Stochastic Machine Learning
Chapter 03 - Estimating risk

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

SS 2024

Risk and Risk measures

- Now we start an important discussion, which is central in the regulation of banks: the measurement of **risk**
- It has become an own strand of research - the study of risk measures.
- Here we only shortly touch upon this topic. For more details or literature we refer to Föllmer and Schied (2011).

Loss Distributions

Risks are represented by **random variables** mapping unforeseen future states of the world into values representing **profits and losses**.

The risks which interest us are **aggregate** risks. In general we consider a **portfolio** which might be

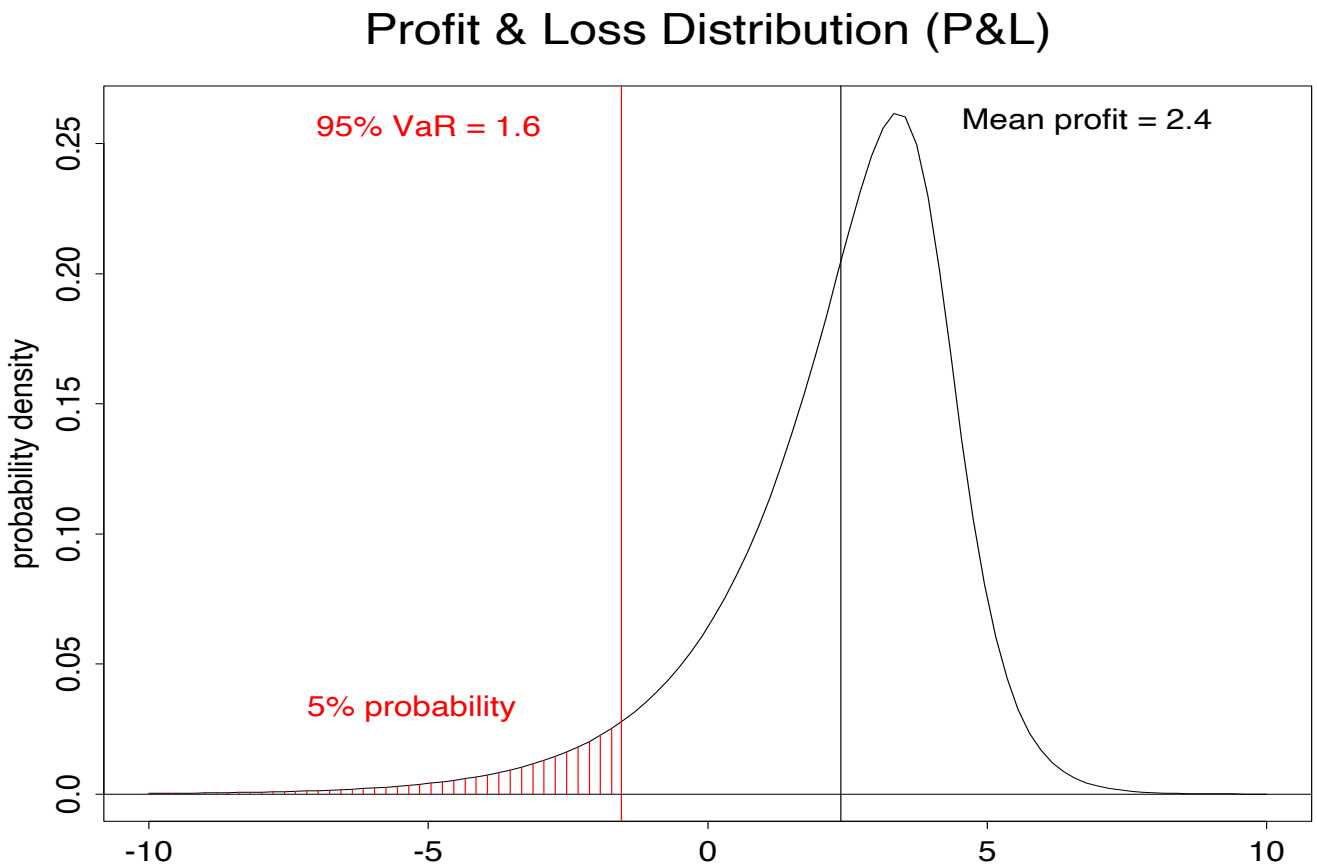
- a collection of **stocks and bonds**;
- a book of **derivatives**;
- a collection of risky **loans**;
- a financial institution's **overall position** in risky assets.

Portfolio Values and Losses

- Consider a portfolio and let V_t denote its **value** at time t; we assume this random variable is **observable** at time t.
- Suppose we look at risk from perspective of time t and we consider the time period $[t, t+1]$. The value V_{t+1} at the end of the time period is unknown to us.
- The distribution of $(V_{t+1} - V_t)$ is known as the profit-and-loss or **P&L distribution**. We denote the **loss** by $L_{t+1} = -(V_{t+1} - V_t)$. By this convention, losses will be positive numbers and profits negative.

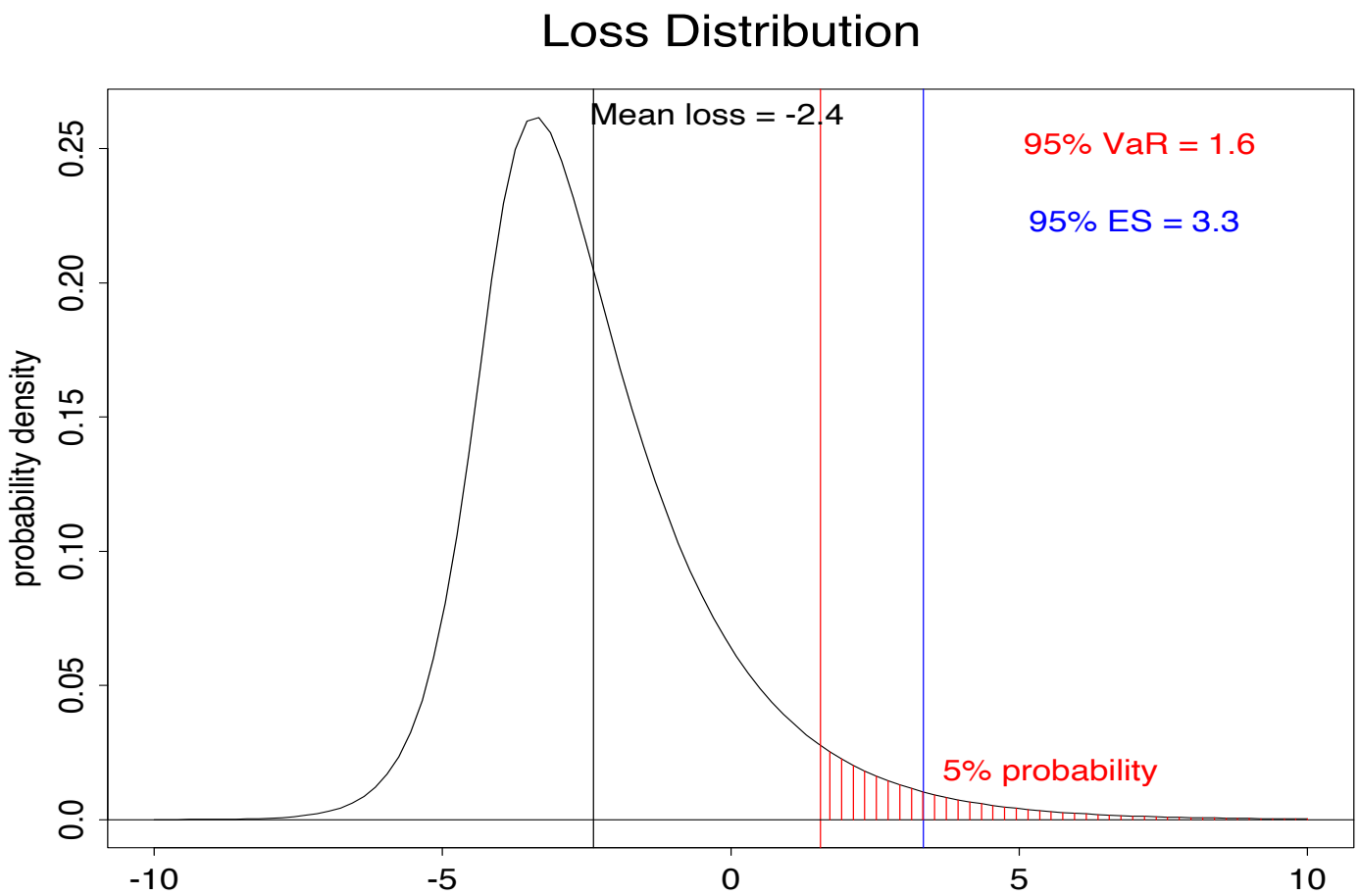
We refer to the distribution of L_{t+1} as the **loss distribution**.

An example



- Profits and losses. Now we turn it around and obtain ...

An example



- Profits and losses. Now we turn it around and obtain ...

Risk measures

Risk measures attempt to quantify the riskiness of a portfolio. The most popular risk measures like value-at-risk **VaR** describe the right tail of the loss distribution of L_{t+1} (or the left tail of the P&L).

- Think of a risk measure as a method of quantifying the amount of capital which is needed to make a financial position acceptable. This can be made mathematically precise (soon) and will lead to **coherent** and **convex** risk measures.
- We will also meet risk measures in the context of **fairness of AI** - here it is used as a deviation of a positive random variable from zero. It will be one of our tasks to critically review this approach and probably suggest improved alternatives !
- So first - what is a risk measure?

Probability and Quantile Transforms

- Denote the distribution function of the loss $L := L_{t+1}$ by F_L so that $P(L \leq x) = F_L(x)$.
- Recall: The (generalized) **inverse** of the cdf F is

$$F^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\}$$

for any $t \in (0, 1)$. $q_\alpha := F^{-1}(\alpha)$ is the α-**Quantile** of F.

