# Spectral Backtests of Forecast Distributions

## with Application to Risk Management

Michael B. Gordy[1], Hsiao Yen Lok[2], Alexander J. McNeil[3]

[1]Federal Reserve Board [2]Heriot-Watt [3]York Management School

Risk Measurement & Regulatory Issues in Business
September 2017

The opinions expressed are our own, and do not reflect the views of the Board of Governors or its staff.

# Traditional setting for backtesting

- We take a new look at the old risk management problem of backtesting.
- Consider a bank with a one-day-ahead forecast of its loss, defined as negative P&L. We use this notation throughout:

  $\mathcal{F}_t$ : Information available at $t$ (filtration).
  $L_t$ : Loss realized at $t$ on portfolio formed at $t-1$.
  $F_t$ : $F_t(y) = \Pr(L_t \leq y | \mathcal{F}_{t-1})$; i.e., the df of the day-ahead forecast distribution.
  $\widehat{F}_t$ : The forecast distribution formed by the bank's risk-manager.

- In the regulatory context, it is generally assumed that the backtest is based on VaR exceedances.
  - $\widehat{\text{VaR}}_{\alpha,t} := \widehat{F}_t^{\leftarrow}(\alpha)$ is an estimate of $\alpha$-VaR constructed at time $t-1$.
  - Bank reports $\widehat{\text{VaR}}_{\alpha,t}$ and realized $L_t$.
  - VaR exceedance is simply $I_t = I[L_t > \widehat{\text{VaR}}_{\alpha,t}]$.

# Probability integral transform

- Increasingly, regulators can observe more than just the VaR exceedances.
- Consider the PIT process given by $P_t = \widehat{F}_t(L_t)$.
- Reported PIT values contain information about VaR exceedances at every level $\alpha$.

$$P_t \geq \alpha \iff L_t \geq \widehat{\mathrm{VaR}}_{\alpha,t}$$

- If the $\{\widehat{F}_t\}$ coincide with the true $\{F_t\}$, then the process $\{P_t\}$ is iid $U[0,1]$.
- In the US, banks on the Internal Models Approach for the trading book have been required to report PIT values to regulators since 2013.
- Motivation: What is the best way to exploit this additional information?

# Simulated example of a backtest dataset

| Days | VaR | Loss | Exceed? | PIT |
|---|---|---|---|---|
| 1 | 2.492 | 0.278 | 0 | 0.602 |
| 2 | 2.968 | 0.716 | 0 | 0.713 |
| 3 | 3.336 | -0.759 | 0 | 0.298 |
| 4 | 3.018 | -0.451 | 0 | 0.364 |
| 5 | 2.654 | 2.955 | 1 | 0.995 |
| 6 | 3.335 | -1.697 | 0 | 0.118 |
| 7 | 3.137 | 0.184 | 0 | 0.554 |
| 8 | 2.641 | 1.091 | 0 | 0.832 |

# Some quantiles of greater interest than others

- Diebold, Gunther, and Tay (1998) develop forecast density tests based on PIT values. They show how to test the null hypothesis that $\{P_t\}$ is iid $U[0,1]$.
- In a risk-management context, some quantiles of the forecast distribution are more important than others.
  - Accuracy in "good tail" of high profits (low $P_t$) is generally much less important than accuracy in the "bad tail" of large losses (high $P_t$).
  - Models generally cannot expected to perform well in the extreme tail of once-per-generation shock.
- We study a class of backtests for forecast distributions in which the test statistic weights exceedance events by a function of the probability level $\alpha$.
- The choice of the kernel function makes explicit the priorities for model performance.

# Power vs Specificity

Applied to the regulator's problem of backtesting bank VaR, our results point to a tradeoff between specificity and power.

- Under the current regulatory framework, capital depends only on 99% VaR.
- So long as the model estimates VaR accurately, the regulator could be indifferent to its performance at other quantiles.
- Taking this narrow view, the sequence of VaR exceedance indicators is sufficient for backtesting.
- Exceedance indicators arise as a limiting case of the kernel function, so many of the traditional backtests are nested in our framework.
- But if the regulator is willing to assign positive mass to probability levels in a neighborhood of $\alpha$, we can construct more powerful backtests.

# A broader view of the backtesting mandate

We take a broader view of the risk-manager's mandate, in which the objective is to forecast probabilities over a *range* of large losses.

- We might readily believe that a model designed to deliver 99%-VaR could fail to deliver an accurate estimate of the $75^{\text{th}}$ percentile.
- But if the bank fails to forecast losses at the 98.5% or 99.5% level, is it plausible to trust that the modeling of the 99% VaR is robust?
- The formal guidance of US regulators to banks on internal model validation (SR Letter 11-7) explicitly requires "checking the distribution of losses against other estimated percentiles."

# Fundamental Review of the Trading Book

- Reforms mandated by the FRTB due to begin parallel run in 2018.
- FRTB replaces 99%-VaR with 97.5%-Expected Shortfall (ES) as determinant of capital requirements.
  - This has led to much debate on whether or not ES is amenable to direct backtesting.
  - The model approval process will continue to be based on VaR exceedances.
- We devise tests of the forecast distribution from which risk measures are estimated and not tests of the risk measure estimates.
  - For ES, a test of the tail of the forecast distribution offers an indirect approach to backtesting.
  - FRTB requires banks to go beyond the mandatory VaR backtesting regime to consider multiple levels or other features of the tail.
  - Basel Committee explicitly mentions the use of PIT values as a possible direction for the extended model validation requirements.
- To fix ideas, we henceforth assume the backtest is conducted by a regulator who is interested primarily in assessing the bank's 99%-VaR forecast, but
  - our conclusions hinge little on the choice of risk measure; and
  - apply equally to internal assessments of forecasting performance.

