Backtesting with correlated data

Nikolai Nowaczyk, Vladimir V. Piterbarg

Frontiers in Quantitative Finance

24/10/2024 London

The views expressed in this presentation are those of the authors alone.

mail@nikno.de
https://uk.linkedin.com/in/niknow
https://github.com/niknow

Content

Backtesting

Use Cases

- Front Office / Market Risk:
  - returns/risks of trading strategies
  - market risk metrics (e.g. VaR)
  - margin models (e.g. SIMM)
- Counterparty Credit Risk (CCR): EAD
  - risk factor evolution
  - portfolio MtMs

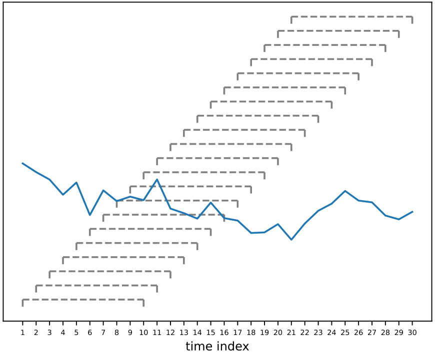
Challenges

- Data scarcity and quality
- Computational intensity
- Legacy infrastructures
- Complex statistical evaluations
  - Which test to choose?
  - Which test is "better"?
  - ...
  - **How to deal (best) with correlations?**

Backtesting

- **Statistical theory** typically starts with:
  Let x_i, $i = 1, \ldots, n$, be the independent samples...
- **Statistical reality** in finance typically starts with the insight that this basic assumption is not met due to
  - **auto-correlation** within a single time series whenever the samples correspond to overlapping returns.
  - **cross-correlation** between any two quantites (e.g. IR/FX).
- **All typical applications are affected**, e.g. CCR backtesting, SIMM backtesting etc.
- **Ignoring the correlations leads to materially incorrect results.**

Example of Auto-correlation



Samples: Given $n = 250$ daily independent time series returns $Y_i \sim \mathcal{N}(0, \sigma^2)$, we consider the $m = 10$-day returns X

$$X_i := \sum_{j=i}^{i+m-1} Y_j \sim \mathcal{N}(0, m\sigma^2),$$

$i = 1, \ldots, n - m + 1$, which slide forward by 1-day.

$\Longrightarrow$ Obtain $n_m := 241$ samples, but with up to 90% correlation.

Content

Content

Hypothesis Test: Ingredients

1. Formulate *null hypothesis:*
   The model's predictive distributions are consistent with market realizations.
2. Collect *sample* $\hat{x}$ from a *sample space* $\mathfrak{X} = (\mathfrak{X}, \mathcal{F}, \mathbb{P}_\vartheta)_{\theta \in \Theta}$.
3. Split $\Theta = \Theta_0 \dot{\cup} \Theta_1$: We call Θ_0 *null hypothesis* and Θ_1 is called *alternative*.
4. Choose a *significance level* α, e.g. $\alpha = 5\%$.
5. Choose a *test statistic* $T : \mathfrak{X} \to \mathbb{R}$ and a critical value $t_{\mathrm{crit}} = Q_{1-\alpha}(T)$.
   This requires the distribution of the test statistic T under the null hypothesis.
6. A *decision rule* $\varphi : (\mathfrak{X}, \mathcal{F}) \to \{0, 1\}$, e.g. for upper-tailed test

$$\varphi(\hat{x}) = \begin{cases} 1, & T(\hat{x}) > t_{\mathrm{crit}} \implies \text{reject null hypothesis} \\ 0, & T(\hat{x}) \leq t_{\mathrm{crit}} \implies \text{retain null hypothesis} \end{cases}$$

Example:

- **Hypothesis**: We want to test the null hypothesis

$$H_0 : \sigma \leq \sigma_0 := 100 \qquad \text{against} \qquad H_1 : \sigma > \sigma_0$$

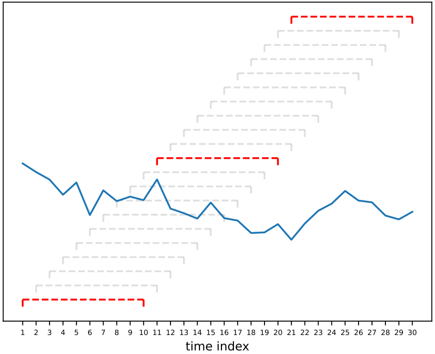
- **Test statistic definition**: *Exceedence counting* at quantile level $\gamma := 95\%$

$$T := \sum_{j=1}^{n_m} 1_{\{X_j > h\}}, \qquad h := Q_\gamma(\mathcal{N}(0, m\sigma_0^2)) = \sigma_0 \sqrt{m} \Phi^{-1}(\gamma).$$

- **Test statistic** T does **not** have a Binomial distribution under null hypothesis due to correlation in the data.

