0092
	V.BIN	0.0084	0.0002	0.0009	0.0513	0.0115
	V.4	0.0062	0.0003	0.0007	0.0921	0.0230
	V.½	0.0658	0.0063	0.0078	0.1376	0.0301
102	DQ	0.0088	0.0975	0.1448	0.1873	0.1443
	V.BIN	0.0119	0.4716	0.3441	0.0201	0.0147
	V.4	0.0000	0.0086	0.0023	0.0003	0.0006
	V.½	0.0000	0.0239	0.0111	0.0004	0.0005
103	DQ	0.7540	0.7429	0.8408	0.1787	0.1879
	V.BIN	0.0773	0.0426	0.1331	0.0068	0.0149
	V.4	0.3064	0.1372	0.2286	0.0074	0.0088
	V.½	0.5209	0.4194	0.5181	0.0318	0.0276
104	DQ	0.8732	0.8501	0.5202	0.0295	0.0138
	V.BIN	0.8700	0.8456	0.5167	0.0289	0.0135
	V.4	0.3225	0.2013	0.1287	0.0034	0.0022
	V.½	0.2085	0.0976	0.0689	0.0016	0.0011
107	DQ	NA	NA	NA	NA	NA
	V.BIN	NA	NA	NA	NA	NA
	V.4	0.1844	0.1071	0.1085	0.0001	0.0000
	V.½	0.1844	0.1071	0.1085	0.0001	0.0000
109	DQ	0.9917	0.6351	0.0548	0.1465	0.1383
	V.BIN	0.8041	0.7510	0.1179	0.4522	0.5170
	V.4	0.8929	0.8603	0.2067	0.5578	0.6225
	V.½	0.9313	0.6389	0.1327	0.4766	0.5266
110	DQ	0.2658	0.3058	0.4121	0.1661	0.1395
	V.BIN	0.0041	0.0006	0.0009	0.0044	0.0108
	V.4	0.0093	0.0008	0.0012	0.0100	0.0352
	V.½	0.1403	0.0513	0.0840	0.1508	0.2823

Table 8: Tests of conditional coverage.
We report test p-values by portfolio, conditioning variable transformation, kernel window and kernel family. The monospectral tests utilize $k = 4$ lags, and for the ZLL bispectral test we set $(k_1 = 4, k_2 = 0)$. Narrow kernel window is [0.985,0.995] and wide kernel window is [0.95,0.995]. Sample period is 2014-01-01 to 2016-12-31. Forecast models for $Pf105$, $Pf106$ and $Pf108$ (not tabulated) are rejected at the 1% level for all choices of CVT and kernel.

.99), in both cases the matrix $\hat{\Sigma}_Y$ in (15) is singular so cannot be inverted. This demonstrates a practical limitation of a binary-valued CVT, as short samples may often contain no tail values. Observe also that the backtest fails to reject for the remaining two CVT on the narrow window, even though the forecast model for this portfolio is strongly rejected by the unconditional tests. Since $P_t < \alpha_1 = 0.985$ for all t, W_t has a degenerate distribution in the sample. In this situation, it may be shown that the conditional test statistic is invariant to the CVT and to k and is equal to the unconditional test statistic. Recalling that the test statistic has distribution χ^2_{1+k} under the null hypothesis, we find that the p-value increases with k. This explains why unconditional backtests may have greater power than conditional backtests in situations where an overly conservative forecast model leads to degeneracy in W_t.

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 kernels facilitates the construction of new tests. From the perspective of the practice of risk management, making explicit the choice of kernel measure may help to discipline the backtesting process because the kernel directly expresses the user's priorities for model performance.

Different kernels are sensitive to different deviations from the null hypothesis. A tester who only cares about systematic under- or overestimation of quantiles within a narrow range is well served by a number of single kernels, discrete and continuous. A tester who wants to ensure maximum fidelity of the forecast models to the true distributions across a wider range of quantiles may worry more about slope violations (overestimation of quantiles at one

end of a window and underestimation at the other). Such a tester may favor a multispectral test. However, to promote a single "best" test from the spectral family would be contrary to the philosophy of our contribution, and we refrain from doing so. The tester should reflect on performance priorities and select her kernel accordingly.