# Spectral transformations of PIT exceedances

- Our tests are based on transformations of indicator variables for PIT exceedances.
  - We mean "spectral" in the integral transform sense, not in the Durlauf (1991) sense of a transformed autocovariance sequence.
- The transformations take the form

$$W_t = \int_0^1 I_{\{P_t > u\}} \mathrm{d}\nu(u) = \nu([0, P_t))$$

  where $\nu$ is a measure defined on $[0, 1]$.
  - $\nu$ is chosen to apply weight to different levels in the unit interval, typically in the region of the standard VaR level $\alpha = 0.99$.
  - We refer to $\nu$ as the kernel measure for the transform.
- $W_t$ is (weakly) increasing in $P_t$.

# Spectral backtests

- Spectral backtests are backtests based on $W_1, \ldots, W_n$.
- Null hypothesis. Let $F_W^0$ denote df of $W_t = \nu([0, P_t))$ when $P_t$ is uniform.

$$H_0 : \quad W_1, \ldots, W_n \text{ are iid with df } F_W^0.$$

- Within the class of spectral backtests, we have
  - tests of unconditional coverage: test for correct distribution $F_W^0$;
  - tests of conditional coverage: correct distribution and serial independence.

# Discrete kernels

$$W_t = \sum_{i=1}^{m} k_i I_{\{P_t > \alpha_i\}}$$

- Special case $m = 1$: $W_t \propto I_{\{P_t > \alpha\}}$ is exceedance indicator for $\alpha$-VaR.
  - $(1/k_1) \sum_i W_i$ is distributed Binomial$(n, 1-\alpha)$ under $H_0$.
  - Possible tests: Z-test (classical binomial score test), LRT (Kupiec '95, Christoffersen '98).
- General case yields multinomial tests.
  - $W_t$ takes values in $0 = q_0 < q_1 < \cdots < q_m$ where $q_j = \sum_{i=1}^{j} k_i$.
  - Form count variables $O_i = \sum_{t=1}^{n} I_{\{W_t = q_i\}}, \quad i = 0, 1 \ldots, m$.
  - Under $H_0$: $(O_0, \ldots, O_m) \sim \text{MN}(n, (\alpha_1, \alpha_2 - \alpha_1, \ldots, \alpha_m - \alpha_{m-1}, 1 - \alpha_m))$.
  - Possible tests: Pearson-Nass score test, LRT.

# Continuous kernels

- A continuous kernel measure has density $\mathrm{d}\nu(u) = g(u)\mathrm{d}u$ on an interval $[\alpha_1, \alpha_2] \subset [0, 1]$, where
  - $g$ is continuous on $[\alpha_1, \alpha_2]$,
  - $g(u) > 0, u \in (\alpha_1, \alpha_2)$, and
  - $g(u) = 0, u \notin [\alpha_1, \alpha_2]$.
- The kernel density $g$ plays the same role as the "kernel function" in the nonparametric statistics literature, and $[\alpha_1, \alpha_2]$ is the kernel window.
- When $g$ satisfies the additional requirement that $\int_{\alpha_1}^{\alpha_2} g(u)\mathrm{d}u = 1$, is a normalized kernel density.
- In nonparametric statistics, the kernel is often defined to be normalized and symmetric, but we do not impose these requirements.
- Writing $G$ for the integral of $g$, we have $W_t = G(P_t^*)$ for truncated PIT-value $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$.
- $G$ strictly increasing inside kernel window, so $W_t$ strictly increasing in $P_t^*$.
- Can also have mixed kernels.

# Intuition for continuous kernel

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

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

- The measures $\nu_{\alpha,\epsilon}$ defined by $g_{\alpha,\epsilon}$ converge to Dirac measure $\delta_\alpha$ as $\epsilon \to 0+$, and $\lim_{\epsilon \to 0} W_t = I_{\{P_t > \alpha\}}$ almost surely.
- Thus, classic tests based on the exceedance indicator $I_{\{P_t > \alpha\}}$ can be seen as limiting cases of more general continuous tests as the width $\epsilon$ of the kernel window vanishes to zero.

# Test of unconditional coverage

- Test for correct distribution $F_W^0$ implied by $H_0$ and choice of kernel.
- Broadly, our tests fall into two categories, Z-tests and likelihood ratio tests.
- Presentation focuses on continuous kernel case, but discrete case works the same way.

# Z-tests

Z-tests are based on the asymptotic normality under $H_0$ of

$$\overline{W}_n = n^{-1} \sum_{t=1}^{n} W_t$$

- Solve for $\mu_W = \mathbb{E}(W_t)$ and $\sigma_W^2 = \text{var}(W_t)$ in the null model $F_W^0$.
- Trivially follows from CLT that, under $H_0$,

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

- Multivariate extensions of Z-tests are chi-squared tests.
- In general, Z-tests are sensitive to the choice of kernel measure.

# Likelihood ratio tests

- LRT are based on parametric models $F_W(w \mid \theta)$ for $W_1, \ldots, W_n$ that nest the model in $H_0$.
  - In other words $F_W^0 = F_W(\cdot \mid \theta_0)$ for some value $\theta_0$.
- Test is based on the asymptotic distribution of the statistic

$$\mathrm{LR}_{W,n} = L_W(\theta_0 \mid \mathbf{W}) / L_W(\hat{\theta} \mid \mathbf{W})$$

  where $\hat{\theta}$ denotes the maximum likelihood estimate (MLE).
- A key difference with Z-tests is that LRT depends only on the support of the kernel measure, and not the distribution of mass within the support.
  - In the continuous case, it is boundaries of the kernel window $[\alpha_1, \alpha_2]$ that determines the test statistic and not the kernel density $g$.
  - This result is a consequence of the well-known invariance of the LRT under strictly increasing tranformations.
- Another key distinction with Z-test is that LRT requires estimation of $\hat{\theta}$ under the alternative.