VaR and Expected Shortfall

Let $0 < \alpha < 1$ (typically $\alpha = 0.95,\ 0.99$). We use

- **Value at Risk** is defined as

$$\mathrm{VaR}_\alpha = q_\alpha(F_L) = F_L^{-1}(\alpha)\,, \tag{1}$$

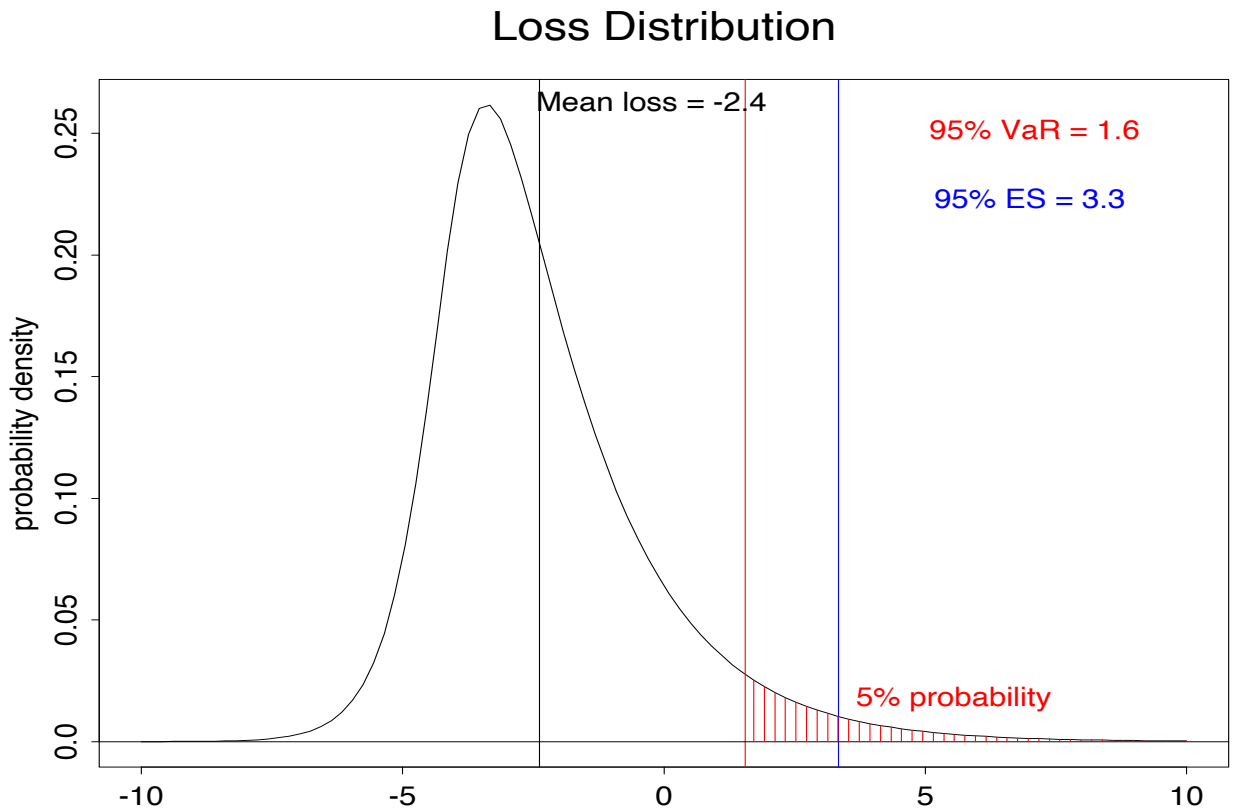
  where we use the notation $q_\alpha(F_L)$ or $q_\alpha(L)$ for a quantile of the distribution of L and F_L^{-1} for the (generalized) inverse of F_L.

- Provided $E(|L|) < \infty$ **expected shortfall** is defined as

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_L)du. \tag{2}$$

- Here - instead of fixing one level α, we average over all levels above α.

VaR in Visual Terms



Expected Shortfall

- For **continuous** loss distributions expected shortfall is the expected loss, given that the VaR is exceeded.

Lemma

Assume that L is continuous. Then, for any $\alpha \in (0,1)$ we have

$$\mathrm{ES}_\alpha = \frac{E\big[L \cdot \mathbb{1}_{L \geq q_\alpha(L)}\big]}{1-\alpha} = E[L \mid L \geq \mathrm{VaR}_\alpha]\,.$$

Proof(blackboard)

For a discontinuous loss df we have the more complicated expression

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\Big(E\big[L \cdot \mathbb{1}_{L \geq q_\alpha(L)}\big] + q_\alpha(1-\alpha-P(L \geq q_\alpha))\Big)\,.$$

Coherent Measures of Risk

- There are many possible measures of the risk in a portfolio such as VaR, ES or stress losses. To decide which are reasonable risk measures a systematic approach is called for.
- The authors in Artzner et al. (1999) achieved a list of properties which a risk measure should have - and called them **coherent**
- Thereafter a study of coherence started (value-at-risk is not always coherent, expected shortfall is, etc.)
- and further extensions have been proposed: for example, convex risk measures.

Purposes of Risk Measurement

Risk measures are used for the following purposes:

- Determination of **risk capital**. Risk measure gives amount of capital needed as a buffer against (unexpected) future losses to satisfy a regulator.
- **Management tool**. Risk measures are used in internal limit systems.
- **Insurance premia** can be viewed as measure of riskiness of insured claims.

Interpretation: Risk measure gives amount of capital that needs to be added to a position with loss L, so that the position becomes **acceptable** to an (internal/external) regulator.

The Axioms

A coherent risk measure is a real–valued function ϱ on some space of rv's (representing losses), s.t.

1. **Monotonicity**. For two rv's with $L_1 \geq L_2$ we have $\varrho(L_1) \geq \varrho(L_2)$.
2. **Subadditivity**. For any L_1, L_2 we have $\varrho(L_1 + L_2) \leq \varrho(L_1) + \varrho(L_2)$.

   This is the most debated property. Necessary for following reasons:

   - Reflects idea that risk can be reduced by **diversification** and that "a merger creates no extra risk".
   - Makes **decentralized** risk management possible.
   - If a regulator uses a non-subadditive risk measure, a financial institution could reduce risk capital by splitting into subsidiaries.
3. **Positive homogeneity**. For $\lambda \geq 0$ we have that $\varrho(\lambda L) = \lambda \varrho(L)$. If there is no diversification we should have equality in subadditivity axiom.
4. **Translation invariance**. For any $a \in \mathbb{R}$ we have that $\varrho(L + a) = \varrho(L) + a$.

Remarks:

- VaR is in general not coherent. ES (as we have defined it) is coherent.
- Non-subadditivity of VaR is relevant in presence of **skewed** loss distributions (credit-risk management, derivative books), or if traders **optimize against VaR**.
- Many recent papers study **convex** risk measures. Here instead of Subadditivity and Positive Homogeneity one has convexity:

$$\rho(\lambda L_1 + (1-\lambda)L_2) \leq \lambda\rho(L_1) + (1-\lambda)\rho(L_2)$$

for all $\lambda \in [0,1]$.

Non-Coherence of VaR: an Example

Consider portfolio of 50 defaultable bonds with independent defaults. Default probability identical and equal to 2%. Current price of bonds equal to 95, face value equal to 100.

1. Portfolio A: buy 100 units of bond 1; current value is $V_0 = 9500$.
2. Portfolio B : buy 2 units of each bond; current value is $V_0 = 9500$.

Common sense. Portfolio B is less risky (better diversified) than Portfolio A. This is wrong if we measure risk with VaR !

Loss of each bond equals

$$L_i := 95 - 100(1 - Y_i) = 100Y_i - 5,$$

where $Y_i = 1$ if default occurs, $Y_i = 0$ else. Y_i are iid Bernoulli(0.02).

Non-Coherence of VaR: an Example II

$$L_i := 95 - 100(1 - Y_i) = 100Y_i - 5.$$

Portfolio A: $L = 100L_1$ and hence

$$\mathrm{VaR}_{0.95}(L) = 100\,\mathrm{VaR}_{0.95}(L_1) = -500,$$

i.e. we may take 500 out of portfolio and still satisfy regulator.

Portfolio B: $L = \sum_{i=1}^{50} 2L_i = 200\sum_{i=1}^{50} Y_i - 500$, and hence

$$\mathrm{VaR}_\alpha(L) = 200\, q_\alpha\left(\sum_{i=1}^{50} Y_i\right) - 500.$$

Inspection shows that $q_{0.95}(\sum_{i=1}^{50} Y_i) = 3$, so that $\mathrm{VaR}_{0.95}(L) = 100$, i.e. **extra capital is needed** to hold the portfolio.

This is directly linked to non-coherence of VaR.

Examples

Further, we want to compute some examples and learn how to elaborate with this context. We focus on the easiest case - when losses are normal.

- Assume that L is normally distributed with mean μ and variance σ^2.
- The value-at-risk is easily computed:

$$\mathrm{VaR}_\alpha = q_\alpha(L) = \Phi^{-1}\Big(\frac{\alpha - \mu}{\sigma}\Big).$$

Lemma

For normally distributed losses we have that

$$\mathrm{ES}_\alpha = \mu + \sigma \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}.$$

Proof (blackboard)

Summary

- We have met risk measures and their two important classes: coherent and convex risk measures.
- We also saw the two important special cases value-at-risk and expected shortfall
- and showed that expected shortfall is an average over the value-at-risk when distribution functions are continuous (inspiring spectral risk measures).

The estimation of risk

Now we look into the paper Pitera and Schmidt (2018)

- Assume that the **P&L** of a portfolio, say X, is normally distributed with mean μ and variance σ^2. Then the value-at-risk at level α is given by

$$\mathrm{VaR}_\alpha(X) = -\Big(\mu + \sigma\Phi^{-1}(\alpha)\Big).$$

- But in practice, μ and σ are unknown and have to be estimated. In this regard, let us consider the simplest case: we have an i.i.d. sample $X_1, \ldots, X_n =: \boldsymbol{X}$ at hand.
- **Efficient** estimators of μ and σ are at hand:

$$\hat{\mu}_n = \bar{\boldsymbol{X}}, \qquad \hat{\sigma}_n = \bar{\sigma}(\boldsymbol{X}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})^2}. \tag{3}$$

- **Common practice** is to use the plug-in estimator

$$\mathrm{VaR}_\alpha^{plugin} := -\Big(\hat{\mu}_n + \hat{\sigma}_n\Phi^{-1}(\alpha)\Big).$$

- Can this be efficient?

