Spectral Backtests of Forecast Distributions

with Application to Risk Management

Alexander J. McNeil
The York Management School, University of York

Joint work with Michael Gordy and Yen H. Lok

Research Seminar in Statistics and Mathematics
Vienna University of Economics and Business
30th November 2018

Overview

Setting for backtesting

- A firm makes a one-step-ahead forecast of its loss distribution.
- Notation used throughout:

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

- In banking regulation backtesting is based on VaR exceedances.
  - $\widehat{\text{VaR}}_{\alpha,t} := \widehat{F}_t^{\leftarrow}(\alpha)$ is an estimate of α-VaR constructed at time $t-1$.
  - Bank reports $\widehat{\text{VaR}}_{\alpha,t}$ and realized L_t.
  - VaR exceedance is simply the indicator $I_t = I_{\{L_t > \widehat{\text{VaR}}_{\alpha,t}\}}$.

Probability integral transform

- Increasingly, regulators observe more than just 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 α.

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

- The ideal forecaster. If the $(\widehat{F}_t)$ coincide with the true (F_t), then the process (P_t) is iid $U[0,1]$ (Rosenblatt, 1952).
- 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

Priorities for model performance

- Diebold, Gunther, and Tay (1998) develop forecast density tests based on testing PIT values for iid $U[0,1]$. See also Blum (2005).
- 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 be expected to perform well in the extreme tail of once-per-generation shock events.
- We study a class of backtests for forecast distributions in which the test statistic weights exceedances by a function of the probability level α.
- The kernel function makes explicit the priorities for model performance.

Regulatory background

- In banking FRTB replaces 99%-VaR with 97.5% expected shortfall (ES) as determinant of capital requirements.
- This has led to debate whether ES is amenable to direct backtesting.
- The model approval process continues to be based on VaR exceedances.
- In insurance the 99.5%-VaR of annual loss distribution is target risk measure.
- We devise tests of the forecast distribution from which risk measures are estimated and not tests of the risk measure estimates.
- For purposes of exposition we focus on testing forecast distributions designed to yield estimates of 99%-VaR.

Spectral transformations

- Our tests are based on transformations of the indicator function for PIT exceedances and are termed "spectral" in the integral transform sense.
- The transformations take the form

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

  where ν is a Lebesgue-Stieltjes measure on $[0, 1]$.
- W_t increases in P_t.
- ν is chosen to apply weight to different levels in the unit interval, typically in the region of the VaR level $\alpha = 0.99$.
- We refer to ν as the kernel measure for the transform.
- The support of the measure describes subsets of $[0, 1]$ that are weighted.
- Note that

$$\mathbb{E}(W_t) = \int_0^1 (1 - u)\mathrm{d}\nu(u) \quad .$$

Spectral backtests

- Univariate spectral backtests are backtests based on $W_1, \ldots, W_n$.
- Multivariate tests based on $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ where $\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,J})'$ and $W_{t,j} = \nu_j([0, P_t])$ for distinct measures $\nu_1, \ldots, \nu_J$.
- Null hypothesis. Let F_W^0 denote df of $\boldsymbol{W}_t$ when P_t is uniform.

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

- Within the class of spectral backtests, we have tests of unconditional and conditional coverage.
- Unconditional coverage: test for correct distribution F_W^0;
- Conditional coverage: correct distribution and independence from $\mathcal{F}_{t-1}$.

Useful product result

Theorem (Gordy-McNeil)

The set of spectrally transformed PIT values defined by $W_{t,i} = \int_0^1 I_{\{P_t \geq u\}} \mathrm{d}\nu_i(u)$ *is closed under multiplication and*

$$W_{t,1} W_{t,2} = \int_0^1 I_{\{P_t \geq u\}} \mathrm{d}\nu^*(u)$$

for a measure ν^* *which satisfies*

$$\mathrm{d}\nu^*(u) = \left(\nu_2([0,u]) - \frac{1}{2}\nu_2(\{u\})\right)\mathrm{d}\nu_1(u) + \left(\nu_1([0,u]) - \frac{1}{2}\nu_1(\{u\})\right)\mathrm{d}\nu_2(u) \,.$$

- Integration by parts for Lebesgues-Stieltjes measures.
- For measures we consider, explicit forms for ν^* are available.
- Hence can calculate $\mathbb{E}(W_{t,1} W_{t,2})$ or $\mathbb{E}(W_t^2)$.