# Continuous spectral Z-test

- Product rule for continuous kernels
  - Say we have two kernel densities $g_1$ and $g_2$ on the same window $[\alpha_1, \alpha_2]$.
  - Let $W_t^{(j)}$, $j = 1, 2$, be the corresponding spectrally-transformed PIT values.
  - Then the product $W_t^* = W_t^{(1)} W_t^{(2)}$ is a spectrally-transformed PIT value generated by density

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

    on the same window $[\alpha_1, \alpha_2]$.

- For a kernel density $g$ on $[\alpha_1, \alpha_2]$, we have $\mu_W = \mathbb{E}(W_t)$ and $\sigma_W^2 = \mathbb{E}(W_t^2) - \mu_W^2$ where

$$\mathbb{E}(W_t) = \int_{\alpha_1}^{\alpha_2} g(u)(1-u)\mathrm{d}u,$$

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

- The spectral Z-test for the sample mean $\overline{W}_n$ follows immediately.

# Continuous bispectral Z-test

Consider $\mathbf{W}_t = (W_t^{(1)}, W_t^{(2)})'$ for two distinct kernel measures on same support.

- Use the product rule to calculate $\mu_W = \mathbb{E}(\mathbf{W}_t)$ and $\Sigma_W = \text{cov}(\mathbf{W}_t)$.
- Under $H_0$, for large $n$, we have approximately that

$$n\left(\overline{\mathbf{W}} - \mu_W\right)' \Sigma_W^{-1} \left(\overline{\mathbf{W}} - \mu_W\right) \sim \chi_2^2.$$

- Straightforward to generalize to higher dimensions ($J$ kernels).

# Truncated probitnormal tests

- We assume $P_t$ is probitnormal: $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$ with df $F_P(p \mid \theta)$ and density $f_P(p \mid \theta)$, $\theta = (\mu, \sigma)'$.
- Nests uniform distribution: $\theta = \theta_0 = (0, 1)'$.
- Tests based on truncated PIT-values $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$.
- Likelihood function for $P_t^*$ has closed-form expression.
- Denote the observed score vector for $P_t^*$ by

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

- We can show that the score can be written as a vector of spectrally-transformed PIT values.
  - $\nu_1$ and $\nu_2$ a bit tedious to write out, but solved analytically.

  Thus, the score test is a special case of the bispectral Z-test.
- LRT is also a spectral test and generalizes a test proposed by Berkowitz (2001).

# Parametric simulation study

- Sample data $L_t$ from 4 distributions $F$: normal, t5, t3, skewed t3.
- All distributions are scaled to have *mean zero and variance one*.

| $F$ | $\text{VaR}_{0.975}$ | $\text{VaR}_{0.99}$ | $\Delta_1$ | $\text{ES}_{0.975}$ | $\Delta_2$ |
|---|---|---|---|---|---|
| Normal | 1.96 | 2.33 | 0.00 | 2.34 | 0.00 |
| t5 | 1.99 | 2.61 | 12.04 | 2.73 | 16.68 |
| t3 | 1.84 | 2.62 | 12.69 | 2.91 | 24.46 |
| st3 ($\gamma = 1.2$) | 2.04 | 2.99 | 28.68 | 3.35 | 43.11 |

- $\Delta_1$ shows the percentage increase in the value of $\text{VaR}_{0.99}$ for each model compared to a normal model. $\Delta_2$ does the same for $\text{ES}_{0.975}$.
- We assume $\widehat{F}_t = \Phi$ and apply tests to data $P_t = \Phi(L_t)$.
  - When $F$ is normal, the data $P_t$ are uniform.
  - Otherwise $P_t$ will show departures from uniformity that would be typical when tail is underestimated.

# Menagerie of discrete kernels

- Set of reasonable candidates for kernel measure $\nu$ is very large. We begin with a few representative discrete examples.
- Discrete kernels on supports of cardinality 1, 3, 5:

  1 point: $\alpha = 0.99$.
  3 points: $\{0.985, 0.99, 0.995\}$.
  5 points: $\{0.985, 0.9875, 0.99, 0.9925, 0.995\}$.

- Discrete tests of five types.