Motivation from Backtesting

- Let us perform a standard backtesting-procedure, i.e. we run several simulations, estimate the value-at-risk and check if the percentage of unsufficient capital does not exceed 5%.
- We show estimates of $\mathrm{VaR}_{0.05}$ with the plug-in procedure. The data is from NASDAQ100 and from a sample from normally distributed random variable with mean and variance fitted to the NASDAQ data ("Simulated", second column), both for 4.000 data points $\rightarrow$ we would expect **200** exceedances.
- **Exceeds** reports the number of exceptions in the sample, where the actual loss exceeded the risk estimate.

Estimator		NASDAQ		Simulated	
		exceeds	percentage	exceeds	percentage
Plug-in	$\hat{\mathrm{VaR}}_{\alpha}^{\mathrm{plugin}}$	241	0.061	221	0.056

Motivation from Statistics

- In the normal case, for **known** σ, the likelihood-ratio test turns out to be the Gauss-test, or, equivalently, the confidence-interval is a normal distribution.
- If σ is **unknown**, one utilizes the t-distribution to obtain an efficient test: consider w.l.o.g. the test for $\mu = 0$ versus $\mu \neq 0$. The standardized test statistic is

$$T(X_1, \ldots, X_n) =: T(\boldsymbol{X}) = \frac{\sqrt{n}\,\bar{\boldsymbol{X}}}{\bar{\sigma}(\boldsymbol{X})}$$

  and the test rejects the null hypothesis if

$$T(X) > t_n(1-\alpha).$$

- Shouldn't there be a similar adjustment towards the t-distribution in the estimator for VaR?

- Our findings suggest that the estimator is biased. In a statistical sense !
- Our goal is to analyse this problem and give a new notion of unbiasedness in an **economic sense**.

The estimation of risk

We begin with well-known results on the measurement of risk.

- Let $(\Omega, \mathcal{A})$ be a measurable space and $(P_\theta : \theta \in \Theta)$ be a family of probability measures.
- For simplicity, we assume that the measures P_θ are equivalent, such that their null-sets coincide.
- For the estimation, we assume that we have a sample $X_1, X_2, \ldots, X_n$ of observations at hand.
- A risk measure ρ is a mapping from L^0 to $\mathbb{R} \cup \{+\infty\}$.
- The value $\rho(X)$ is a quantification of risk for a future position: it is the amount of money one has to add to the position X such that the position becomes acceptable.

A priori, the definition of a risk measure is formulated without any relation to the underlying probability. However, in most practical applications one typically considers law-invariant risk-measures. Denote by the convex space of cumulative distribution functions of real-valued random variables.

Definition

The family of risk-measures $(\rho_\theta)_{\theta\in\Theta}$ is called **law-invariant**, if there exists a function $R :\to \mathbb{R}\cup\{+\infty\}$ such that for all $\theta \in \Theta$ and $X \in L^0$

$$\rho_\theta(X) = R(F_X(\theta)), \tag{4}$$

$F_X(\theta) = P_\theta(X \leq \cdot)$ denoting the cumulative distribution function of X under the parameter θ.

Estimation

We aim at estimating the risk of the future position when $\theta \in \Theta$ is unknown and needs to be estimated from a data sample $x_1, \ldots, x_n$.

Definition

An **estimator** of a risk measure is a Borel function $\hat{\rho}_n : \mathbb{R}^n \to \boldsymbol{R} \cup \{+\infty\}$.

Sometimes we will call $\hat{\rho}_n$ also risk estimator.

The following definition introduces an economically motivated formulation of unbiasedness.

Definition

The estimator $\hat{\rho}_n$ is called **unbiased** for $\rho(X)$, if for all $\theta \in \Theta$,

$$\rho_\theta(X + \hat{\rho}_n) = 0. \tag{5}$$

- If the estimator is unbiased, adding the estimated amount of risk capital $\hat{\rho}_n$ to the position X makes the position $X + \hat{\rho}_n$ acceptable under all possible scenarios $\theta \in \Theta$.
- Requiring equality in Equation (5) ensures that the estimated capital is not too high.
- Except for the i.i.d. case, the distribution of $X + \hat{\rho}_n$ does also depends on the dependence structure of $X, X_1, \ldots, X_n$ and not only on the (marginal) laws.

Relation to the statistical definition of unbiasedness

Our Definition differs from unbiasedness in the statistical sense!

- The estimator $\hat{\rho}_n$ is called **statistically unbiased**, if

$$E_\theta[\hat{\rho}_n] = \rho_\theta(X), \qquad \text{for all } \theta \in \Theta, \tag{6}$$

- One point why the statistical unbiasedness is not reasonable here is that it does not behave well in various backtesting or stress-testing procedures.

Unbiased estimation of value-at-risk under normality

- Let $X \sim \mathcal{N}(\theta_1, \theta_2^2)$ and denote $\theta = (\theta_1, \theta_1) \in \Theta = \mathbb{R} \times \mathbb{R}_{>0}$.
- The value-at-risk is

$$\rho_\theta(X) = \inf\{x \in \boldsymbol{R} \colon P_\theta[X + x < 0] \leq \alpha\}, \qquad \theta \in \Theta, \tag{7}$$

- Unbiasedness as defined in Equation (5) is equivalent to

$$P_\theta[X + \hat{\rho} < 0] = \alpha, \qquad \text{for all } \theta \in \Theta. \tag{8}$$

- We define estimator $\hat{\rho}$, as

$$\hat{\rho}(x_1, \ldots, x_n) = -\bar{x} - \bar{\sigma}(\boldsymbol{x}) \sqrt{\frac{n+1}{n}} t_{n-1}^{-1}(\alpha), \tag{9}$$

This estimator is **unbiased**: first, note that

$$X + \hat{\rho} = X - \bar{X} - \bar{\sigma}(\boldsymbol{X})\sqrt{\frac{n+1}{n}}t_{n-1}^{-1}(\alpha) \leq 0$$

$$\Leftrightarrow \qquad \sqrt{\frac{n}{n+1}} \cdot \frac{X - \bar{X}}{\bar{\sigma}(\boldsymbol{X})} \leq t_{n-1}^{-1}(\alpha).$$

Using the fact that X, $\bar{X}$ and $\bar{\sigma}(\boldsymbol{X})$ are independent for any $\theta \in \Theta$, we obtain

$$T := \sqrt{\frac{n}{n+1}} \cdot \frac{X - \bar{X}}{\bar{\sigma}(\boldsymbol{X})} = \frac{X - \bar{X}}{\sqrt{\frac{n+1}{n}}\theta_2} \cdot \sqrt{\frac{n-1}{\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{\theta_2})^2}} \sim t_{n-1}.$$

Thus, the random variable T is a pivotal quantity and

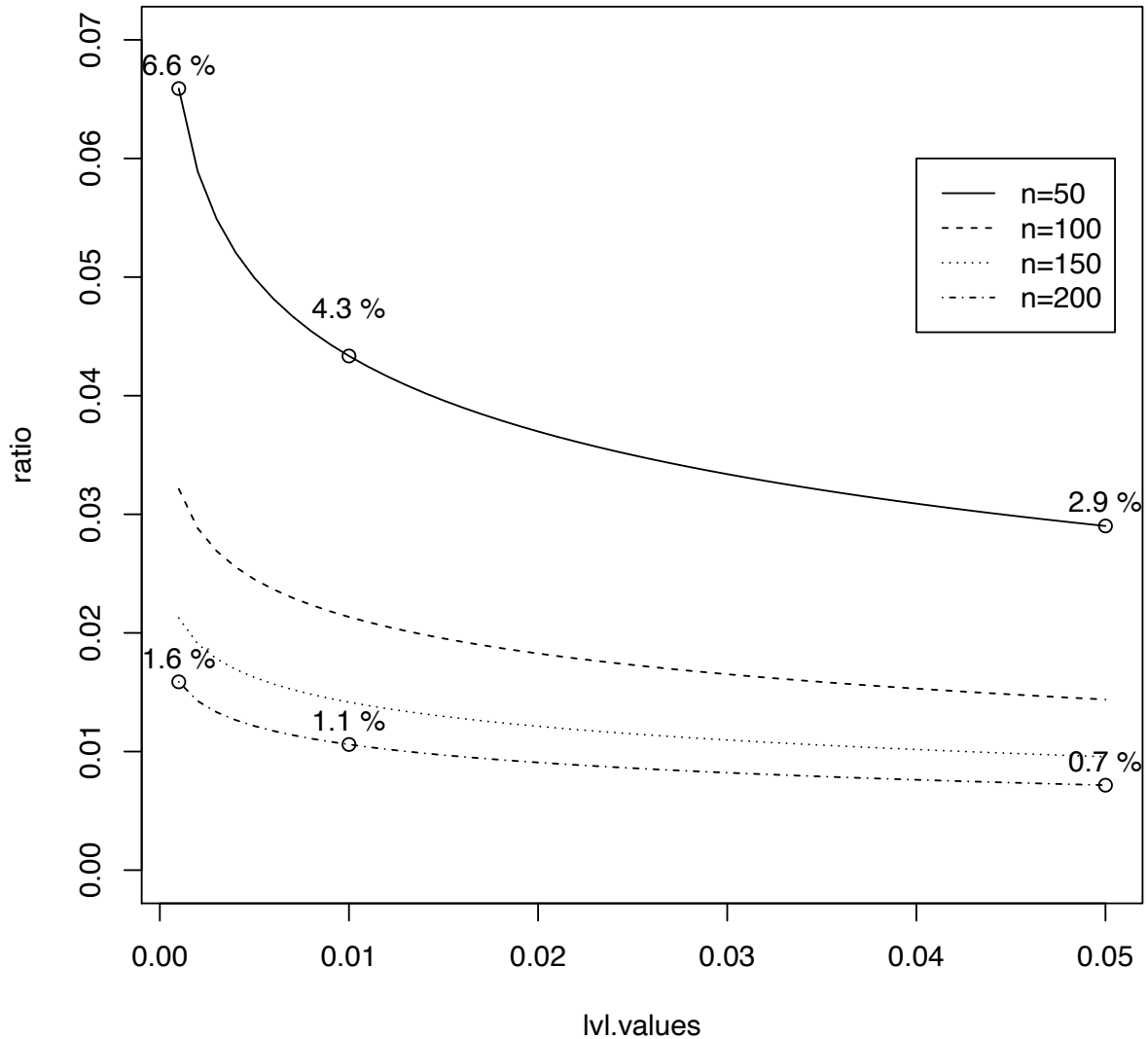
$$P_\theta[X + \hat{\rho} < 0] = P_\theta[T < q_{t_{n-1}}(\alpha)] = \alpha.$$

Let us elaborate a little bit on the difference between the plug-in and the unbiased estimator.