Content

Strategy 1: Filtering



Throw away the correlated samples, i.e. only work with the 25 independent samples

$$X_{mi}, \qquad i = 1, \ldots, n/m.$$

Then

$$T_0 := \sum_{j=1}^{n/m} 1_{\{X_{mj} > h\}},$$

has Binomial distribution $\mathrm{Bin}_{n/m}(1-\gamma)$.

Strategy 2: Correlate null distributions via Monte Carlo Simulation

- Generate Monte Carlo paths of the samples

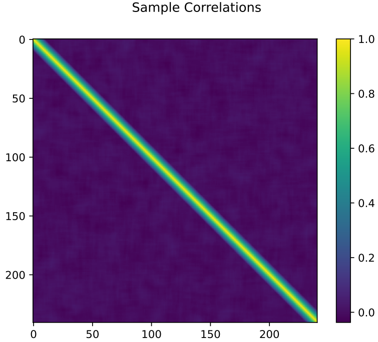
$$Y_i(\omega) \sim \mathcal{N}(0, \sigma_0^2), \qquad X_i(\omega) := \sum_{j=i}^{i+m-1} Y_j(\omega)$$

- and the test statistic via

$$T(\omega) := \sum_{j=1}^{n_m} 1_{\{X_j(\omega) > h\}},$$

and calculate the quantile $t_{\text{crit}} := Q_{1-\alpha}(T)$ empirically.

Strategy 3: Decorrelation



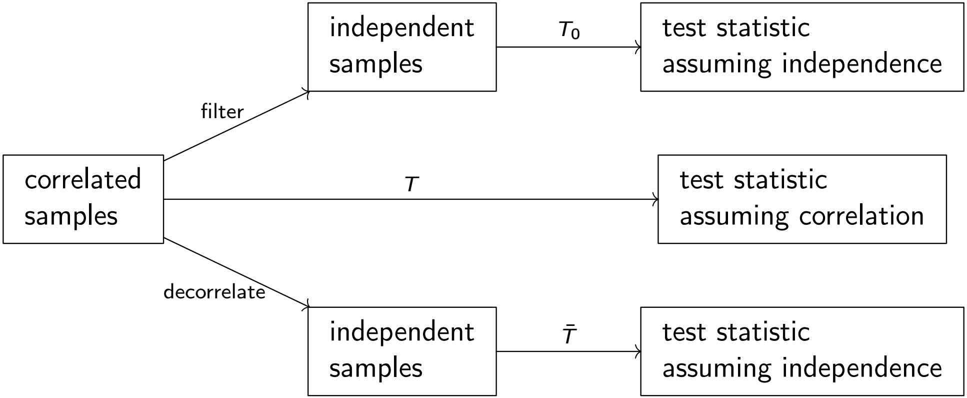
Under null hypothesis, correlation matrix C of sample vector $X = (X_1, \dots, X_{n_m})$ is known (in this case analytically).
Hence:

- Compute Cholesky decomposition $C = LL^\top$
- Decorrelate samples to $\bar{X} := L^{-1}X$.
- $\implies$ Exceedence count statistic

$$\bar{T} := \sum_{j=1}^{n_m} 1_{\{\bar{X}_j > h\}},$$

now has Binomial distribution with sample size n_m and success probability $1 - \gamma$.

How to decide which strategy is best?



Content

Hypothesis Test: Evaluation

		Test Result	
		retain H_0	reject H_0
Assumption	H_0	correct retention	incorrect rejection (α), type I
	H_1	incorrect retention (β), type II	correct rejection