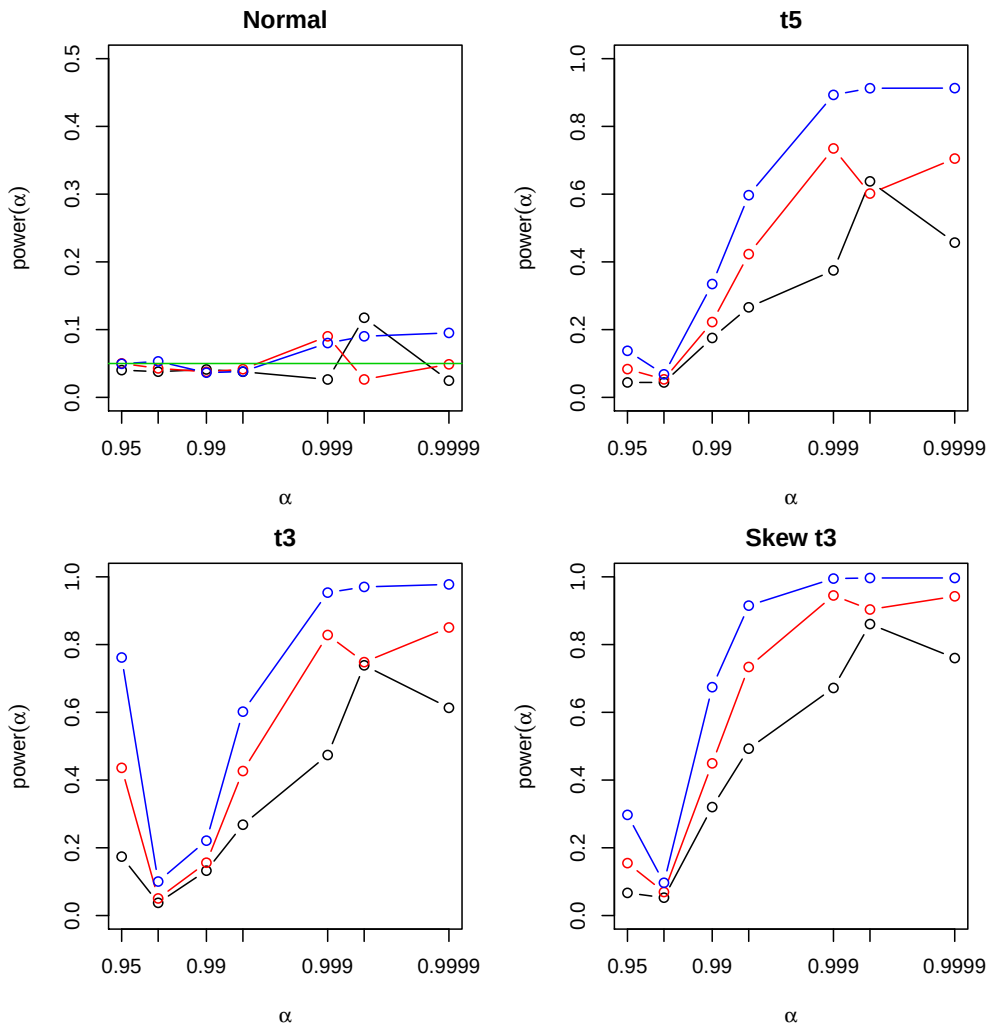
  B99: Binomial score test on VaR exceedance at $\alpha = 0.99$.
  PN$m$: Pearson-Nass multinomial score test on $m$ points.
  LR$m$: Multinomial LRT on $m$ points.
  ZU$m$: Spectral test with uniform weights on $m$ points.
  ZE$m$: Spectral test with exponential weights rising from $k_1 = 1$ to $k_m = exp(2) \approx 7.4$.

# Size and power of unconditional discrete tests

| $F$ | $n$ \| test | B99 | PN3 | LR3 | ZU3 | ZE3 | PN5 | LR5 | ZU5 | ZE5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal | 250 | 4.0 | 3.7 | 3.0 | 4.1 | 4.5 | 4.3 | 2.0 | 3.7 | 4.1 |
| | 500 | 3.7 | 5.1 | 5.7 | 4.6 | 4.5 | 5.0 | 4.9 | 4.6 | 4.8 |
| | 1000 | 3.8 | 4.5 | 5.9 | 5.0 | 4.9 | 4.5 | 7.3 | 5.4 | 5.0 |
| t5 | 250 | 17.7 | 15.4 | 13.5 | 20.0 | 26.6 | 13.3 | 8.3 | 19.0 | 23.4 |
| | 500 | 22.4 | 30.7 | 25.7 | 27.8 | 36.6 | 21.6 | 19.2 | 27.4 | 34.4 |
| | 1000 | 33.0 | 48.0 | 42.0 | 40.3 | 55.0 | 38.6 | 38.4 | 40.6 | 51.6 |
| t3 | 250 | 13.5 | 14.3 | 15.4 | 15.6 | 25.0 | 12.4 | 8.1 | 14.5 | 20.9 |
| | 500 | 16.2 | 32.1 | 33.6 | 20.5 | 33.5 | 19.6 | 22.9 | 20.5 | 30.5 |
| | 1000 | 22.3 | 55.4 | 56.4 | 28.0 | 49.1 | 40.8 | 53.2 | 28.7 | 44.8 |
| st3 | 250 | 31.2 | 31.5 | 30.2 | 35.0 | 46.3 | 27.6 | 20.8 | 33.2 | 41.6 |
| | 500 | 44.2 | 60.9 | 55.1 | 51.3 | 64.8 | 48.0 | 45.7 | 50.8 | 62.0 |
| | 1000 | 66.2 | 85.9 | 82.8 | 73.8 | 87.5 | 79.3 | 78.5 | 74.3 | 85.2 |

# Power of binomial test in the extreme tail

- Test most powerful in far tail.
- But also most erratic: At $\alpha = 0.9999$, one exceedance in $n = 1000$ enough to reject.
- Nonmonotonic in $\alpha$.
- ZE3 weights very heavily on $\alpha_3 = 0.995$, so similar to Binomial($\alpha_3$) test.
- Poor performance of ZU3 due to weight at $\alpha_1 = 0.985$.
- Again, choose $\nu$ to fit priorities, not to chase power.




$n = 250$, $n = 500$, $n = 1000$.