$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{u}} = -\bar{x} - \bar{\sigma}(\boldsymbol{x})\sqrt{\frac{n+1}{n}}\, t_{n-1}^{-1}(\alpha)$$
$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{plugin}} = -\bar{x} - \bar{\sigma}(\boldsymbol{x})\Phi^{-1}(\alpha)$$

The percentage of additional capital over the mean needed for the unbiased estimator is given by

$$\sqrt{\frac{n+1}{n}}\,\frac{t_{n-1}^{-1}(\alpha)}{\Phi^{-1}(\alpha)}. \tag{10}$$



The percentage of additional capital over the mean from Equation (10).

Empirical evidence

- Consider a large i.i.d. sample from the standard normal distribution and measure the risk using VaR at level 1%. We follow a rolling window approach with fixed learning period ($l := 250$) and backtesting period ($b := 1$).
- We **backtest** as follows: we construct secured position

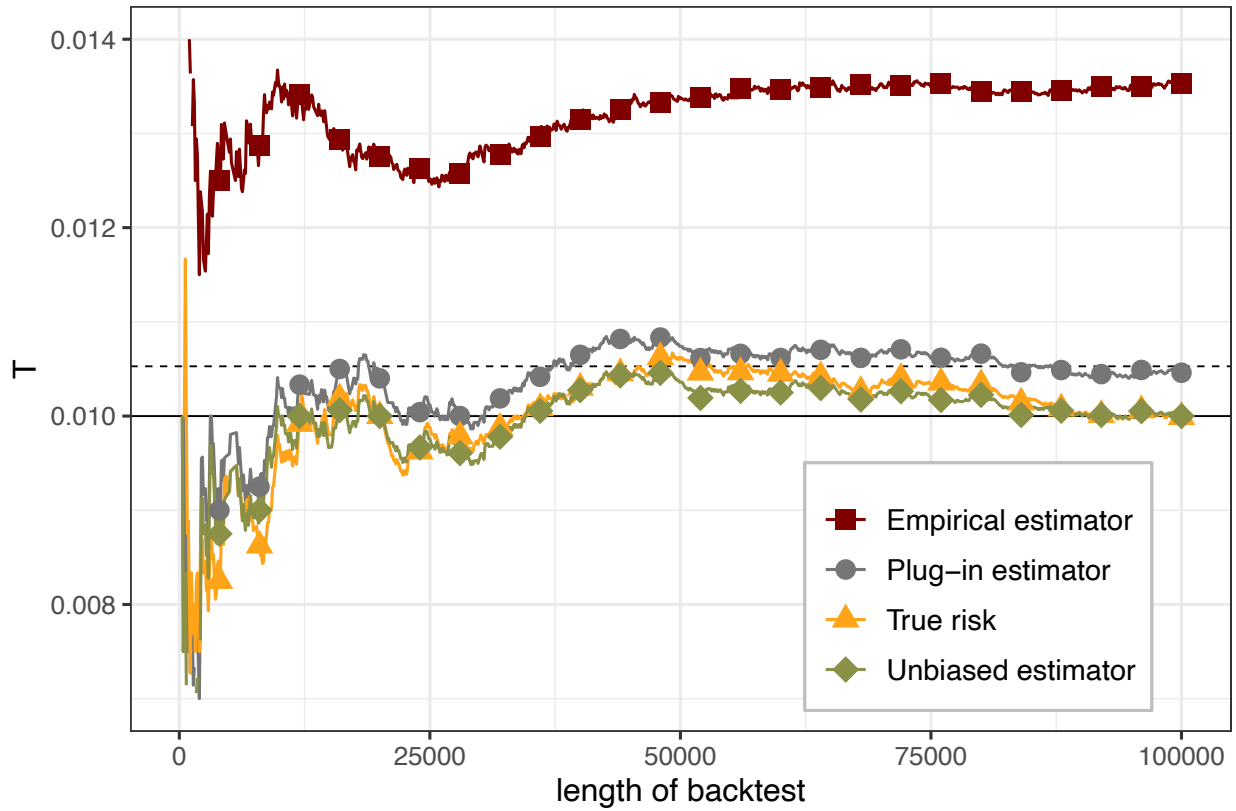
$$Y_t := X_t + \hat{\mathrm{VaR}}_t, \qquad t = l+1, \ldots, n,$$

where $\hat{\mathrm{VaR}}$ is the estimator (plug-in, empirial, unbiased ...)

- Then we compute the **average number of exceptions** up to day h,

$$T_h := \frac{1}{h} \sum_{i=1}^{h} \mathbb{1}_{\{Y_{l+i} < 0\}} \tag{11}$$

and show this sequence for increasing h in the following graph.


0.014
0.012
0.010
0.008
T
0
25000
50000
75000
100000
length of backtest
Empirical estimator
Plug−in estimator
True risk
Unbiased estimator

- Indeed, we are able to precisely compute this deviation: for $l = 250$, it is equal to

$$t_{l-1}\left[\sqrt{\frac{l}{l+1}}\Phi^{-1}(0.01)\right] \approx 1.05\%.$$

- Of course, this effect will get more pronounced, the heavier the tails are! (and the smaller α is).

Unbiased estimation of expected shortfall under normality

- We continue in the previous setting,
- The expected shortfall at level α under a continuous distribution is

$$\rho_\theta(X) = E_\theta[-X | X \leq q_X(\theta, \alpha)],$$

where $q_X(\theta, \alpha)$ is α-quantile of X under P_θ.

- We consider estimators of the form

$$\hat{\rho}(x_1, \ldots, x_n) = -\bar{x} - \bar{\sigma}(x) a_n, \tag{12}$$

for some $(a_n)_{n \in \boldsymbol{N}}$, where $a_n \in \mathbb{R}$.

- We can show that there exists a_n which makes $\hat{\rho}$ unbiased. This a_n can easily be computed numerically.

Empirical study

- It is the aim of this section to analyse the performance of selected estimators on various sets of real market data (Market) as well as on simulated data (Simulated). Our focus is on the practically most relevant risk measures, VaR and ES.
- The market data we use are returns from the data library **FamFre2015**, containing returns of 25 portfolios formed on book-to-market and operating profitability in the period from 27.01.2005 to 01.01.2015. We obtain exactly 2500 observations for each portfolio.
- The sample is split into 50 separate subsets, each consisting of 50 consecutive trading days. For $i = 1, 2, \ldots, 49$, we estimate the risk measure using the i-th subset and test it's adequacy on $(i + 1)$-th subset.
- The simulation study uses i.i.d. normally distributed random variables whose mean and variance was fitted to each of the 25 portfolios. The sample size was set to 2500 for each set of parameters. In this way we are able to exclude difficulties due to dependencies in the data or bad model fit.

Backtesting VaR

- We considered the unbiased estimator $\hat{\mathrm{VaR}}_\alpha^{\mathrm{u}}$, the empirical sample quantile $\hat{\mathrm{VaR}}_\alpha^{\mathrm{emp}}$, the modified Cornish-Fisher estimator $\hat{\mathrm{VaR}}_\alpha^{\mathrm{CF}}$, the plug-in estimator $\hat{\mathrm{VaR}}_\alpha^{\mathrm{norm}}$ and the GPD plug-in estimator[1] $\hat{\mathrm{VaR}}_\alpha^{\mathrm{GPD}}$.

$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{emp}}(x) := -\left(x_{(\lfloor h \rfloor)} + (h - \lfloor h \rfloor)(x_{(\lfloor h+1 \rfloor)} - x_{(\lfloor h \rfloor)})\right),$$

$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{CF}}(x) := -\left(\bar{x} + \bar{\sigma}(x)\bar{Z}_{CF}^{\alpha}(x)\right),$$

$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{norm}}(x) := -\left(\bar{x} + \bar{\sigma}(x)\Phi^{-1}(\alpha)\right),$$

$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{GPD}} := -u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{\alpha n}{k}\right)^{-\hat{\xi}} - 1\right),$$

$$\hat{\mathrm{VaR}}_\alpha^{\mathrm{u}}(x_1, \ldots, x_n) := -\left(\bar{x} + \bar{\sigma}(x)\sqrt{\frac{n+1}{n}}\, t_{n-1}^{-1}(\alpha)\right),$$

where $x_{(k)}$ is the k-th order statistic of $x = (x_1, \ldots, x_n)$, the value $\lfloor z \rfloor$ denotes the integer part of $z \in \boldsymbol{R}$, $h = \alpha(n-1)+1$, Φ denotes the cumulative distribution function of the standard normal distribution and $\bar{Z}_{CF}^{\alpha}$ is a standard Cornish-Fisher α-quantile estimator.

[1] For each portfolio, we set the threshold value u to match the 0.7-empirical quantile of the corresponding sample.

Table: Top: the results for portfolios in the period from 27.01.2005 to 01.01.2015 from the Fama & French dataset. Bottom: the results on simulated Gaussian data. We perform the standard backtest, splitting into intervals of length 50 and computing average rate of exceptions.