First test type: Z-test

- Univariate 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 Z-tests are based on

$$T_n = n\left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right)' \Sigma_W^{-1} \left(\overline{\boldsymbol{W}}_n - \boldsymbol{\mu}_W\right) \xrightarrow[n\to\infty]{d} \chi_J^2.$$

Second test type: likelihood ratio test

- LR-tests employ continuous parametric family $F_P(p \mid \theta)$ for P_t which nests uniformity for $\theta = \theta_0$.
- Examples are the probitnormal and beta distributions.
- Let $F_W(w \mid \theta)$ denote implied distribution of W_t so that $F_W^0 = F_W(\cdot \mid \theta_0)$.
- Test is based on the asymptotic chi-squared distribution of the statistic

$$\text{LR}_{W,n} = \frac{L_W(\theta_0 \mid \boldsymbol{W})}{L_W(\hat{\theta} \mid \boldsymbol{W})}$$

  where $\hat{\theta}$ denotes the maximum likelihood estimate (MLE).
- LR-test requires estimation of $\hat{\theta}$ under the alternative.

Dirac kernel

- A Dirac kernal $\nu = \delta_\alpha$ (point mass) yields $W_t = I_{\{P_t \geq \alpha\}}$, the α-VaR exceedance indicator.
- The (W_t) are iid Bernoulli$(1-\alpha)$ under H_0.
- This case corresponds to standard VaR backtesting.
- The Z-test is the classical binomial score test.
- The LR-test in this context was proposed by Kupiec (1995) and Christoffersen (1998) and is very widely applied in practice.
- Kratz, Lok, and McNeil (2016) show that the score test has best performance in typical regulatory context.

Discrete kernels

- Univariate
  - A general discrete kernal $\nu = \sum_{i=1}^{m} k_i \delta_{\alpha_i}$ yields $W_t = \sum_{i=1}^{m} k_i I_{\{P_t \geq \alpha_i\}}$.
  - W_t satisfies
$$\mathbb{P}(W_t = q_i) = \alpha_{i+1} - \alpha_j = i, \quad i = 0, \ldots, m \tag{2}$$
where $q_i = \sum_{j=1}^{j} k_j$, $q_0 = 0$, $\alpha_0 = 1$ and $\alpha_{m+1} = 1$.
  - The Z-test is a new test which allows user to vary the weights k_i.
  - The LR-test nests the distribution (2) in a general multinomial model: $\mathbb{P}(W_t = q_i) = \theta_i$, $\sum_{i=0}^{m} \theta_i = 1$. The cell counts $O_i = \sum_{t=1}^{n} I_{\{W_t = q_i\}}$ are sufficient statistics so actual values of k_i play no role. Test proposed by Pérignon and Smith (2010) and Colletaz, Hurlin, and Perignon (2013).
- Multivariate
  - A set of m distinct Dirac kernels $\nu_1 = \delta_{\alpha_1}, \ldots, \nu_m = \delta_{\alpha_m}$ yields multivariate tests based on $\boldsymbol{W}_t = (I_{\{P_t \geq \alpha_1\}}, \ldots, I_{\{P_t \geq \alpha_m\}})'$.
  - The Z-test is identical to Pearson's chi-squared test. This test has been proposed by Campbell (2007).
  - The LR-test depends on count variables $O_i = \sum_{t=1}^{n} I_{\{\mathbf{1}' \boldsymbol{W}_t = i\}}$ and coincides with LR-test in univariate case.