# Menagerie of continuous kernels

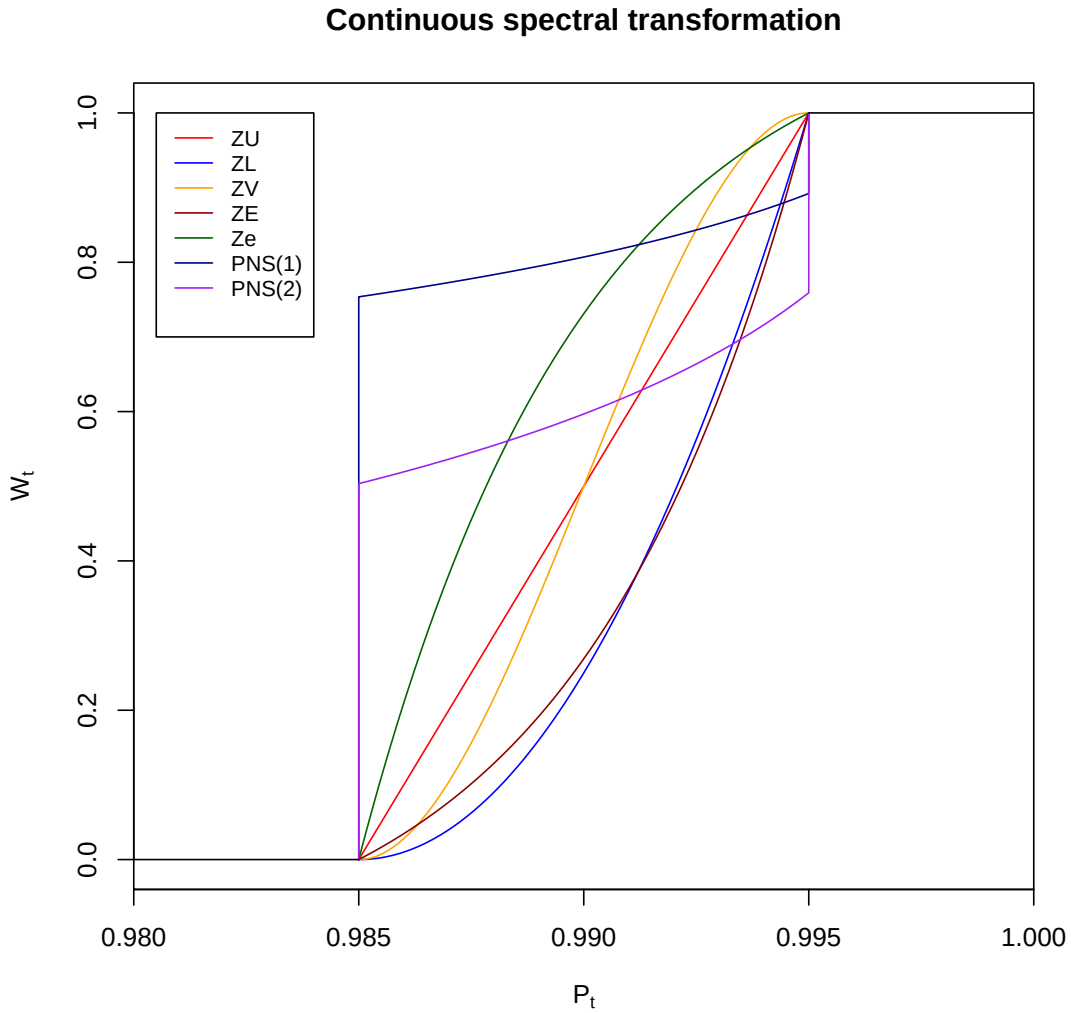
- Kernel densities
  - Ze: Exponential kernel, decreasing $g(u) = \exp\left(\kappa\left(\frac{u-\alpha_1}{\alpha_2-\alpha_1}\right)\right)$ for $\kappa = -2$.
  - ZE: Exponential kernel, increasing $\kappa = 2$.
  - ZV: Epanechnikov kernel. Hump-shaped, symmetric around $\alpha = 0.99$.
  - ZU: Uniform kernel.
  - ZL: Linear kernel $g(u) = u - \alpha_1$
- Continuous bispectral tests: ZeE, ZUE, ZLE.
- Probitnormal LRT (PNL) and score test (PNS).

# Size and power of unconditional continuous tests

| $F$ | $n$ \| test | Ze | ZV | ZU | ZE | ZeE | ZUE | ZLE | PNL | PNS |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal | 250 | 3.7 | 3.8 | 4.0 | 4.1 | 5.0 | 5.0 | 5.1 | 7.0 | 5.0 |
| | 500 | 4.6 | 4.3 | 4.7 | 4.7 | 4.6 | 4.7 | 4.7 | 5.6 | 4.5 |
| | 1000 | 5.1 | 5.2 | 5.4 | 5.3 | 4.8 | 4.9 | 4.9 | 5.3 | 4.8 |
| t5 | 250 | 15.9 | 18.2 | 19.2 | 22.4 | 21.1 | 21.3 | 17.4 | 17.8 | 22.8 |
| | 500 | 22.7 | 25.8 | 26.9 | 31.5 | 30.5 | 31.1 | 25.6 | 27.1 | 33.5 |
| | 1000 | 32.8 | 38.6 | 39.9 | 47.4 | 49.4 | 50.0 | 39.2 | 47.1 | 53.8 |
| t3 | 250 | 11.0 | 13.7 | 14.5 | 19.6 | 21.2 | 21.5 | 13.2 | 23.0 | 23.3 |
| | 500 | 14.4 | 18.9 | 20.1 | 26.5 | 32.3 | 33.0 | 20.2 | 35.8 | 37.2 |
| | 1000 | 19.7 | 26.9 | 27.9 | 38.7 | 55.1 | 56.2 | 31.1 | 61.4 | 62.1 |
| st3 | 250 | 27.8 | 31.9 | 33.3 | 39.5 | 39.9 | 40.1 | 30.7 | 35.8 | 42.7 |
| | 500 | 41.6 | 48.4 | 50.3 | 58.0 | 59.6 | 60.3 | 49.9 | 57.3 | 64.0 |
| | 1000 | 64.0 | 72.2 | 73.5 | 81.7 | 86.1 | 86.6 | 74.9 | 86.0 | 89.2 |

Kernel window is $[0.985, 0.995]$.