Whereas most of the widely-used tests of unconditional coverage are LR-tests, our findings suggest that Z-tests should be preferred in spectral backtesting. First, we show that Z-tests perform at least as well as LR-tests in power and size over the range of backtest sample sizes used in practice. Second, our Z-tests are much easier to implement and faster to run than the LR-tests because the Z-test does not require estimation of a nesting model. Third, our Z-test framework is parsimonious, in the sense that the unconditional Z-test is nested as a special case within the broader class of conditional Z-test.

Finally, 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 enabling the formation of spectral test statistics, lagged PIT-values are especially effective as conditioning variables in regression-based tests of conditional coverage.

A Proofs

A.1 Proof of Theorem 3.1

Since G_1 and G_2 are increasing, right-continuous distribution functions it follows that $G^*(u) = G_1(u)G_2(u)$ is also a distribution function corresponding to some probability measure. The formula for G^* is obtained by applying the integration-by-parts formula for the Lebesgue-Stieltjes integral (Hewitt, 1960, Theorem A).

A.2 Proof of Theorem 3.2

Let p_t denote the realized value of P_t and $w_{t,j} = G_j(p_t)$ the corresponding realized value of $W_{t,j}$ for $t = 1, \ldots, n$ and $j = 1, 2$. There are two cases to consider. Either p_t occurs in an interval where the right derrivative of G_j is 0 or in an interval where the right derivative is positive. Let $\mathcal{G}_j$ denote the subset of $[0, 1]$ consisting of all points for which the right derivative of G_j equals zero.

If $p_t \in \mathcal{G}_j$ then, by the right-continuity of G_j, p_t must occur in an interval of the form $[a_{t,j}, b_{t,j})$ (if there is a jump in G_j at $b_{t,j}$) or $[a_{t,j}, b_{t,j}]$ (if G_j is continuous at $b_{t,j}$). In either case the contribution of $w_{t,j}$ to the likelihood is

$$\mathbb{P}\left(W_{t,j} = w_{t,j}\right) = \mathbb{P}\left(G_j(P_t) = G_j(p_t)\right) = F_P(b_{t,j}) - F_P(a_{t,j})\,.$$

If $P_t \notin \mathcal{G}_j$ then $w_{t,j}$ satisfies $\mathbb{P}(W_{t,j} \leqslant w_{t,j}) = \mathbb{P}(P_t \leqslant p_t)$ and $p_t = G_j^{-1}(w_{t,j})$, the unique inverse of G_j at $w_{t,j}$. The contribution to the likelihood is a density contribution given by

$$f_W(w_{t,j} \mid \boldsymbol{\theta}) = \frac{f_P(p_t \mid \boldsymbol{\theta})}{G_j'(p_t)}\,.$$

The general form of the realized likelihood given $\boldsymbol{w}_j = (w_{1,j}, \ldots, w_{n,j})'$ is thus

$$\mathcal{L}_{W_j}(\boldsymbol{\theta} \mid \boldsymbol{w}_j) = \prod_{p_t \in \mathcal{G}_j} \left(F_P(b_{t,j}) - F_P(a_{t,j})\right) \prod_{p_t \notin \mathcal{G}_j} \frac{f_P(p_t) \mid \boldsymbol{\theta})}{G_j'(p_t)}$$

For the measures ν_1 and ν_2 the sets $\mathcal{G}_1$ and $\mathcal{G}_2$ may differ at most by a null set. Let us assume that each realized point p_t is either in both of the sets $\mathcal{G}_1$ and $\mathcal{G}_2$ or in neither of the sets.

If $p_t \in \mathcal{G}_1$ and $p_t \in \mathcal{G}_2$ then the agreement of the supports on $(0, 1)$ implies that $a_{t,1} = a_{t,2}$ and $b_{t,1} = b_{t,2}$. Thus the likelihood contributions are identical.