Type of data:	MARKET				
Portfolio	Estimator type				
	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{emp}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{norm}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{CF}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{GPD}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{u}}$
LoBM.LoOP	0.071	0.073	0.067	0.067	0.069
BM1.OP2	0.076	0.070	0.069	0.069	0.065
BM1.OP3	0.071	0.064	0.063	0.064	0.061
BM1.OP4	0.069	0.071	0.067	0.067	0.068
LoBM.HiOP	0.071	0.071	0.070	0.067	0.068
...	...	...	...	...	
mean	0.073	0.071	0.068	0.067	0.067

Type of data:	SIMULATED				
	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{emp}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{norm}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{CF}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{GPD}}$	$\hat{\mathrm{VaR}}_\alpha^{\mathrm{u}}$
LoBM.LoOP	0.065	0.057	0.055	0.056	**0.051**
BM1.OP2	0.064	0.053	0.053	0.053	**0.050**
BM1.OP3	0.069	0.058	0.058	0.060	**0.052**
BM1.OP4	0.069	0.057	0.058	0.062	**0.053**
LoBM.HiOP	0.060	0.054	0.053	0.056	**0.047**
...	...	...	...	...	
mean	0.066	0.057	0.057	0.058	**0.051**

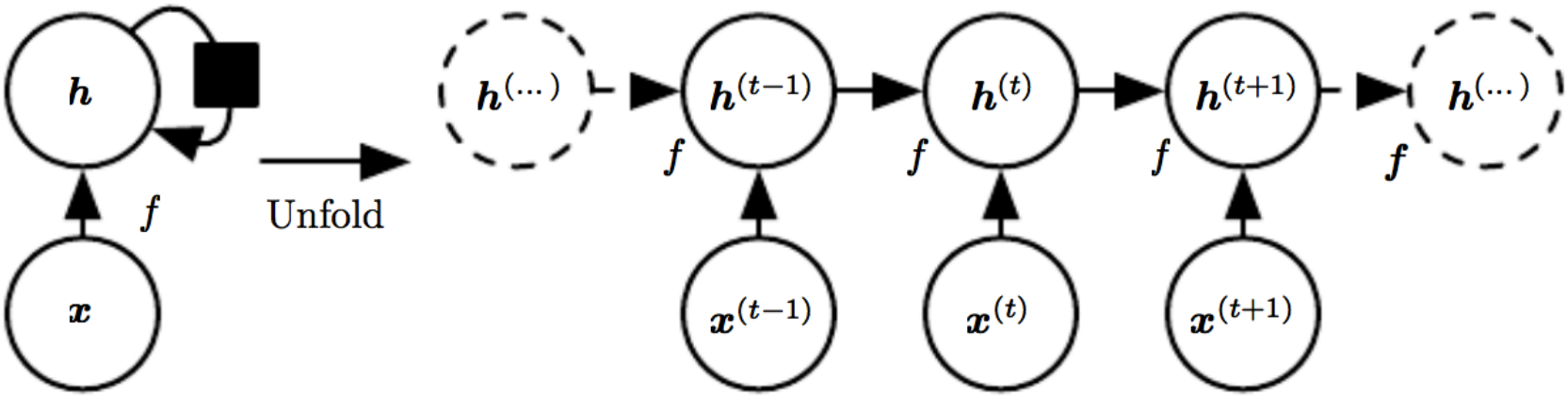
Summary

- In this part of the lecture we studied the estimation of risk, with a particular view on **unbiased** estimators and backtesting.
- The new notion of unbiasedness introduced is motivated from economic principles rather than from statistical reasoning, which links this concept to a better performance in backtesting.
- Some unbiased estimators, for example the unbiased estimator for value-at-risk in the Gaussian case, can be computed in closed form while for many other cases numerical methods are available.
- A small empirical analysis underlines the outperformance of the unbiased estimators with respect to standard backtesting measures.

Recurrent neural networks (RNNs)

- Neural Networks that allow for cycles in the connectivity graph
- Cycles let information persist in the network for some time (state), and provide a time-context or (fading) memory
- Very powerful for processing sequences
- Implement dynamical systems rather than function mappings, and can approximate any dynamical system with arbitrary precision

Unfolding the Computational Graph of an RNN



$$\begin{aligned} \boldsymbol{h}^{(t)} &= f(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}; \boldsymbol{\theta}) \\ &= f(f(\boldsymbol{h}^{(t-2)}, \boldsymbol{x}^{(t-1)}; \boldsymbol{\theta}), \boldsymbol{x}^{(t)}; \boldsymbol{\theta}) \end{aligned}$$

Regardless of the sequence length, the learned model always has the same input size!

Long Short Term Memory (LSTM) Networks

- Introduced in Hochreiter and Schmidhuber, 1997
- Address the vanishing gradient problem through units with special structure
- Memory cell with forget, input and output gates
- By now: standard RNN model in many applications

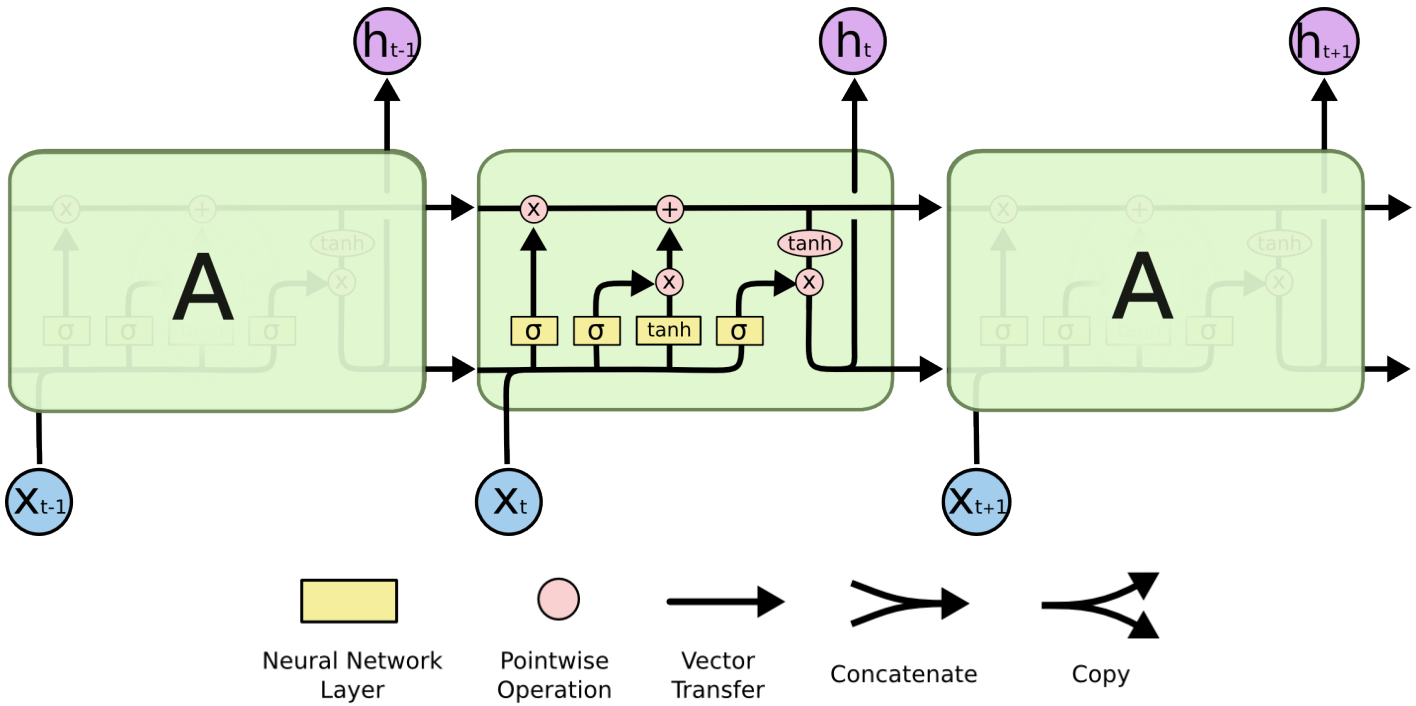


image credit: Christopher Olah

LSTM - Cell State

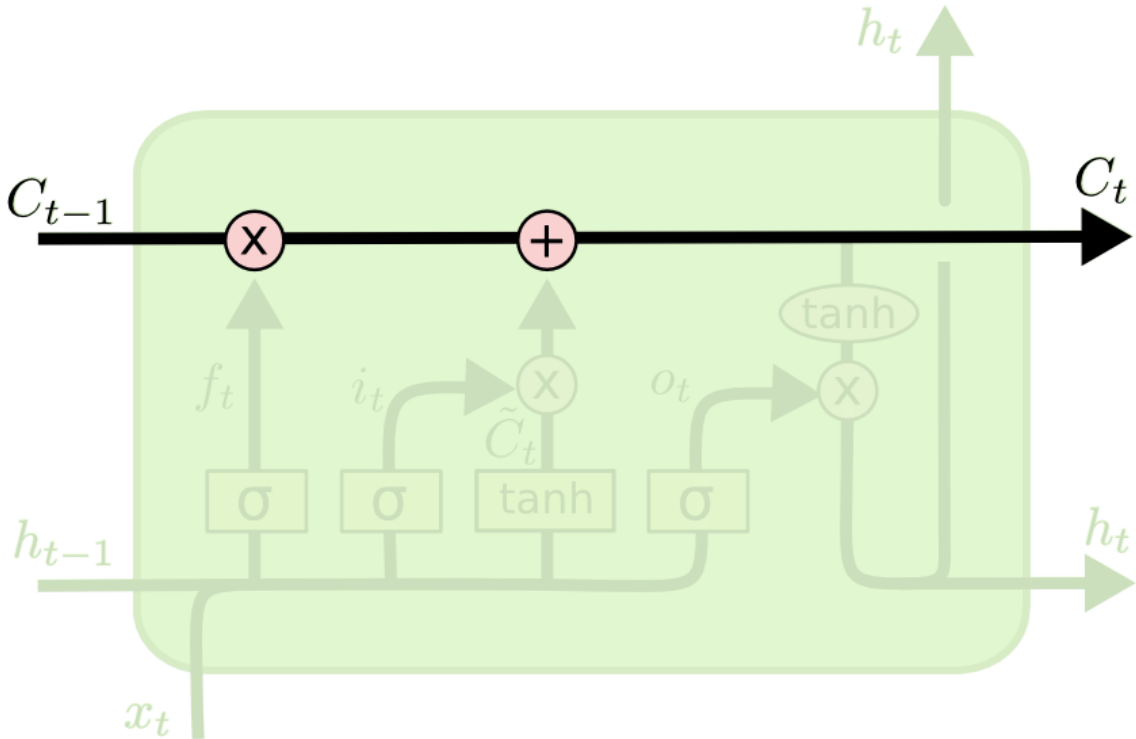
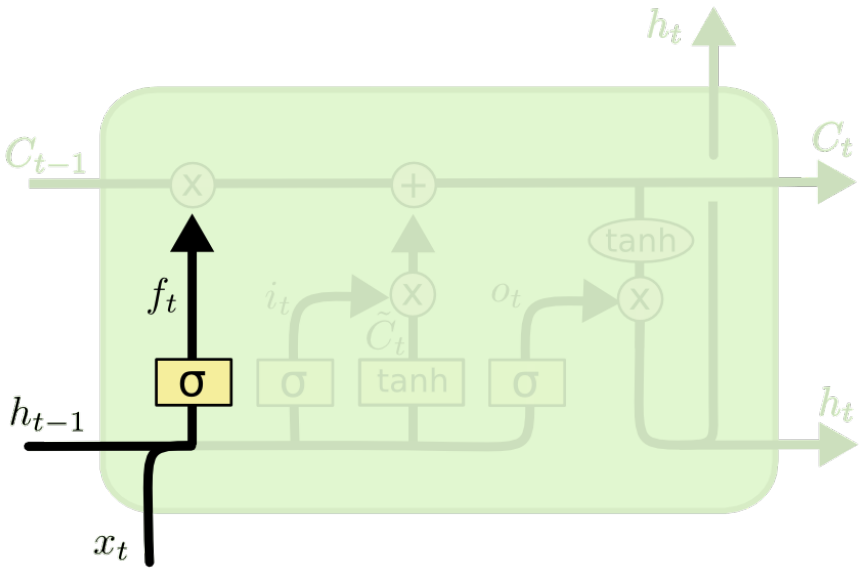