# Normalized kernel measures



# Block tests of conditional coverage

- While the expected value of a sum does not depend on the dependence structure of the sample, the expected value of a product does. This can give power to detect deviations from serial independence.
- Let $(\tilde{W}_t)_{t\in\mathbb{N}}$ denote the sequence of centred values $\tilde{W}_t = W_t - \mu_W$.
- Divide the $n$ reported PIT-values into $N_B$ blocks of size $B$ and define block sums and products of $(W_t)$:

$$\mathbf{Y}_b = \left( \sum_{t=s_b}^{bB} \tilde{W}_t, \prod_{t=s_b}^{B} (\tilde{W}_t + \zeta) \right)', \quad s_b = (b-1)B + 1,$$

  for some chosen constant $\zeta \neq \mu_W$.
- Let $\overline{\mathbf{Y}} = N_B^{-1} \sum_{b=1}^{N_B} \mathbf{Y}_b$. Under the null, for fixed $B$ as $N_B \to \infty$,

$$N_B \left(\overline{\mathbf{Y}} - \mu_{\mathbf{Y}}\right)' \Sigma_{\mathbf{Y}}^{-1} \left(\overline{\mathbf{Y}} - \mu_{\mathbf{Y}}\right) \sim \chi_2^2,$$

  where $\mu_{\mathbf{Y}} = (0, \zeta^B)'$ and $\Sigma_{\mathbf{Y}}$ is a covariance matrix which can also be explicitly determined.
- Block bispectral tests and block probitnormal score tests can also be constructed.

# Martingale difference tests of conditional coverage

- We test for the martingale difference (MD) property $\mathbb{E}(\tilde{W}_t \mid \mathcal{F}_{t-1}) = 0$.
- For any $\mathcal{F}_{t-1}$-measurable $\mathbf{h}_{t-1}$ we must have $\mathbb{E}(\mathbf{h}_{t-1}\tilde{W}_t) = 0$.
- Form the *lagged* vector $\mathbf{h}_{t-1} = (1, h(P_{t-1}), \ldots, h(P_{t-k}))'$ for some function $h$ and test whether $\mathbb{E}\left(\mathbf{h}_{t-1}\tilde{W}_t\right) = \mathbf{0}, \quad t = k+1, \ldots, n.$
- Let $\mathbf{Y}_t = \mathbf{h}_{t-1}\tilde{W}_t$ for $t = k+1, \ldots, n$. Let $\overline{\mathbf{Y}} = (n-k)^{-1}\sum_{t=k+1}^{n}\mathbf{Y}_t$ and let $\hat{\Sigma}_Y$ denote a consistent estimator of $\Sigma_Y := \mathrm{cov}(\mathbf{Y}_t)$.
- Giacomini and White (2006) show that under very weak assumptions, for large enough $n$ and fixed $k$,

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

- The MD test generalizes the Dynamic Quantile test of Engle and Manganelli (2004).

# Designing experiment for tests of conditional coverage

- Idea is to capture behavior of reported PIT values when DGP $F_t$ features stochastic volatility, but SV is ignored in the bank's model $\widehat{F}_t$.
- Draw sequence of uniform rv with serial dependence generated by an AR(1) process.
- Generate losses $L_t = F^{-1}(U_t)$, where $F$ is normal, t5, t3, or st3.
- The bank reports PIT values from normal distribution: $P_t = \widehat{F}(L_t) = \Phi(L_t)$.
- When $F = \Phi$, PIT-values are a correlated sequence of uniformly distributed data.
- When $F \neq \Phi$, PIT-values are correlated and non-uniform. Tails will be underestimated.
- In MD tests, choose $h(P_{t-j}) = \Phi^{-1}(|2P_{t-j} - 1|)$ to target SV.