If $p_t \notin \mathcal{G}_1$ and $p_t \notin \mathcal{G}_2$ then the likelihoods differ only by the scaling factor $G_j'(p_t)$ which does not involve the parameters $\boldsymbol{\theta}$. This factor will appear in the log-likelihood only as an unimportant additive term and cancel out of the LR test statistic.

It follows that the likelihoods $\mathcal{L}_{W_j}(\boldsymbol{\theta} \mid \boldsymbol{w}_j)$ are maximized by the same values $\hat{\boldsymbol{\theta}}$ and the LR-test statistics are identical.

A.3 Proof of Theorem 4.1

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 (4) $\boldsymbol{X}_t$ 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 $i \neq j$.

Now 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. Let $\overline{\boldsymbol{Y}} = n^{-1}\sum_{t=1}^n \boldsymbol{Y}_t$. A standard approach to the asymptotics of the Pearson test is to show that

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

and hence to argue that $S_m \sim \chi^2_m$ in the limit as $n \to \infty$ by the central limit theorem. It remains to show that the right-hand side of (A.1) has the spectral test representation (6).

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. It may be easily verified that $\boldsymbol{Y}_t = A\boldsymbol{W}_t$, $\boldsymbol{\theta} = A\boldsymbol{\mu}_W$ and $\Sigma_Y = A\Sigma_W A'$. It follows that

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

B Conditional bispectral Z-test

The conditional spectral Z-test generalizes to a conditional multispectral Z-test. In the bispectral case, 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

$$\boldsymbol{Y}_t = \left(\boldsymbol{h}'_{t-1,1}(W_{t,1} - \mu_{W,1}), \boldsymbol{h}'_{t-1,2}(W_{t,2} - \mu_{W,2})\right)', \tag{B.1}$$

where $\mu_{W,i}$ is the mean of $W_{t,i}$ under the null hypothesis and $\boldsymbol{h}_{t-1,i} = (1, h_i(P_{t-1}), \ldots, h_i(P_{t-k_i}))'$. Parallel to the univariate case, 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,

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

Under the null hypothesis $\Sigma_Y = A_W \circ H$, where $\circ$ denotes element-by-element multiplication (Hadamard product). The matrices are

$$H = \begin{pmatrix} \mathbb{E}\left(\boldsymbol{h}_{t-1,1}\boldsymbol{h}'_{t-1,1}\right) & \mathbb{E}\left(\boldsymbol{h}_{t-1,1}\boldsymbol{h}'_{t-1,2}\right) \\ \mathbb{E}\left(\boldsymbol{h}_{t-1,2}\boldsymbol{h}'_{t-1,1}\right) & \mathbb{E}\left(\boldsymbol{h}_{t-1,2}\boldsymbol{h}'_{t-1,2}\right) \end{pmatrix}, \quad A_W = \begin{pmatrix} \sigma^2_{W,1} J_{k_1+1,k_1+1} & \sigma_{W,12} J_{k_1+1,k_2+1} \\ \sigma_{W,12} J_{k_2+1,k_1+1} & \sigma^2_{W,2} J_{k_2+1,k_2+1} \end{pmatrix} \tag{B.3}$$

where $J_{m,n}$ denotes the $m \times n$ matrix of ones and $\sigma_{W,12} = \mathbb{E}\left((W_{t,1} - \mu_{W,1})(W_{t,2} - \mu_{W,2})\right)$. To estimate Σ_Y we generalize (15) by setting

$$\hat{\Sigma}_Y = (n - (k_1 \vee k_2))^{-1} A_w \circ \sum_{t=(k_1 \vee k_2)+1}^{n} (\boldsymbol{h}'_{t-1,1}, \boldsymbol{h}'_{t-1,2})'(\boldsymbol{h}'_{t-1,1}, \boldsymbol{h}'_{t-1,2}). \tag{B.4}$$

References

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.

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

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.

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

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.