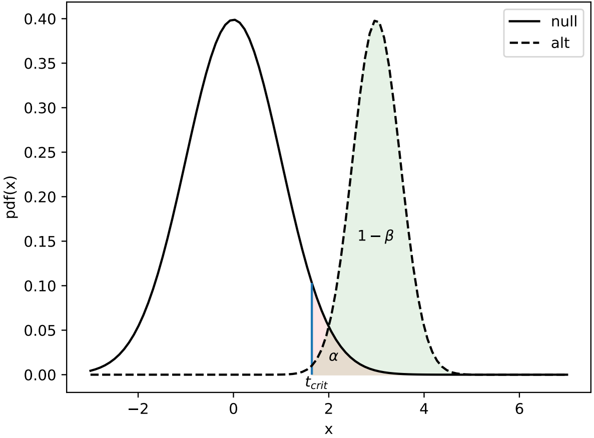
Using the *Discriminatory Power* function:

$$G_\varphi : \Theta \to [0,1], \qquad \vartheta \mapsto \mathbb{P}_\vartheta[\{\varphi = 1\}] = 1 - \mathbb{P}_\vartheta[T \leq t_{\text{crit}}]$$

we obtain for

- Type I: $\mathbb{P}_{\vartheta_0}[\{\varphi = 1\}] \leq \alpha$ by construction ($\Longrightarrow$ no choice)
- Type II: $\beta_\varphi(\vartheta_1) = 1 - G_\varphi(\vartheta_1)$ ($\Longrightarrow$ **natural metric to optimize**)

Visualizing Type I & Type II error



Which Alternative should we choose?

- In our case the **null hypothesis pertains to one probability measure** $\mathbb{P}_{\vartheta_0}$ given by the model.
- Notice that the **power** of testing the null hypothesis ϑ_0 against an alternative ϑ_1 **depends on the alternative**. What alternative should we choose?
- **Theoretically, every other probability measure** $\mathbb{P}_{\vartheta_1}$ could be an alternative.
- **Practically** evaluating this is not really feasible.
- **Pragmatic** approach (common in empirical research, medicine, psychology etc.) is to assess the power on a 1-parameter family of interesting alternatives.

Optimizing Hypothesis Tests

We evaluate the strategies

- filtering,
- correlating the test statistic,
- decorrelating the samples,

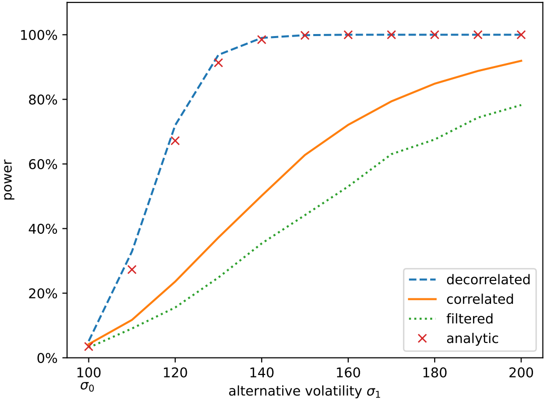
by

- constructing prototypical hypothesis tests,
- calculating power curves for pragmatic family of alternatives,
- compare the results.

Content

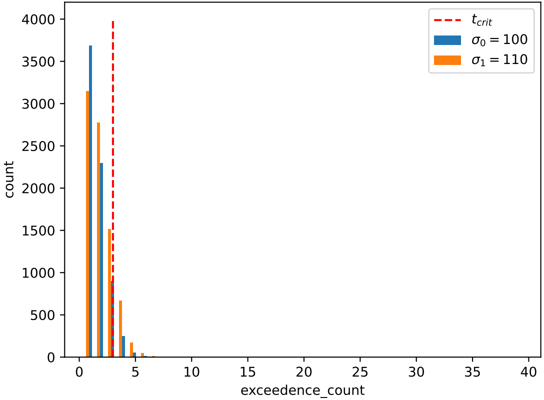
Content

Impact of correlation strategy on power



Why does this work?

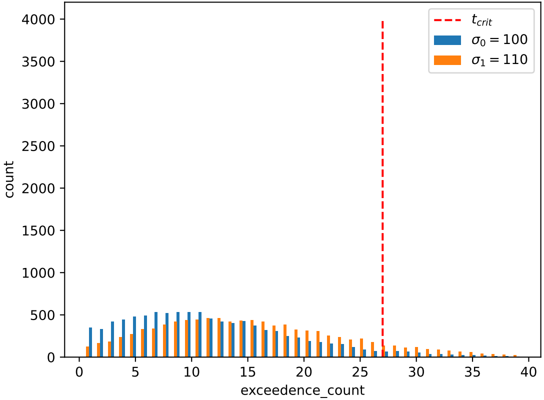
PDF of Null Hypothesis vs. Alternative: Filtered