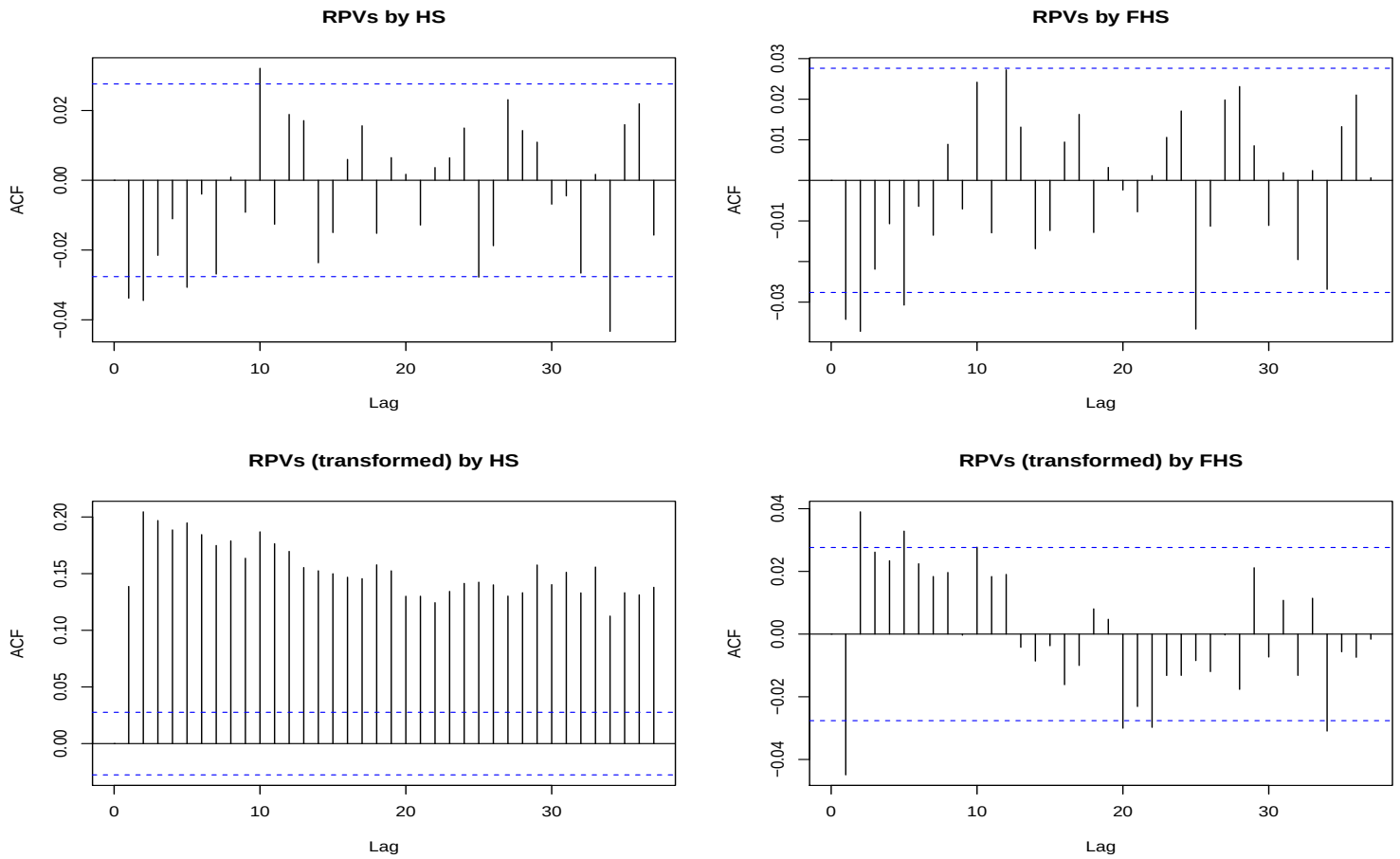
# Power of tests of conditional coverage

| $F$ | $n$ | ITT \| test | B99 | ZU | ZE | ZeE | ZUE | PNS |
|---|---|---|---|---|---|---|---|---|
| Normal | 250 | None | 5.6 | 5.3 | 5.3 | 6.4 | 6.3 | 6.3 |
| | | Block | 19.4 | 18.1 | 15.4 | 20.3 | 18.3 | 15.5 |
| | | MD | 27.6 | 29.5 | 28.0 | 29.6 | 28.7 | 30.3 |
| | 500 | None | 5.9 | 7.4 | 7.0 | 6.3 | 6.3 | 6.1 |
| | | Block | 34.3 | 25.9 | 23.4 | 20.9 | 24.8 | 22.4 |
| | | MD | 42.1 | 46.2 | 42.5 | 46.6 | 44.4 | 49.7 |
| | 1000 | None | 5.8 | 7.7 | 7.5 | 6.8 | 6.8 | 6.8 |
| | | Block | 25.3 | 30.8 | 31.5 | 30.2 | 29.1 | 29.6 |
| | | MD | 70.3 | 77.5 | 71.5 | 78.2 | 74.5 | 82.0 |
| t5 | 250 | None | 18.8 | 20.2 | 22.7 | 21.9 | 22.0 | 23.3 |
| | | Block | 32.9 | 33.5 | 32.2 | 35.5 | 34.1 | 33.8 |
| | | MD | 49.5 | 52.6 | 53.3 | 51.5 | 52.2 | 51.3 |
| | 500 | None | 24.1 | 28.4 | 32.4 | 31.2 | 31.7 | 33.9 |
| | | Block | 53.9 | 49.2 | 49.2 | 44.5 | 50.1 | 45.0 |
| | | MD | 70.7 | 75.3 | 75.3 | 74.0 | 74.7 | 74.7 |
| | 1000 | None | 33.6 | 40.0 | 46.8 | 48.3 | 49.1 | 53.0 |
| | | Block | 52.6 | 59.6 | 65.0 | 64.2 | 65.0 | 66.0 |
| | | MD | 92.8 | 95.1 | 94.5 | 95.3 | 95.2 | 95.8 |

# Historical Simulation

- Historical simulation (HS) is a nonparametric method based on the empirical distribution function of historical risk-factor changes (returns) affecting P&L.
- Perignon and Smith (2010) report that 73% of US and international banks use historical simulation for $\widehat{F}_t$.
- *Filtered HS* (FHS) as refinement: first apply EWMA volatility estimation to risk-factor changes, then use edf of volatility-filtered returns.
- Assuming stationarity of risk-factors, PIT values are unconditionally $U[0,1]$, but $P_t$ very likely to be serially dependent under HS.
- Experiment: generate samples from a GJR-GARCH(1,1) model with skew t innovations.
  - Many features of a real financial return series: SV, leverage effect, skewed and heavy-tailed shocks.
  - Fit to historical sample of S&P log-returns.
  - The EWMA volatility filter can capture GARCH(1,1) volatility effect but cannot capture the leverage effect.