Continuous kernels

- A continuous kernel measure has density $\mathrm{d}\nu(u) = g(u)\mathrm{d}u$ for some non-negative function g on $[0, 1]$.
- We study measures with support given by a window $[\alpha_1, \alpha_2] \subset [0, 1]$.
- Univariate Z-tests considered by Costanzino and Curran (2015) and Du and Escanciano (2017) (with focus on uniform kernel $g(u) = I_{\{\alpha_1 \leq u \leq \alpha_2\}}$).
- We are particularly interested in bispectral Z-tests using two measures on same support.
- The LR-test requires a parametric family $F_P(p \mid \theta)$ for P_t which nests uniformity for $\theta = \theta_0$.
- By taking the probitnormal family $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$ with $\boldsymbol{\theta}_0 = (0, 1)$ we obtain generalization of an LR-test proposed by Berkowitz (2001).

Mixed kernel

- Suppose $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$ and let $P_t^* = \alpha_1 \vee (P_t \wedge \alpha_2)$.
- The likelihood function for P_t^* has closed-form expression.
- Denote the observed score vector for P_t^* by

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

- This can be written as vector of spectrally-transformed PIT values.
- The two measures ν_1 and ν_2 have continuous and discrete parts which can be written explicitly.
- Thus probitnormal score test of $(\mu, \sigma^2) = (0, 1)$ yields a bispectral Z-test.

Description of study

- Concentrate on continuous kernels which give more stable results.
- Sample data L_t from 4 distributions F (all mean zero, variance 1):

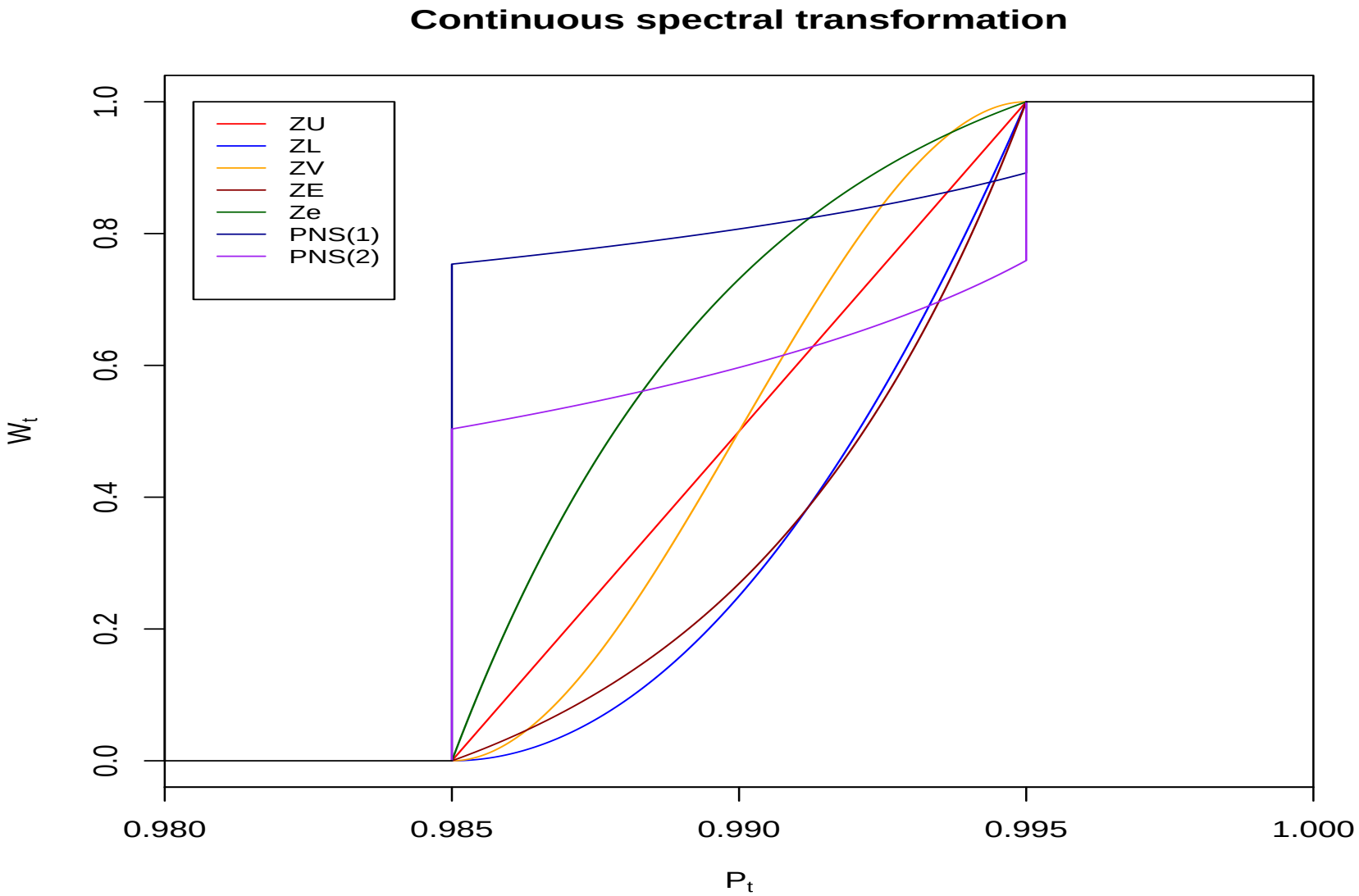
F	$\text{VaR}_{0.975}$	$\text{VaR}_{0.99}$	Δ_1	$\text{ES}_{0.975}$	Δ_2
Normal	1.96	2.33	0.00	2.34	0.00
t5	1.99	2.61	12.04	2.73	16.68
t3	1.84	2.62	12.69	2.91	24.46
st3 ($\gamma = 1.2$)	2.04	2.99	28.68	3.35	43.11