Test parameters

- $n = 250$ days of returns
- $m = 10$ window size

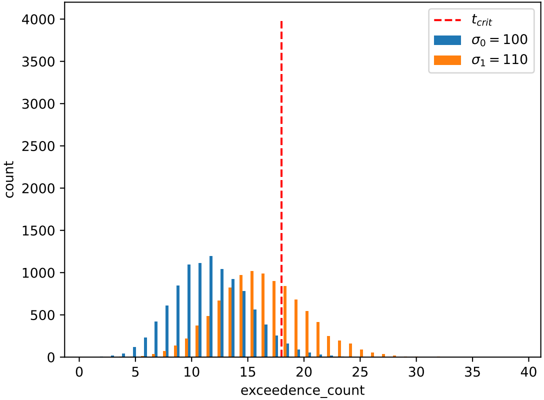
PDF of Null Hypothesis vs. Alternative: Correlated



Test parameters

- $n = 250$ days of returns
- $m = 10$ window size

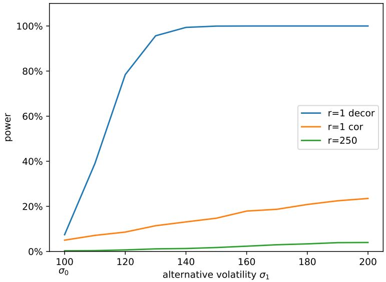
PDF of Null Hypothesis vs. Alternative: Decorrelated



Test parameters

- $n = 250$ days of returns
- $m = 10$ window size

Extreme example of parameters



We test

- using $n = 2Y$ of history
- with a window size of $m = 1Y$

in the three set ups:

- $r = 250$: 0.5% power (2 samples)
- $r = 1$ (correlated samples): 7.7% power (251 samples)
- $r = 1$ (decorrelated samples): 77% power (251 samples)

Content

A two-sided alternative & Chi Squared test

- Hypothesis: We want to test the null hypothesis

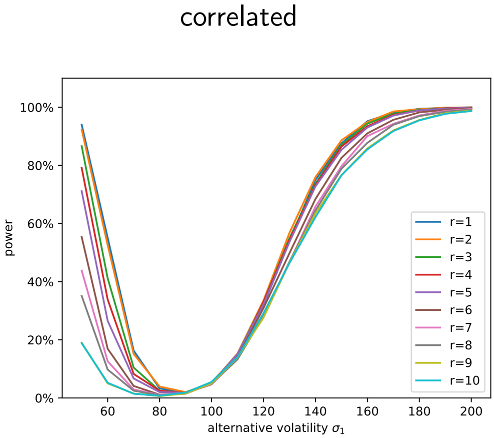
$$H_0 : \sigma = \sigma_0 \qquad \text{against} \qquad H_1 : \sigma \neq \sigma_0$$

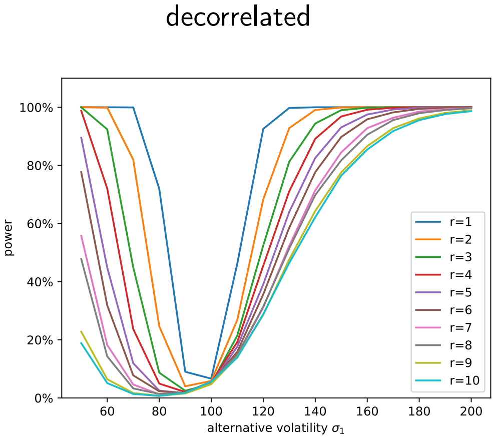
- Test statistic: Choose some quantile level grid of length k, say $\gamma = \{0\%, 1\%, 5\%, 20\%, 50\%, 80\%, 95\%, 99\%, 100\%\}$, construct the associated thresholds $h_j := Q_{\gamma_k}(\mathcal{N}(0, m\sigma_0^2))$ and for each bin $[h_{j-1}, h_j[$, compare the observed samples o_j in the bin with the expected samples $e_j = n_r(\gamma_j - \gamma_{j+1})$ via the *chi squared* as test statistic:

$$\chi_r^2 := \sum_{j=1}^{k} \frac{(e_j - o_j)^2}{e_j}.$$

For $r = m$, χ_r^2 is asymptotically distributed as $\chi^2(k-1)$. But for $r < m$, this distribution needs to be estimated via Monte Carlo simulation.