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.

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.

Corradi, Valentina, and Norman R. Swanson, 2006a, Bootstrap conditional distribution tests in the presence of dynamic misspecification, *Journal of Econometrics* 133, 779–806.

Corradi, Valentina, and Norman R. Swanson, 2006b, Predictive density and conditional confidence interval accuracy tests, *Journal of Econometrics* 135, 187–228.

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.

Diks, Cees, Valentyn Panchenko, and Dick van Dijk, 2011, Likelihood-based scoring rules for comparing density forecasts in tails, *Journal of Econometrics* 163, 215–230.

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

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

Escanciano, J.C., and J. Olmo, 2010, Backtesting parametric value-at-risk with estimation risk, *Journal of Business & Economic Statistics* 28, 36–51.

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

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.

Hewitt, E., 1960, Integration by parts for Stieltjes integrals, *The American Mathematical Monthly* 67, 419–423.

Hong, Yongmiao, and Haitao Li, 2005, Nonparametric specification testing for continuous-time models with applications to term structure of interest rates, *Review of Financial Studies* 18, 37–84.

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., S. Laurent, R. Quaedvlieg, and S. Smeekes, 2017, Risk measure inference, *Journal of Business & Economic Statistics* 35, 499–512.

Kratz, M., Y.H. Lok, and A.J. McNeil, 2018, Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall, *Journal of Banking and Finance* 88, 393–407.

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.

Nolde, N., and J.F. Ziegel, 2017, Elicitability and backtesting: Perspectives for banking regulation, *Annals of Applied Statistics* 11, 1833–1874.

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.

Shaffer, J. P., 1995, Multiple hypothesis testing, *Annual Review of Psychology* 46, 561–584.

Online Supplement
Spectral backtests of forecast distributions with application to risk management

Michael B. Gordy and Alexander J. McNeil*

March 30, 2020

Abstract

We elaborate on certain results in our paper published in *Journal of Banking and Finance*. To avoid confusion with references to tables, figures and equations in the main paper, we prepend "S" when numbering tables, figures and equations contained in this supplement.

Code

All Z-tests described below are implemented in the spectralBacktest package for `R`. It is available for download at `https://github.com/ajmcneil/spectralBacktest`.

Contents

*The opinions expressed here are our own, and do not reflect the views of the Board of Governors or its staff. Email: `michael.gordy@frb.gov` and `alexander.mcneil@york.ac.uk`.

S.A 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

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

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

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

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

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

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 form in (S.1), numerical solution to ${}_3F_2(c_1, c_2, c_3; d_1, d_2; 1)$ 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:

$$(a+b)W_{a,b} = aW_{a+1,b} + bW_{a,b+1} \quad \text{(S.2)}$$

An immediate implication is that the uniform, linear increasing and linear decreasing trans-

formations (parameter sets (1,1), (2,1) and (1,2), respectively) are linearly dependent. 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.

S.B Backtest sample size in the test of unconditional coverage

In this appendix, we explore the effect of backtest sample size n on the size and power of unconditional Z-tests and LR-tests. Our purpose is to demonstate that the findings of Section 4.3 of the main paper are robust to the choice of n. Recall that the baseline sample size is $n = 750$, which corresponds approximately to a three-year sample of bank data. Here we consider $n = 250$ (one year) and $n = 500$ (two years) as well. All mnemonic kernel identifiers are as defined in the main paper.

We report in Table S.1 the percentage of rejections of the null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. The bottom panel, for $n = 750$, replicates Table 2 in the main paper. Comparing to the results for $n = 250$ (upper panel) and $n = 500$ (middle panel), we confirm that size and power improve as n increases, as one would expect. Comparing across columns, we find that the qualitative description of results in Section 4.3 holds for each value of n.