# Reported PIT-values derived from HS/FHS




ACF plots of PIT-values ($P_t$) and transformed PIT-values ($|2P_t - 1|$) obtained from HS and FHS applied to S&P 500 returns (250-day window).

# Rejection rates for Historical Simulation

| $m$ | $n$ | ITT \| test | B99 | ZU | ZE | ZeE | ZUE | PNS |
|---|---|---|---|---|---|---|---|---|
| 250 | 250 | None | 30.3 | 32.1 | 33.4 | 31.6 | 31.6 | 32.7 |
| | | Block | 39.2 | 41.0 | 39.6 | 42.1 | 41.0 | 40.9 |
| | | MD | 44.2 | 46.2 | 46.5 | 45.7 | 46.2 | 46.4 |
| | 500 | None | 35.5 | 41.4 | 44.0 | 39.4 | 39.8 | 41.9 |
| | | Block | 63.3 | 59.4 | 58.4 | 54.7 | 58.7 | 56.0 |
| | | MD | 64.0 | 67.0 | 67.3 | 66.1 | 66.3 | 66.6 |
| | 1000 | None | 51.9 | 61.4 | 65.6 | 59.3 | 59.8 | 63.4 |
| | | Block | 68.3 | 76.5 | 79.1 | 76.7 | 77.0 | 78.0 |
| | | MD | 84.6 | 87.8 | 88.3 | 87.7 | 87.8 | 88.3 |
| 500 | 250 | None | 24.8 | 25.6 | 26.1 | 25.6 | 25.4 | 25.5 |
| | | Block | 31.7 | 32.2 | 30.4 | 33.6 | 32.1 | 31.6 |
| | | MD | 34.1 | 35.2 | 35.2 | 34.9 | 35.1 | 35.2 |
| | 500 | None | 31.7 | 36.8 | 36.3 | 31.5 | 31.6 | 31.8 |
| | | Block | 52.8 | 50.3 | 46.5 | 43.5 | 46.8 | 44.3 |
| | | MD | 50.8 | 53.2 | 52.9 | 52.7 | 52.5 | 52.8 |
| | 1000 | None | 32.1 | 37.4 | 38.8 | 34.5 | 34.8 | 36.4 |
| | | Block | 57.9 | 64.1 | 65.3 | 63.1 | 63.1 | 63.0 |
| | | MD | 75.6 | 78.6 | 77.9 | 77.8 | 77.5 | 78.5 |

# Rejection rates for Filtered Historical Simulation

| $m$ | $n$ | ITT \| test | B99 | ZU | ZE | ZeE | ZUE | PNS |
|---|---|---|---|---|---|---|---|---|
| 250 | 250 | None | 8.6 | 9.3 | 10.6 | 9.3 | 9.5 | 10.3 |
| | | Block | 17.2 | 17.5 | 16.1 | 19.0 | 17.7 | 16.9 |
| | | MD | 20.2 | 21.7 | 22.7 | 20.7 | 21.4 | 20.4 |
| | 500 | None | 6.5 | 8.9 | 11.0 | 8.5 | 8.9 | 10.3 |
| | | Block | 31.3 | 23.1 | 22.8 | 16.4 | 22.7 | 18.5 |
| | | MD | 21.0 | 23.1 | 25.6 | 20.9 | 22.4 | 20.9 |
| | 1000 | None | 9.3 | 14.8 | 19.6 | 12.9 | 13.4 | 17.2 |
| | | Block | 17.8 | 24.3 | 30.8 | 24.4 | 25.0 | 25.5 |
| | | MD | 24.5 | 27.9 | 31.3 | 25.5 | 27.5 | 26.9 |
| 500 | 250 | None | 6.4 | 6.4 | 7.1 | 7.1 | 7.2 | 7.1 |
| | | Block | 14.5 | 13.7 | 12.1 | 15.5 | 14.1 | 12.7 |
| | | MD | 15.4 | 15.8 | 16.9 | 15.0 | 16.1 | 15.1 |
| | 500 | None | 3.8 | 4.2 | 5.0 | 4.3 | 4.3 | 4.6 |
| | | Block | 25.0 | 16.8 | 16.1 | 10.8 | 15.9 | 11.7 |
| | | MD | 14.8 | 15.6 | 17.0 | 13.3 | 14.6 | 13.2 |
| | 1000 | None | 1.8 | 2.3 | 3.4 | 2.4 | 2.4 | 2.8 |
| | | Block | 11.4 | 14.3 | 16.2 | 13.5 | 13.4 | 12.4 |
| | | MD | 14.3 | 15.4 | 16.8 | 12.8 | 14.0 | 12.7 |

# Summary

- Tests based on spectral transformations of *reported PIT-values* can yield more power than simple VaR exception tests.
- The spectral class of backtests provides a *unifying framework* encompassing many widely-used backtests.
  - VaR exception tests (Kupiec), probitnormal LTR (Berkowitz), DQ (Engle-Manganelli).
  - Among widely-used backtests, only duration-based tests (Christoffersen and Pelletier, 2004) fall outside the spectral class.
- Expressing tests in this form facilitates construction of new tests and encourages thinking about the implied kernel.
- The tests are available in *unconditional and conditional* variants. Conditional tests can be based on *blocking* or testing the *martingale difference property*.
- Spectral backtests increase rejection rates for HS/FHS procedures.
  - Backtests of length $n = 500$ (2 years) would be preferable to $n = 250$.
- Application to proprietary bank-reported data is in progress.