Impact of step size: Power

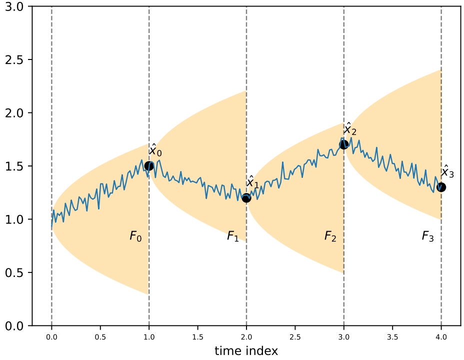




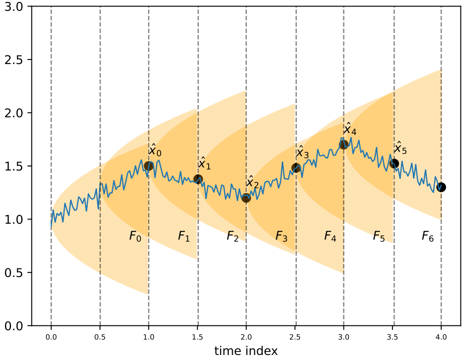
Content

CCR backtesting

non-overlapping



overlapping



Test Setup and Null Hypothesis

We assume the quantity to be backtested is given by

$$dX_t = \sigma dW_t \qquad X_t = X_0 + \sigma\sqrt{t}Z,\ Z \sim \mathcal{N}(0,1)$$

No recalibration of σ, but initialization of start value.

- Hypothesis: We want to test the null hypothesis

$$H_0 : \sigma = \sigma_0 \qquad \text{against} \qquad H_1 : \sigma \neq \sigma_0$$

- Simulation setup:
  - Fix backtesting date grid $t_1 < \ldots < t_n$ of width e.g. $\delta = 2W$ over observation window, e.g. 5Y
  - Fix horizon, e.g. $\tau = 1Y$ and generate simulations $X_i := X(t_i, t_i + \tau)$ with N_{sim} paths
  - Obtain their distribution $\hat{F}_i$ (does not need simulation in this case)
  - For any given sample $\hat{x}$ test if the resulting *PITs* $\pi_i := \hat{F}_i(\hat{x}_i)$ are *uniform*

Probability Integral Transform to the Uniform

A key trick is to make the statistical framework independent of the underlying distribution via the following.

Lemma

Let X be a real valued random variable with continuous CDF F. Then $F(X)$ is uniformly distributed on $[0, 1]$.

Definition

For any sample $\hat{x}$ of X, we call $\pi(\hat{x}) := F(\hat{x})$ the *probability integral transform (PIT) of $\hat{x}$ with respect of F*.

$\Longrightarrow$ We can work with $\pi_i := F_i(\hat{x}_i)$ where F_i is the CDF of $X(t_i, T_i)$ and test for uniformity *if* π_i are independent.

Uniformity metrics

- Exceedence counting over some quantile
- χ^2 with some binning
- Cramer-von-Mises metric (CvM):

$$\int_{\mathbb{R}} |F(x) - \hat{F}(x)|^2 dF(x)$$

- Anderson-Darling (AD)

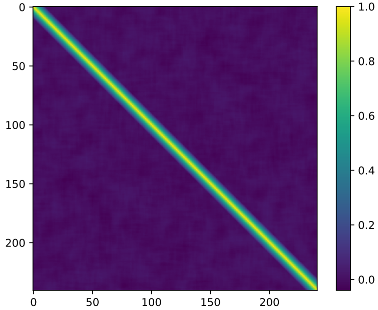
$$\int_{\mathbb{R}} |F(x) - \hat{F}(x)|^2 w(x) dF(x), \qquad w(x) = \frac{1}{F(x)(1 - F(x))}$$

- Kolmogorov-Smirnoff (KS)

$$\sup_{x \in \mathbb{R}} |F(x) - \hat{F}(x)|$$

Here $\hat{F}$ is an estimated ECDF and $F(x) = x$ is the CDF of $\mathcal{U}(0,1)$.

Decorrelation of uniformly distributed PITs



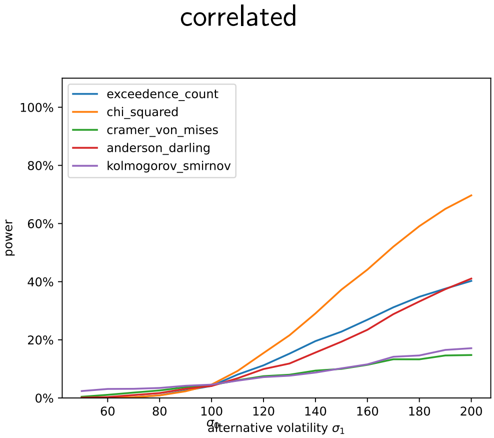
Under the null hypothesis, the correlation matrix C of the pits is known as well.
Hence:

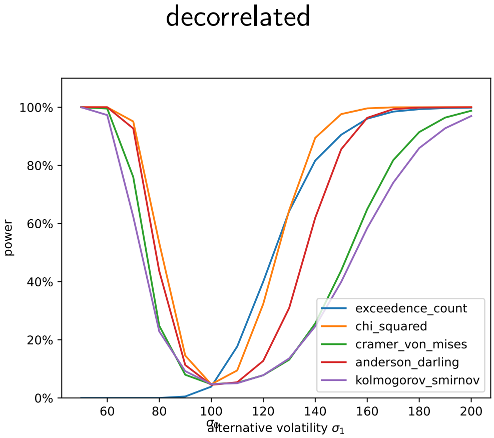
- Compute Cholesky decomposition $C = LL^\top$
- PITs are on $[0, 1]$ and hence L cannot be applied directly.
- Hence, decorrelate samples via

$$\bar{\pi} := \Phi(L^{-1}(\Phi^{-1}(\pi))),$$

where Φ is the CDF of the standard normal distribution.

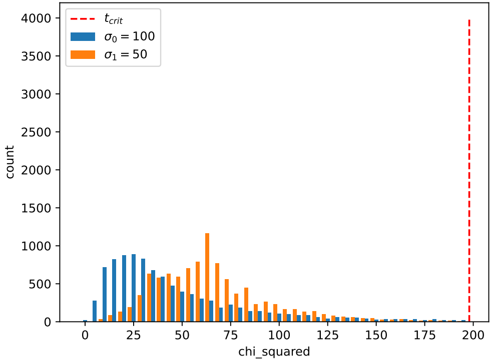
Power Analysis CCR: correlated & decorrelated





(5Y observation window, 1Y simulation horizon, 2W grid width)

PDF of Chi Squared Null vs. Alt in correlated case



- Null distribution of χ^2 has a very long tail
- Alternative distribution clearly different from null distribution
- Very hard to detect for the test though due to shape of distributions

Content

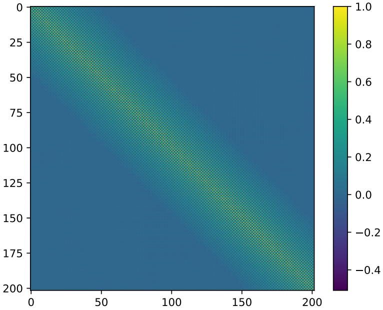
Multi-variate time series setting

Assume we have two correlated daily returns $Y_i^{(1)}$, $Y_i^{(2)}$ such that

$$Y_i^{(1)} \sim \mathcal{N}(0, \sigma_{0,1}^2), \qquad Y_i^{(2)} \sim \mathcal{N}(0, \sigma_{0,2}^2), \qquad \rho_0 := \rho(Y_i^{(1)}, Y_i^{(2)})$$

- This means that their corresponding m-day returns $X_i^{(1)}$, $X_i^{(2)}$ now have auto-correlation and cross-correlation.
- The null hypothesis now has three parameters $(\sigma_{0,1}, \sigma_{0,2}, \rho_0)$ and hence testing for canonical alternatives can also be performed in 3 dimensions.

Decorrelation



Decorrelation can be applied as well:

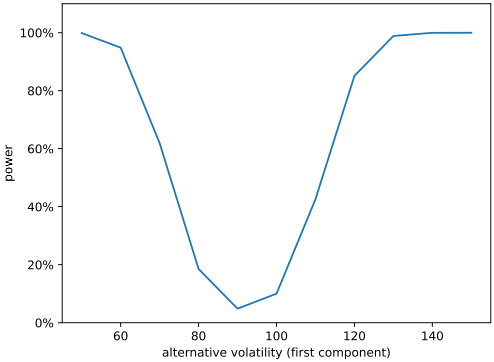
- Zip together the components $X_i^{(1)}$ and $X_i^{(2)}$ into one big vector

$$X = (X_1^{(1)}, X_1^{(2)}, X_2^{(1)}, X_2^{(2)}, \ldots, X_n^{(1)}, X_n^{(2)}).$$

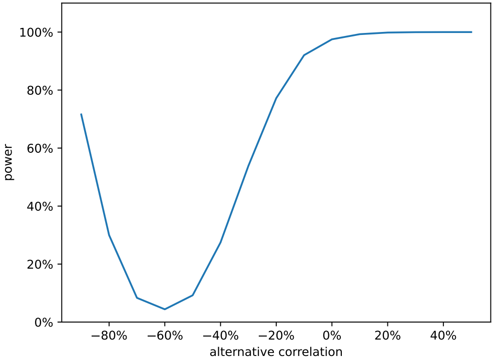
- Vector has correlation matrix $C = LL^\top$.
- Perform same tests (exceedence count, χ^2...) on zipped vector.
- Notice that we now have twice as many samples (not all equally sensitive to all alternatives though).