In Table S.2, we compare the size and power of LR-tests against Z-test counterparts. The bottom panel, for $n = 750$, replicates Table 3 in the main paper. Whereas for $n = 750$ the three-point multinomial LR-test is most oversized, for $n = 250$ and $n = 500$ it is the one-point Kupiec (1995) LR-test (LR1) that is most oversized. When the true model F is the scaled t_5, the Z-tests in each case offer greater power than the corresponding LR-test with a minor exception for the BIN vs. LR1 comparison under $n = 500$. When the true model F is the scaled t_3, the Z-tests typically offer more power than the LR-tests on the narrow window but less power on the wide window.

		Monospectral							Bispectral			Trispectral	
window	F | kernel	BIN	ZU3	ZU	ZA	ZE	ZL_+	ZL_-	PE2	ZLL	ZPP	PE3	ZPUP
						$n = 250$							
narrow	Normal	4.1	4.2	3.9	3.9	3.9	4.1	3.7	6.1	5.3	5.2	5.0	6.0
	Scaled t5	17.4	19.6	18.5	18.9	18.0	22.0	14.6	27.3	20.9	22.6	18.0	19.6
	Scaled t3	13.4	15.3	14.3	14.7	13.8	19.2	9.7	27.2	20.8	23.4	17.5	19.1
wide	Normal	4.1	4.4	4.8	4.8	4.8	4.7	4.8	6.2	4.8	5.0	5.2	5.3
	Scaled t5	17.4	8.1	5.9	6.3	5.7	8.9	4.9	32.0	17.2	25.6	23.0	20.9
	Scaled t3	13.4	9.1	7.7	9.1	6.8	6.3	10.9	49.7	30.2	42.7	36.1	31.3
						$n = 500$							
narrow	Normal	3.9	4.6	4.6	4.6	4.5	4.6	4.6	3.8	4.7	4.7	5.4	5.4
	Scaled t5	22.1	27.1	26.5	26.9	25.7	31.5	21.6	31.7	30.2	33.1	30.9	30.0
	Scaled t3	15.9	20.2	19.6	20.1	18.7	26.4	14.0	35.3	31.0	35.7	31.8	30.5
wide	Normal	3.9	4.7	4.9	4.9	4.8	4.7	4.9	4.5	4.8	4.8	5.1	4.9
	Scaled t5	22.1	9.7	6.3	6.5	6.0	10.6	5.4	44.2	31.3	43.3	40.3	36.6
	Scaled t3	15.9	11.3	12.8	14.8	11.1	6.8	21.5	78.9	64.9	77.5	70.9	67.7
						$n = 750$							
narrow	Normal	6.1	4.9	4.7	4.7	4.7	4.6	4.8	4.8	4.8	4.8	5.3	5.2
	Scaled t5	33.9	35.0	33.8	34.4	33.0	40.3	27.1	44.0	40.0	45.3	40.3	39.3
	Scaled t3	24.0	24.8	23.9	24.3	23.3	32.7	16.5	50.7	43.3	50.9	43.4	42.7
wide	Normal	6.1	5.0	4.9	4.9	4.9	4.9	4.9	4.8	5.0	4.9	5.1	5.0
	Scaled t5	33.9	10.7	6.4	6.6	6.1	11.9	5.8	60.7	45.1	59.2	55.5	51.8
	Scaled t3	24.0	13.5	17.7	20.4	15.4	7.4	31.9	94.0	85.8	93.0	90.6	88.4

Table S.1: Estimated size and power of unconditional Z-tests.
We report the percentage of rejections of the null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. The narrow window is [0.985, 0.995] and the wide window is [0.95, 0.995].