- 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 typical for tail underestimation.

Choice of kernels

- Kernel window is $[0.985, 0.995]$.
- Kernel densities
  - B99: Dirac kernel at $\alpha = 0.99$ (benchmark).
  - Ze: Exponential kernel, decreasing $g(u) = \exp\left(\kappa\left(\frac{u-\alpha_1}{\alpha_2-\alpha_1}\right)\right)$, $\kappa = -2$.
  - ZE: Exponential kernel, increasing $\kappa = 2$.
  - ZV: Epanechnikov kernel - humped, symmetric around $\alpha = 0.99$.
  - ZU: Uniform kernel.
  - ZL: Linear kernel $g(u) = u - \alpha_1$
- Continuous bispectral tests: ZeE, ZUE, ZLE.
- Probitnormal (Berkowitz) LR-test (PNL) and Z-test (PNS).

Normalized kernel measures

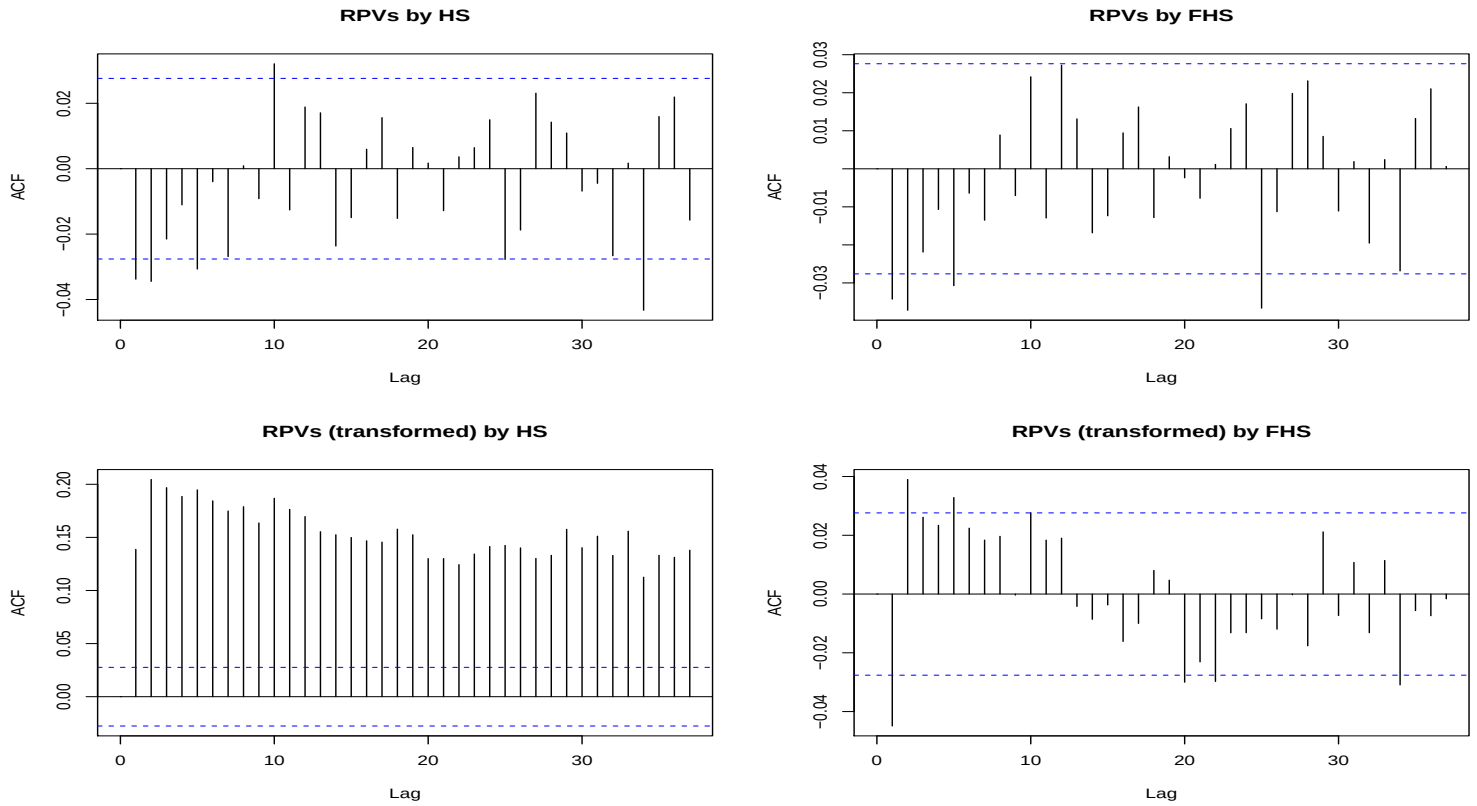


Size and power of tests

F	n | test	B99	Ze	ZV	ZU	ZE	ZeE	ZUE	ZLE	PNL	PNS
Normal	250	4.0	3.7	3.8	4.0	4.1	5.0	5.0	5.1	7.0	5.0
	500	3.7	4.6	4.3	4.7	4.7	4.6	4.7	4.7	5.6	4.5