Power of Chi Squared at alternatives

alternative volatility



alternative correlation



Content

Summary

- The impact of correlations on the distribution of the test statistics and the power of the test is **very high** and hence **must not be ignored**.
- The impact of how choosing a strategy how to handle correlations is **very high**, often higher than the choice of test statistic.
- **Decorrelating the samples**
  - leads to higher power than correlating the test statistics,
  - avoids long-tailed distributions,
  - allows to re-use established statistical tests,
  - leads to natural generalizations for backtesting correlations itself or joint distributions of multiple quantities.

Thank you!

Pre-Print:

Nowaczyk, Piterbarg. *Backtesting Correlated Quantities*, 09/2023,
https://ssrn.com/abstract=4571812

Risk publication:

Nowaczyk, Piterbarg. *Backtesting Correlated Quantities*, 09/2024,
https://www.risk.net/cutting-edge/7959963/backtesting-correlated-quantities

mail@nikno.de
https://uk.linkedin.com/in/niknow
https://github.com/niknow

Content

Content

Derivation of Correlation Matrix

- **Question:** Given that the decorrelated test of the *m*-day returns has the same power curve as the 1-day return test, are those the same tests?
- **Answer: No.**
- Let $X = (X_1, \ldots, X_{m-n+1})$ the vector of *m*-day returns, let $Y = (Y_1, \ldots, Y_n)$ the vector of 1-day returns.

$$A \in \mathbb{R}^{(n-m+1)\times n} \qquad A_{ij} := \begin{cases} 1, & i \leq j \leq i+m-1, \\ 0, & \text{otherwise.} \end{cases}$$

Then Consequently $X = AY$ and

$$\mathbb{V}[X] = \mathbb{V}[AX] = A\mathbb{V}[X]A^\top = \sigma^2 AA^\top \in \mathbb{R}^{n\times n}$$

$$C = \frac{1}{m}AA^\top = LL^\top,$$

but this does not imply that $A = \sqrt{m}L$.

Linear Algebra

Matrix A is

- rectangular, $A \in \mathbb{R}^{(n-m+1)\times n}$,
- upper-triangular,
- surjective, but not injective since $\dim \ker A = m - 1$, hence not invertible.
- Example $(n = 5,\ m = 2)$:

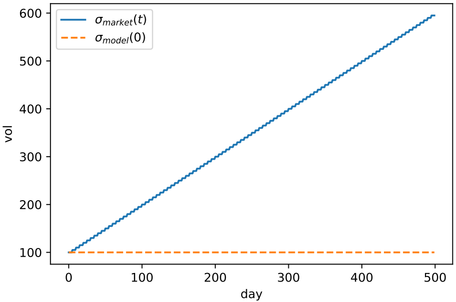
$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Matrix $\sqrt{m}\mathrm{L}$ is

- square, $\sqrt{m}\mathrm{L} \in \mathbb{R}^{n\times n}$,
- lower-triangular,
- invertible.
- Example $(n = 5,\ m = 2)$:

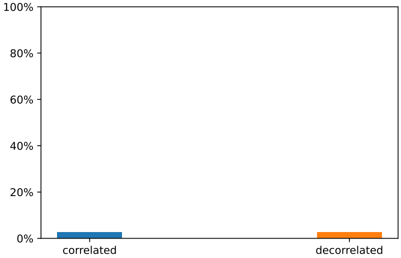
$$\sqrt{m}\mathrm{L} = \begin{pmatrix} 1.41 & 0 & 0 & 0 \\ 0.70 & 1.22 & 0 & 0 \\ 0 & 0.81 & 1.15 & 0 \\ 0 & 0 & 0.86 & 1.11 \end{pmatrix}$$

Statistical Example: Setup

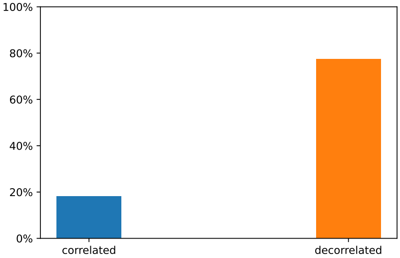


- Let the model be $dX_t = \sigma dW_t$, but with weekly recalibration.
- Assume the market follows an ABM but every week there is a regime change and the vol increases.
- Expect perfect performance of model at weekly horizon, but bad performance at yearly horizon.