window	F | test	BIN	LR1	PE2	LR2	PE3	LR3	ZPP	LRB
				$n = 250$					
narrow	Normal	4.1	9.3	6.1	5.7	5.0	3.2	5.2	7.0
	Scaled t5	17.4	10.6	27.3	17.9	18.0	13.5	22.6	17.8
	Scaled t3	13.4	9.0	27.2	22.5	17.5	15.1	23.4	22.5
wide	Normal	4.1	9.3	6.2	4.6	5.2	4.2	5.0	5.7
	Scaled t5	17.4	10.6	32.0	24.5	23.0	19.6	25.6	24.0
	Scaled t3	13.4	9.0	49.7	53.1	36.1	43.6	42.7	52.2
				$n = 500$					
narrow	Normal	3.9	7.0	3.8	5.6	5.4	5.9	4.7	5.8
	Scaled t5	22.1	22.5	31.7	25.4	30.9	25.6	33.1	26.9
	Scaled t3	15.9	16.5	35.3	33.3	31.8	33.0	35.7	35.3
wide	Normal	3.9	7.0	4.5	6.1	5.1	6.0	4.8	5.2
	Scaled t5	22.1	22.5	44.2	41.1	40.3	38.2	43.3	41.3
	Scaled t3	15.9	16.5	78.9	82.4	70.9	78.1	77.5	82.5
				$n = 750$					
narrow	Normal	6.1	4.1	4.8	6.3	5.3	8.2	4.8	5.5
	Scaled t5	33.9	24.0	44.0	36.5	40.3	34.3	45.3	37.6
	Scaled t3	24.0	16.1	50.7	47.7	43.4	46.5	50.9	49.2
wide	Normal	6.1	4.1	4.8	5.9	5.1	7.3	4.9	5.1
	Scaled t5	33.9	24.0	60.7	57.1	55.5	53.0	59.2	57.7
	Scaled t3	24.0	16.1	94.0	94.8	90.6	93.0	93.0	95.0

Table S.2: Estimated size and power of unconditional Z-tests and LR-tests.
We report the percentage of rejections of the null hypothesis at the 5% confidence level based on $2^{16} = 65{,}536$ replications. The narrow window is [0.985, 0.995] and the wide window is [0.95, 0.995].

S.C A regression derivation of the conditional spectral Z-test

Consider the regression model

$$W_t - \mu_W = \boldsymbol{\beta}' \boldsymbol{h}_{t-1} + \epsilon_t, \quad t = k+1, \ldots, n.$$

The least squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = X'(\boldsymbol{W} - \mu_W \mathbf{1})$ where X is the $(n-k) \times (k+1)$ matrix whose rows are given by $\boldsymbol{h}_{t-1}$ for $t = k+1, \ldots, n$, $\boldsymbol{W} = (W_{k+1}, \ldots, W_n)'$ and $\mathbf{1}$ is the $(n-k)$-vector of ones. Moreover, under the usual assumptions of time series regression (see, for example, Hayashi (2000)) $\hat{\boldsymbol{\beta}}$ is asymptotically normal with covariance matrix $\sigma_W^2 (X'X)^{-1}$. Under our null hypothesis $\boldsymbol{\beta} = \mathbf{0}$ and μ_W and σ_W^2 are known and a test may be based on the statistic

$$\sigma_W^{-2} (\boldsymbol{W} - \mu_W \mathbf{1})' X (X'X)^{-1} X' (\boldsymbol{W} - \mu_W \mathbf{1}) \sim \chi^2_{k+1} \, .$$

This may be shown to be identical to (14) on observing that $\overline{\boldsymbol{Y}}_{n,k} = (n-k)^{-1} X' (\boldsymbol{W} - \mu_W \mathbf{1})$ and $\hat{\Sigma}_Y = \sigma_W^2 (n-k)^{-1} X'X$.

S.D Size and power of tests of conditional coverage

In this appendix, we explore the effect of kernel choice and true model F on the size and power of MD-tests of conditional coverage. Our purpose is to demonstate that the findings of Section 5.2 of the main paper are robust. The CVT choices are defined in Table 4 in the main paper. Note that when CVT takes the value *None*, the test is an unconditional test.

Table S.3 estimates the size of the tests when samples are uniformly distributed and serially independent. The same messages emerge as in the excerpt in Section 5.2: when the CVT is DQ or V.BIN the tests are quite badly oversized, particularly for the former; choosing V.4 or V.½ as CVT substantially mitigates (but does not eliminate) oversizing.