	1000	3.8	5.1	5.2	5.4	5.3	4.8	4.9	4.9	5.3	4.8
t5	250	17.7	15.9	18.2	19.2	22.4	21.1	21.3	17.4	17.8	22.8
	500	22.4	22.7	25.8	26.9	31.5	30.5	31.1	25.6	27.1	33.5
	1000	33.0	32.8	38.6	39.9	47.4	49.4	50.0	39.2	47.1	53.8
t3	250	13.5	11.0	13.7	14.5	19.6	21.2	21.5	13.2	23.0	23.3
	500	16.2	14.4	18.9	20.1	26.5	32.3	33.0	20.2	35.8	37.2
	1000	22.3	19.7	26.9	27.9	38.7	55.1	56.2	31.1	61.4	62.1
st3	250	31.2	27.8	31.9	33.3	39.5	39.9	40.1	30.7	35.8	42.7
	500	44.2	41.6	48.4	50.3	58.0	59.6	60.3	49.9	57.3	64.0
	1000	66.2	64.0	72.2	73.5	81.7	86.1	86.6	74.9	86.0	89.2

Green indicates good results ($\leq$ 6% for the size; $\geq$ 70% for the power); pink indicates poor results ($\geq$ 9% for the size; $\leq$ 30% for the power).

Unmodelled volatility and PIT values



ACF plots of PIT-values (P_t) and transformed PIT-values ($|2P_t - 1|$). In left pictures volatility of returns is not explicitly modelled.