image credit: Christopher Olah

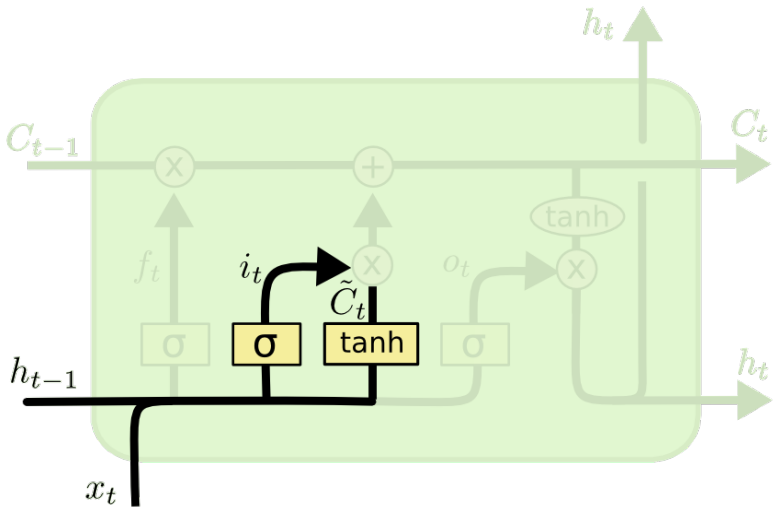
LSTM - Forget Gate



$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

image credit: Christopher Olah

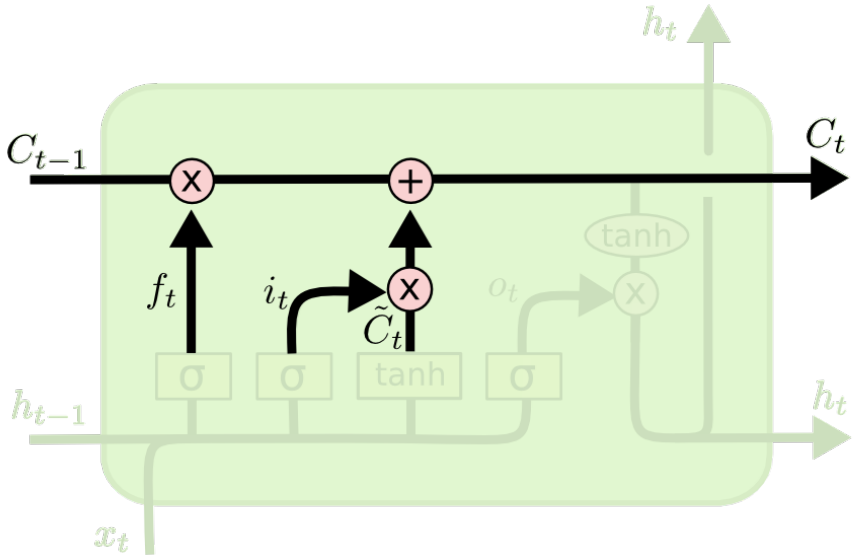
LSTM - Input Gate and Candidate Value



$$
\begin{aligned}
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)
\end{aligned}
$$

image credit: Christopher Olah

LSTM - Cell Update



$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

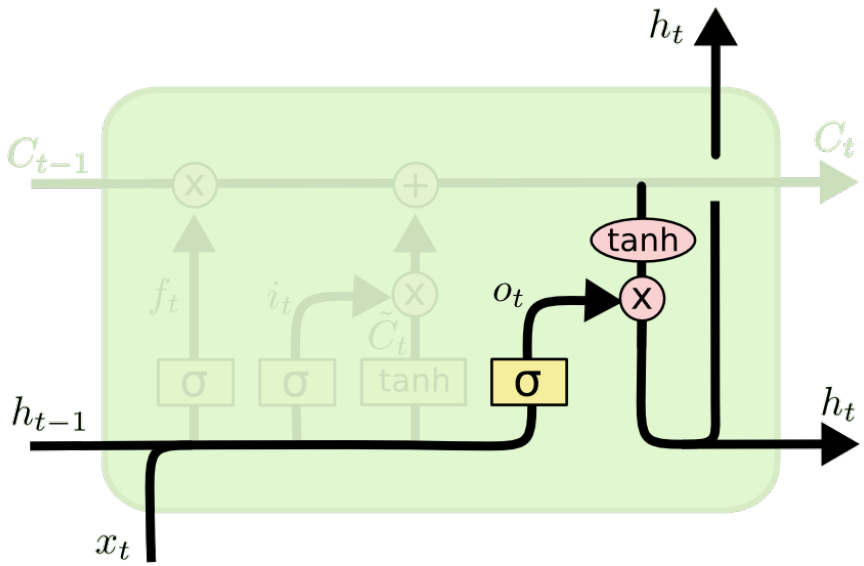
image credit: Christopher Olah

Here $\odot$ is the Hadamard product of two matrices $A = (a_{ij}),\ B = (b_{ij})$ gives

$$A \odot B = (a_{ij} \cdot b_{ij})_{ij}$$

i.e. it corresponds to componentwise multiplication.

LSTM - Output Gate and Hidden State



$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \odot \tanh(C_t)$$

image credit: Christopher Olah

Power of the Gating Mechanism

- LSTM can make variables *internal*, just for further computation
- LSTM can extract features of current input and add them selectively to its state
- LSTM can selectively forget and remember features
- LSTM variables already have their meaning: less training and less variables needed
- LSTM weights can more easily be interpreted
- fewer gradient decay problems because information flows by default. GRU (Gated recurrent units) are even a bit simpler.

End of theory

Now lets go to practice.

Machine learning

- If we have a large enough sample, we expect deep neural networks can learn an efficient estimation.
- Yes, but ....
- we will need additional care.
- Up to now we considered the i.i.d. case, and now we go for time series.
- An appropriated neural network for this is an LSTM.

- We fix the observation period i with (enhanced) data $\tilde{X}^i$.
- In each layer $h \in \{1, \dots, L+H\}$ we will have $d_h \in \mathbb{N}$ neurons.
- Each layer contains the *input gate*, the *forget gate*, and the *output gate* and the *cell input layer* (g): these are maps

$$ {}^k_h(\boldsymbol{x}) = \sigma^k(\boldsymbol{a}^k_h + \boldsymbol{x} A^k_h), $$

where σ^k is a sigmoid function if $k \in \{i, f, o\}$, an σ^k is tanh if $k = g$

- Given LSTM^h_{j-1}, we set

$$ \mathrm{LSTM}^h_j := o^h_j \odot \sigma(c^h_j), \quad \text{for } h = 1, 2, \dots, H. $$

where σ is the *activation function* tanh, and $a \odot b = (a_1 b_1, a_2 b_2, \dots)^\top$, and

$$ c^h_j := f^h_j \odot c^h_{j-1} + i^h_j \odot g^h_j $$

$$ \begin{aligned} o^h_j &:= {}^o_h([\mathrm{LSTM}^h_{j-1}, \mathrm{LSTM}^{h-1}_j]), \\ f^h_j &:= {}^f_h([\mathrm{LSTM}^h_{j-1}, \mathrm{LSTM}^{h-1}_j]), \\ i^h_j &:= {}^i_h([\mathrm{LSTM}^h_{j-1}, \mathrm{LSTM}^{h-1}_j]), \\ g^h_j &:= {}^g_h([\mathrm{LSTM}^h_{j-1}, \mathrm{LSTM}^{h-1}_j]), \end{aligned} $$

where we have used $[\cdot, \cdot]$ for row vector with dimension $d_h + d_{h-1}$.