Table S.4 is an examination of power for the same kernels and CVT functions. The aim of the underlying simulation is to produce pseudo PIT values that are (i) serially dependent with a dependence structure that is typical when stochastic volatility in the data is ignored and (ii) possibly non-uniform with the same distributions used for the unconditional tests. The data generating process used is a VT-ARMA(1,1) model as described in McNeil (2020). This process is designed to mimic the behavior of volatile financial return series, such as daily log-returns on a stock index, and may be motivated as follows.

Given empirical loss data $L_1, \ldots, L_n$, suppose we form a version of the empirical cdf by taking $F_n(x) = (n+1)^{-1} \sum_{t=1}^{n} \mathbb{1}_{\{L_t \leqslant x\}}$ and we use this to construct data $\widehat{U}_t = F_n(L_t)$.

		Monospectral				Bispectral	
window	CVT | kernel	BIN	ZU	ZL_+	ZL_-	ZLL	ZPP
narrow	None	6.2	4.8	4.6	4.9	4.9	4.9
	DQ	13.3	14.4	16.0	11.5	11.8	14.2
	V.BIN	8.0	9.0	10.4	8.4	8.5	14.6
	V.4	6.8	6.7	7.2	6.4	6.6	7.9
	V.½	6.7	6.7	7.0	6.4	6.6	7.7
wide	None	6.2	4.9	5.0	4.9	4.9	4.8
	DQ	13.3	8.5	9.2	8.3	8.3	14.8
	V.BIN	8.0	7.3	8.1	7.0	6.9	12.1
	V.4	6.8	5.3	5.6	5.2	5.3	7.4
	V.½	6.7	5.5	5.7	5.3	5.5	7.3

Table S.3: Estimated size of tests of conditional coverage.
Replications = 65,536. The number of days in the backtest sample is $n = 750$. The narrow window is [0.985, 0.995] and the wide window is [0.95, 0.995].

The transformed returns $(\widehat{U}_t)$ are close to uniformly distributed in the open interval $(0, 1)$ and show negligible serial correlation. However, under the v-shaped transformation $V(u) = |2u - 1|$, we obtain data $(V(\widehat{U}_t))$ which remain approximately uniformly distributed but show strong serial correlation. For simplicity, assume n is even, as this guarantees that the transformed data $(V(\widehat{U}_t))$ are in the open interval $(0, 1)$.[1] Now consider applying the normal quantile function Φ^{-1} to the transformed data to get values that are approximately standard normally distributed. The net effect of these three transformations is described by the function $T_n(l) = \Phi^{-1}(V(F_n(l)))$. If we fit Gaussian ARMA models to the transformed data $(T_n(L_t))$ we often find that an ARMA(1,1) process fits well and typical values for the AR and MA parameters are around 0.95 and -0.85.

A VT-ARMA(1,1) process (L_t) with marginal distribution F is a stochastic process that mimics this behavior. If we define $T(l) = \Phi^{-1}(V(F(l)))$, then the transformed process $(T(L_t))$ is an ARMA(1,1) process. Such a process is straightforward to construct as shown in McNeil (2020). In Table S.4 we simulate VT-ARMA(1,1) processes with different marginal distributions F but the same underlying ARMA(1,1) serial dependence structure with AR parameter 0.95 and MA parameter -0.85 under the transformation T. As for the unconditional tests, pseudo PIT values are obtained by the transformation $P_t = \Phi(L_t)$.

In Table S.4 we observe that there is generally a very large increase in power when we move from the unconditional tests (CVT = *None*) to the conditional tests. This is evident even when the distribution of the simulated data is uniform (F = Normal). The most

[1] To cover the case where n is odd we can use the transformation $V_n(u) = |2u - 1| + \mathbb{1}_{\{n \text{ is odd}\}}/(n + 1)$.

powerful tests are bispectral tests applied to the wider window using the CVT functions V.4 and V.½.[2]