Martingale difference tests

- Let $\tilde{W}_t = W_t - \mu_W$ for all t. Under H_0 the martingale difference (MD) property hold: $\mathbb{E}(\tilde{W}_t \mid \mathcal{F}_{t-1}) = 0$.
- For any $\mathcal{F}_{t-1}$-measurable $\boldsymbol{h}_{t-1}$ vector this implies $\mathbb{E}(\boldsymbol{h}_{t-1}\tilde{W}_t) = \mathbf{0}$.
- We consider $\boldsymbol{h}_{t-1} = (1, h(P_{t-1}), \ldots, h(P_{t-k}))'$ for some choice of h.
- Let $\boldsymbol{Y}_t = \boldsymbol{h}_{t-1}\tilde{W}_t$ for $t = k+1, \ldots, n$. Let $\overline{\boldsymbol{Y}} = (n-k)^{-1}\sum_{t=k+1}^{n}\boldsymbol{Y}_t$ and let $\hat{\Sigma}_Y$ denote a consistent estimator of $\Sigma_Y := \mathrm{cov}(\boldsymbol{Y}_t)$.
- Giacomini and White (2006) show that under very weak assumptions, for large enough n and fixed k,

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

Martingale difference tests (II)

- Generalizes the dynamic quantile test of Engle and Manganelli (2004) which corresponds to $\nu = \delta_\alpha$ and $h(p) = I\{p \geq \alpha\}$.
- Case $k = 0$ is ordinary spectral Z-test.
- Case $k = 1$ is an analog of Markov chain LR-test of Christoffersen (1998).
- Martingale-difference extensions of the bispectral Z-test (including probitnormal score test) are also available.
- We choose $h(p) = (|2p - 1|)^c$ for $c > 0$ to target unmodelled stochastic volatility.

Design of experiment

- Idea is to capture behaviour of PIT values when DGP F_t features stochastic volatility, but this is ignored in the firm's model $\widehat{F}_t$.
- Create sequence of uniform rvs (U_t) such that $(|2U_t - 1|)$ follows an ARMA 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 dependent but uniform.
- When $F \neq \Phi$, PIT-values are dependent and non-uniform.

Power of tests

F	n	MD? | test	B99	ZU	ZE	ZeE	ZUE	PNS
Normal	250	No	5.6	5.3	5.3	6.4	6.3	6.3
		Yes	27.6	29.5	28.0	29.6	28.7	30.3
	500	No	5.9	7.4	7.0	6.3	6.3	6.1
		Yes	42.1	46.2	42.5	46.6	44.4	49.7
	1000	No	5.8	7.7	7.5	6.8	6.8	6.8
		Yes	70.3	77.5	71.5	78.2	74.5	82.0
t5	250	No	18.8	20.2	22.7	21.9	22.0	23.3
		Yes	49.5	52.6	53.3	51.5	52.2	51.3
	500	No	24.1	28.4	32.4	31.2	31.7	33.9
		Yes	70.7	75.3	75.3	74.0	74.7	74.7
	1000	No	33.6	40.0	46.8	48.3	49.1	53.0
		Yes	92.8	95.1	94.5	95.3	95.2	95.8

Green indicates good results ($\geq 70\%$); pink indicates poor results ($\leq 30\%$); red indicates terrible results ($\leq 10\%$).

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.
- 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.
- The conditional tests are particularly powerful at revealing dependencies in the PIT data caused by unmodelled stochastic volatility.
- The tests have been applied to proprietary bank-reported data and results are available in Gordy and McNeil (2018).

For Further Reading

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

Blum, P., 2005, *On Some Mathematical Aspects of Dynamic Financial Analysis*, Ph.D. thesis, ETH Zurich.

Campbell, S., 2007, A review of backtesting and backtesting procedures, *Journal of Risk* 9, 1–18.

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

Colletaz, G., C. Hurlin, and C. Perignon, 2013, The risk map: a new tool for validating risk models, *Journal of Banking and Finance* 37, 3843–3854.

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

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.

For Further Reading (cont.)

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.

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

Gordy, M.B., and A.J. McNeil, 2018, Spectral backtests of forecast distributions with applications to risk management, arXiv:1708.01489.

Kratz, M., Y.H. Lok, and A.J. McNeil, 2016, Multinomial VaR backtests: a simple implicit approach to backtesting expected shortfall, Technical report, ESSEC Working Paper 1617 & arXiv1611.04851v1.

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

For Further Reading (cont.)

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.

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