The objective function

- We focus on a rolling-windows test. Have a time series $(X_1, X_2, \ldots, X_{m+n})$ of length $m+n$ at hand.
- The i-th estimation (training) dataset is $X^i := (X_i, X_{i+1}, \ldots, X_{i+n-1})$.
- The testing variable is $Y^i := X_{i+n}$.
- The neural network depends on parameters, which we denote by A.
- The LSTM therefore delivers the estimates

$$\hat{\rho}(X, A) = (\hat{\rho}(\tilde{X}^1, A), \ldots, \hat{\rho}(\tilde{X}^m, A)).$$

Data imputation

- The application to risk estimation will not work without **data imputation** !
- Note that we are in a small data environment ... !
- We enhance the dataset and use

$$\tilde{X}^i = \left(\begin{bmatrix} \bar{X}^i \\ \bar{X}^i + \bar{\sigma}^i \Phi^{-1}(\alpha) \\ X^i_{(n\alpha)+1} \\ f_1(X_i - \bar{X}^i) \\ f_2(X_i - \bar{X}^i) \\ f_3(X_i - \bar{X}^i) \\ f_4(X_i - \bar{X}^i) \end{bmatrix}, \begin{bmatrix} \bar{X}^i \\ \bar{X}^i + \bar{\sigma}^i \Phi^{-1}(\alpha) \\ X^i_{(n\alpha)+1} \\ f_1(X_{i+1} - \bar{X}^i) \\ f_2(X_{i+1} - \bar{X}^i) \\ f_3(X_{i+1} - \bar{X}^i) \\ f_4(X_{i+1} - \bar{X}^i) \end{bmatrix}, \ldots, \begin{bmatrix} \bar{X}^i \\ \bar{X}^i + \bar{\sigma}^i \Phi^{-1}(\alpha) \\ X^i_{(n\alpha)+1} \\ f_1(X_{i+n-1} - \bar{X}^i) \\ f_2(X_{i+n-1} - \bar{X}^i) \\ f_3(X_{i+n-1} - \bar{X}^i) \\ f_4(X_{i+n-1} - \bar{X}^i) \end{bmatrix} \right), \tag{13}$$

where f_i are centered Chebyscheff polynomials (think of x^k)

Results

- To begin with, we consider the classical statistic, the **exception rate** for the estimator $\hat{\mathrm{Var}}$

$$\mathrm{ER}(\hat{\mathrm{Var}}) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{Y^i + \hat{\mathrm{Var}}_\alpha(X^i) < 0\}}.$$

  The closer the value of ER to $\alpha = 0.05$, the closer the exception rate is to the true rate.
- Values of ER **larger** than α indicate underestimation of risk,

The i.i.d. case

- Can the LSTM compete in the i.i.d. case ?
- we simulate $100\,000$ i.i.d. $\mathcal{N}(0,1)$ and $t(0,1,\nu)$, with $\nu \in \{5, 10, 15\}$ degrees of freedom.
- We split the data into 90/5/5, estimate the parameters using the first two subsets and then evaluate performance of the estimators on a test subset of length 5 000.

Data	Model	ER (in %)			
		true	lstm	emp	u
Full	$\mathcal{N}(0,1)$	4.89	5.07	5.62	**4.85**
	$t_5(0,1)$	5.23	**5.23**	5.82	4.79
	$t_{10}(0,1)$	5.15	5.01	6.14	**5.09**
	$t_{15}(0,1)$	5.68	**5.76**	5.90	5.21

- Best performance for ER is probably the value closest to the true risk output.
- Even if very intuitive, ER might not be the very best statistic.

Elicitability

- So what is this thing about elicitability ? Actually, it is the quest for statistics which you can not trick:
- The exception rate can easily be tricked: think you have 100 samples and need an exception rate of 5% - easy: you provide 95 very high values and 5 very low ones. Without a look at the data !
- Since the score measures the height of the exceedance, this will easily result in a bad score.
- Recall that we also used the score to train the LSTM !
- We use once more the average score given by

$$\bar{S}_\alpha = \frac{1}{m}\sum_{i=1}^{m}(\alpha - \mathbb{1}_{\{Y^i+\hat{\rho}(X^i)\leq 0\}})\,(Y^i + \hat{\rho}(X^i)), \tag{14}$$

Data	Model	$\bar{S}_\alpha$			
		(true)	LSTM	emp	u
Full	$\mathcal{N}(0,1)$	0.104	0.104	0.109	0.106
	$t_5(0,1)$	0.150	0.149	0.159	0.154
	$t_{10}(0,1)$	0.118	0.118	0.123	0.121
	$t_{15}(0,1)$	0.118	0.118	0.124	0.121

GARCH time series

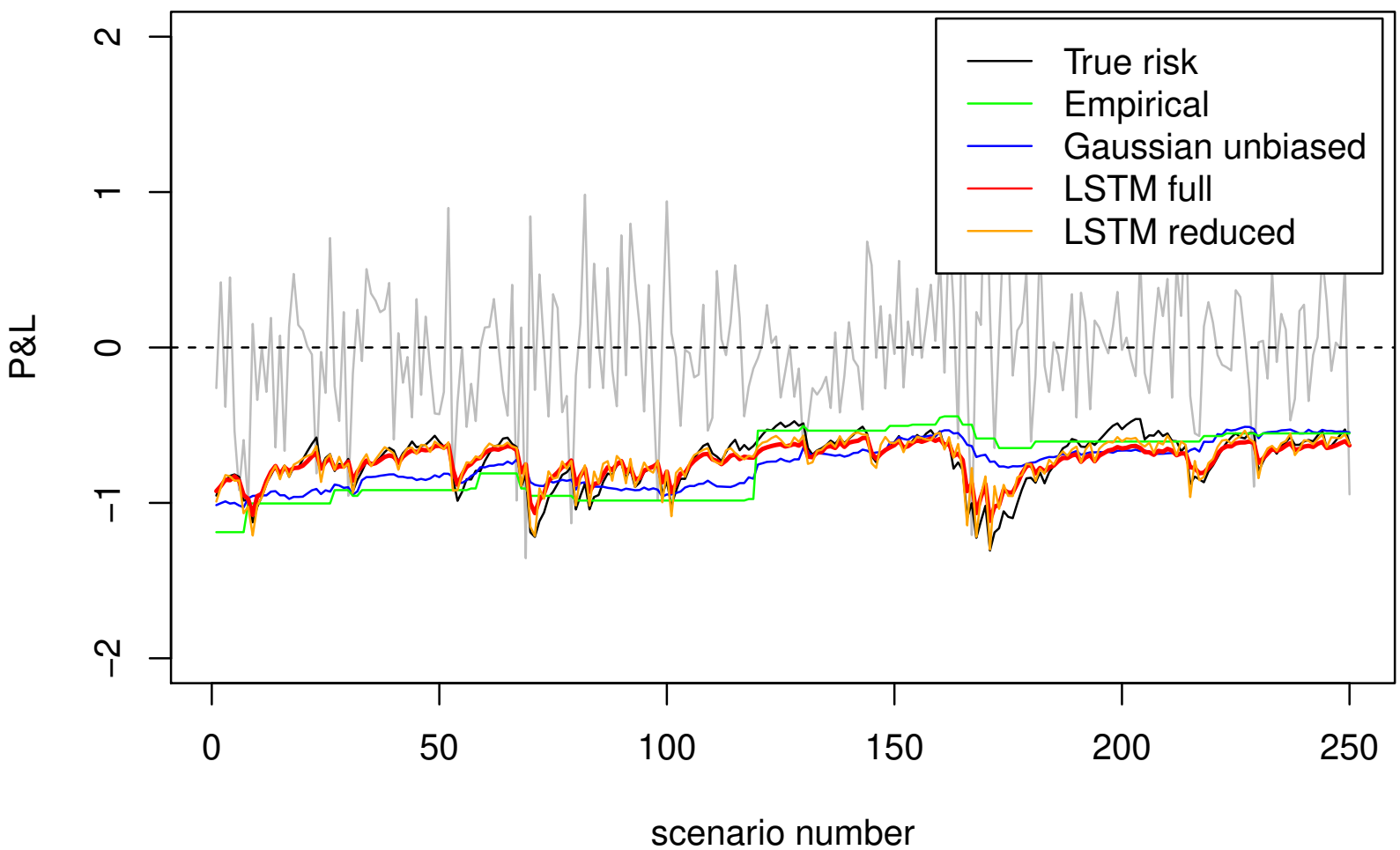
- Probably much more suited for financial time series is a GARCH model

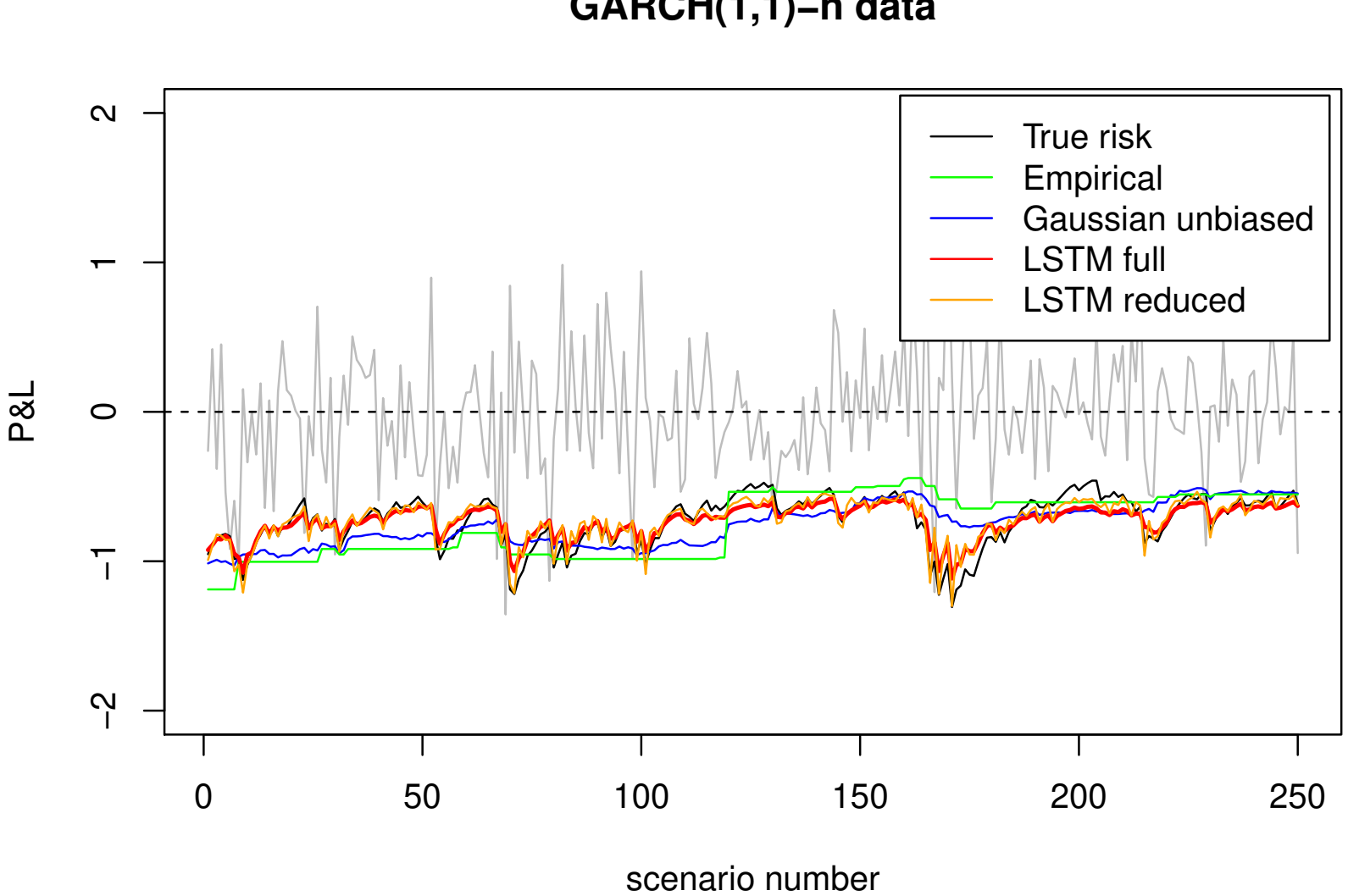
$$\begin{cases} r_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \end{cases}, \tag{15}$$