Statistical Example: Probability of rejection



short horizon



long horizon

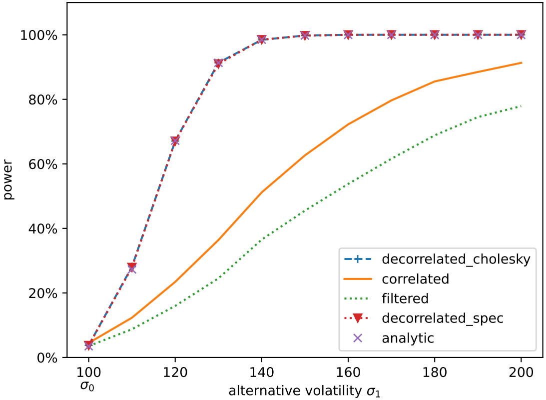
Content

Choice of Decorrelator

- **Question:** Is the Cholesky decomposition the only possibility to decorrelate the samples?
- **Answer: No.**
- The Cholesky decomposition $C = LL^\top$ is one possible choice that is computationally efficient, canonical as it is used in the Monte Carlo simulation to produce the correlation in the first place and it preserves **temporal consistency** as L is lower triangular.
- The spectral decomposition $C = O\Lambda O^\top$ with Λ a diagonal matrix and O an orthogonal matrix is an alternative decomposition that leads to the decorrelator M^{-1}, where $M = O\Lambda^{\frac{1}{2}}O^\top$. This decorrelator has the advantage that it is a symmetric matrix and that the resulting samples are as close as possible to the original samples, i.e.

$$M = \operatorname{argmin}_{AA^\top = C} \mathbb{E}[\|A^{-1}X - X\|^2]$$

Impact of decorrelator on power



No impact on power as power only depends on distribution

Content

Higher order Interactions

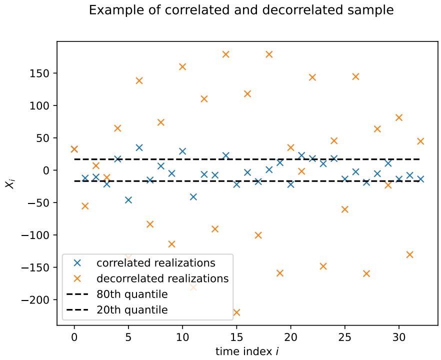
- **Question:** Are the decorrelated samples always independent?
- **Answer: No.**
- For Gaussian distributions uncorrelated and independent is the same, for many distributions it is similar, but it is in general not the same.
- It might still be helpful to decorrelate to remove first order interaction.

Content

Practical Interpretation

- **Question:** How to interpret the decorrelated samples? Can I plot them against a time series and do exception analysis?
- **Answer: Not really.**
- The intended purpose of the decorrelated samples is for calculation of the *p*-value only.
- Analysing the cause of a rejected model is still much easier using the original correlated sample but keeping in mind that the rejection can be caused by wrong volatility, wrong correlation or both.

Example



- Example: $n = 36$, $m = 4$, and a sample drawn from distribution with correct vol, but wrong (=0) correlation
- Correlated: $T = 10 < t_{\text{crit}}$, i.e. H_0 is retained
- Decorrelated: $\bar{T} = 16 > \bar{t}_{\text{crit}}$, hence H_0 is rejected

Content

Theoretical Background

- **Question:** Is there really no analytic formula for the distribution of a correlated exceedence counter?
- **Answer: No.**
- The distribution of the correlated exceedence counter has this neat compact formula:

$$\forall 0 \leq m \leq n : \mathbb{P}[T \leq m] = \sum_{k=0}^{m} \sum_{\substack{I \dot{\cup} J = \underline{n} \\ |I|=k}} \sum_{\nu=0}^{k} \sum_{\substack{L \subset I \\ |L|=\nu}} (-1)^{\nu} F_{J,L}(h_{J,L}),$$

where $\underline{n} := \{1, \ldots, n\}$ and for any multi-index I, F_I is the CDF of $X_I := (X_{i_1}, \ldots, X_{i_k})$ and X is any n-dimensional random variable with continuous CDF and $T := \sum_{i=1}^{n} 1_{X_i > h_i}$ is its exceedence counter.