S.E Identification of spurious PIT values

Consider a stylized Gaussian model in which loss is given by $L_t = \sigma_{t-1} Z_t$, 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

$$Z_t = \Phi^{-1}(0.99) \times L_t / \widehat{\mathrm{VaR}}_t. \tag{S.3}$$

Furthermore, the PIT values can be expressed as

$$P_t = \widehat{F}_{t-1}(L_t) = \Phi(L_t/\sigma_{t-1}) = \Phi(Z_t). \tag{S.4}$$

In general, we would not expect the Z_t to be Gaussian, so (S.4) will not hold. However, so long as (Z_t) is iid, there will still be a monotonic relationship between Z_t (as defined by (S.3)) 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 (S.4) by

$$P_{i,t} = H(Z_{i,t};\theta_i) + \epsilon_{i,t} \tag{S.5}$$

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:

[2]The only exception to these observations is the case of the bispectral tests on the wide window kernel when the true model is the scaled t_3 distribution. In this case, the unconditional test is already so powerful that the presence of additional moments in the conditional test tends to dilute the test.

			Monospectral				Bispectral	
window	F	CVT | kernel	BIN	ZU	ZL_+	ZL_-	ZLL	ZPP
narrow	Normal	None	12.1	10.8	10.0	11.2	9.1	9.4
		DQ	30.0	31.5	32.0	29.4	29.1	28.7
		V.BIN	28.1	30.9	31.8	30.2	29.5	34.0
		V.4	30.3	32.6	30.7	33.6	32.3	26.3
		V.½	19.3	21.7	19.9	22.5	22.0	16.1
	Scaled t5	None	36.0	36.2	40.9	31.4	41.2	45.6
		DQ	47.4	54.9	59.7	46.1	53.4	59.1
		V.BIN	48.4	52.7	56.8	49.5	54.7	64.3
		V.4	56.4	60.7	63.1	57.2	61.1	65.6
		V.½	49.6	54.5	57.3	50.5	55.6	59.9
	Scaled t3	None	28.1	28.3	35.0	22.2	44.1	50.8
		DQ	42.2	50.4	56.1	39.7	53.3	57.4
		V.BIN	42.5	47.3	53.5	42.6	53.3	67.0
		V.4	50.1	54.8	58.9	49.7	58.8	67.5
		V.½	44.1	49.5	54.1	43.7	54.2	63.5
wide	Normal	None	12.1	17.8	15.9	18.3	14.6	15.2
		DQ	30.0	31.1	30.1	31.6	30.4	30.9
		V.BIN	28.1	36.0	34.5	36.2	34.6	34.4
		V.4	30.3	52.1	46.4	54.1	51.2	30.7
		V.½	19.3	44.9	36.9	48.0	44.7	20.2
	Scaled t5	None	36.0	19.5	22.3	19.2	52.9	65.3
		DQ	47.4	35.4	38.9	31.5	49.8	63.8
		V.BIN	48.4	41.0	45.3	37.7	55.1	70.5
		V.4	56.4	55.2	56.7	52.8	67.2	74.8
		V.½	49.6	51.1	51.4	49.2	65.7	71.0
	Scaled t3	None	28.1	28.4	18.5	39.3	87.6	93.9
		DQ	42.2	32.8	32.6	34.2	71.1	79.7
		V.BIN	42.5	38.8	38.7	40.1	75.3	87.6
		V.4	50.1	50.7	48.4	52.7	82.5	91.5
		V.½	44.1	48.1	43.4	51.2	83.6	91.0

Table S.4: Estimated power of tests of conditional coverage when DGP is a VT-ARMA(1,1) model. Replications = 65,536. The number of days in the backtest sample is $n = 750$. The narrow window is [0.985, 0.995] and the wide window is [0.95, 0.995].

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 6.3, we use the imputed values to fill in for spurious values in forming regressors in the tests of conditional coverage.

References

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

Hayashi, F., 2000, *Econometrics* (Princeton University Press, Princeton, NJ).

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

McNeil, Alexander J., 2020, Modelling volatility with v-transforms, Working Paper 2002.10135, arXiv.

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