- We test on different parameter sets

Model	GARCH model specification						
	α_0	α_1	α_2	α_3	α_4	β_1	ϵ_t
GARCH(1, 1)-n	0.01	0.17	-	-	-	0.8	$\mathcal{N}(0,1)$
GARCH(2, 1)-n	0.01	0.12	0.05	-	-	0.8	$\mathcal{N}(0,1)$
GARCH(3, 1)-n	0.01	0.12	0.10	0.05	-	0.7	$\mathcal{N}(0,1)$
GARCH(4, 1)-n	0.01	0.12	0.05	0.05	0.05	0.7	$\mathcal{N}(0,1)$
GARCH(1, 1)-t	0.01	0.17	-	-	-	0.8	$t(0,1,5)$
GARCH(2, 1)-t	0.01	0.12	0.05	-	-	0.8	$t(0,1,5)$
GARCH(3, 1)-t	0.01	0.12	0.10	0.05	-	0.7	$t(0,1,5)$
GARCH(4, 1)-t	0.01	0.12	0.05	0.05	0.05	0.7	$t(0,1,5)$

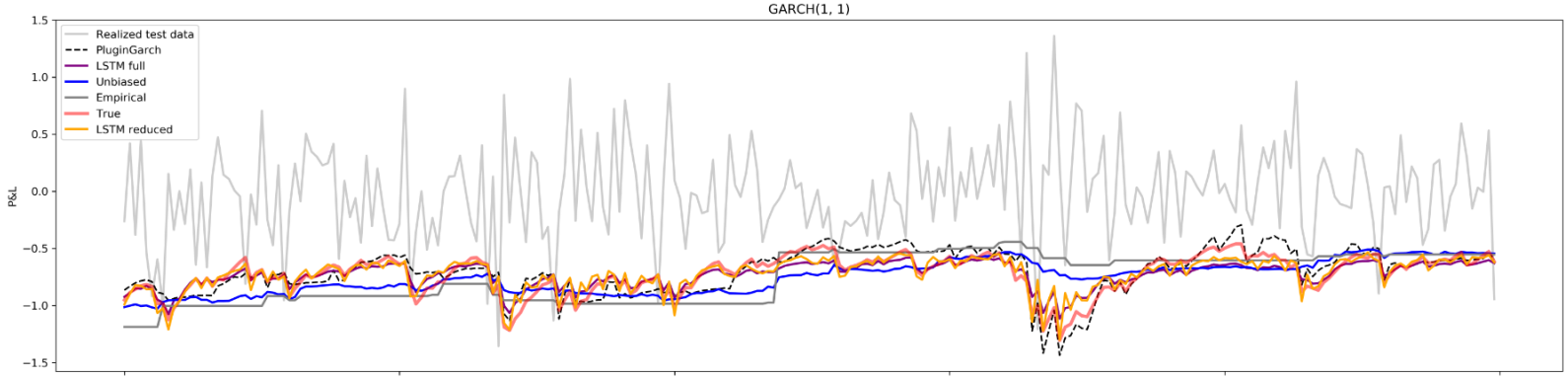
The chosen specifications of the GARCH models used for simulation. We fitted the parameter from a dataset of S&P 500 data.

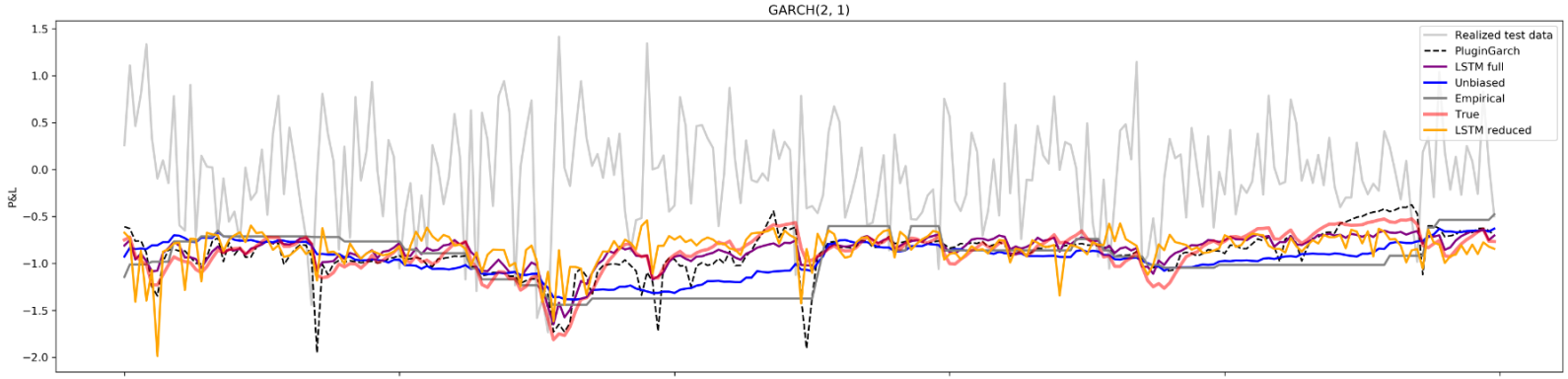

GARCH(1,1)–n data
True risk
Empirical
Gaussian unbiased
LSTM full
LSTM reduced
P&L
2
1
0
−1
−2
0
50
100
150
200
250
scenario number

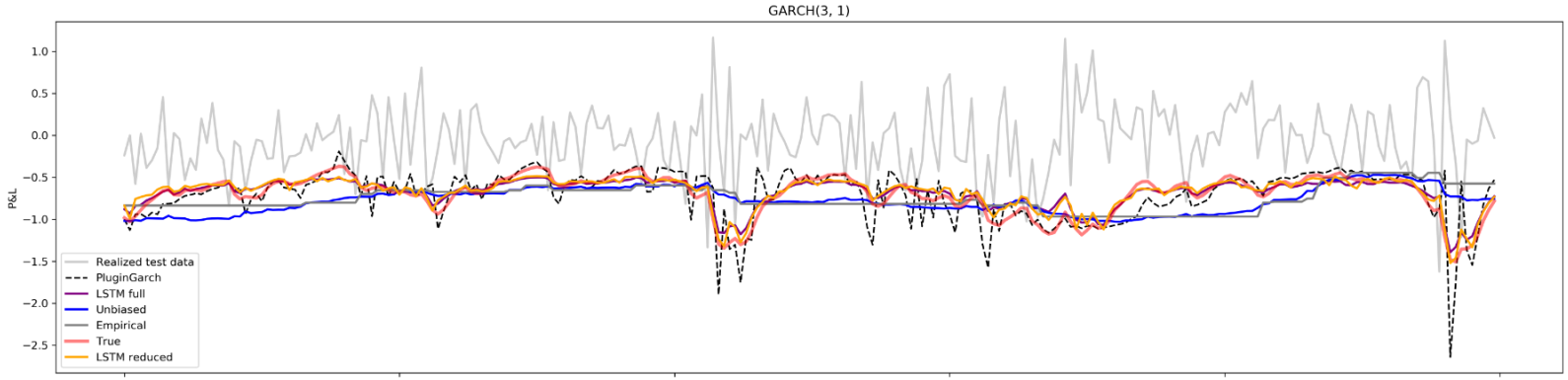

GARCH(1,1)–n data
True risk
Empirical
Gaussian unbiased
LSTM full
LSTM reduced
P&L
2
1
0
−1
−2
0
50
100
150
200
250
scenario number

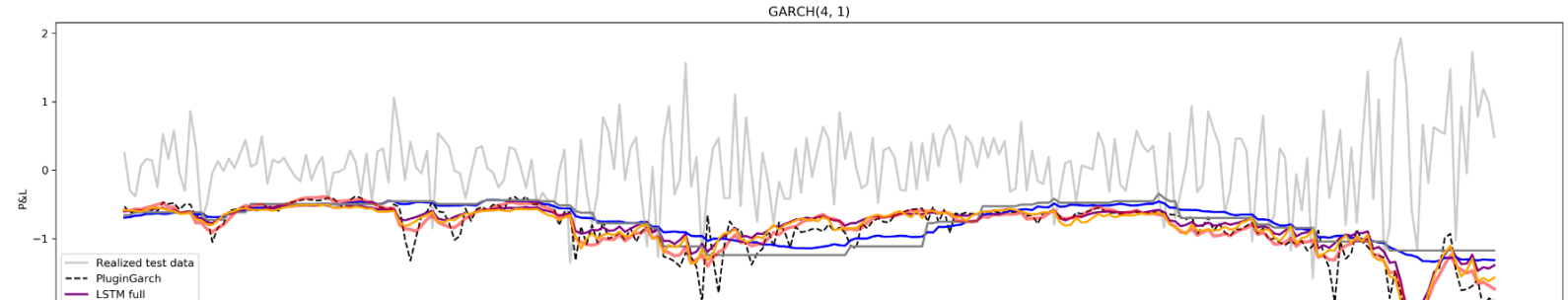
- This is actually a remarkable result:
- Note that the usual procedure is a numerical quasi-maximum-likelihood estimation
- and then the plug-in estimation of the value-at-risk.
- We call this the plugin-GARCH estimator and use an approximately unbiased version of it:

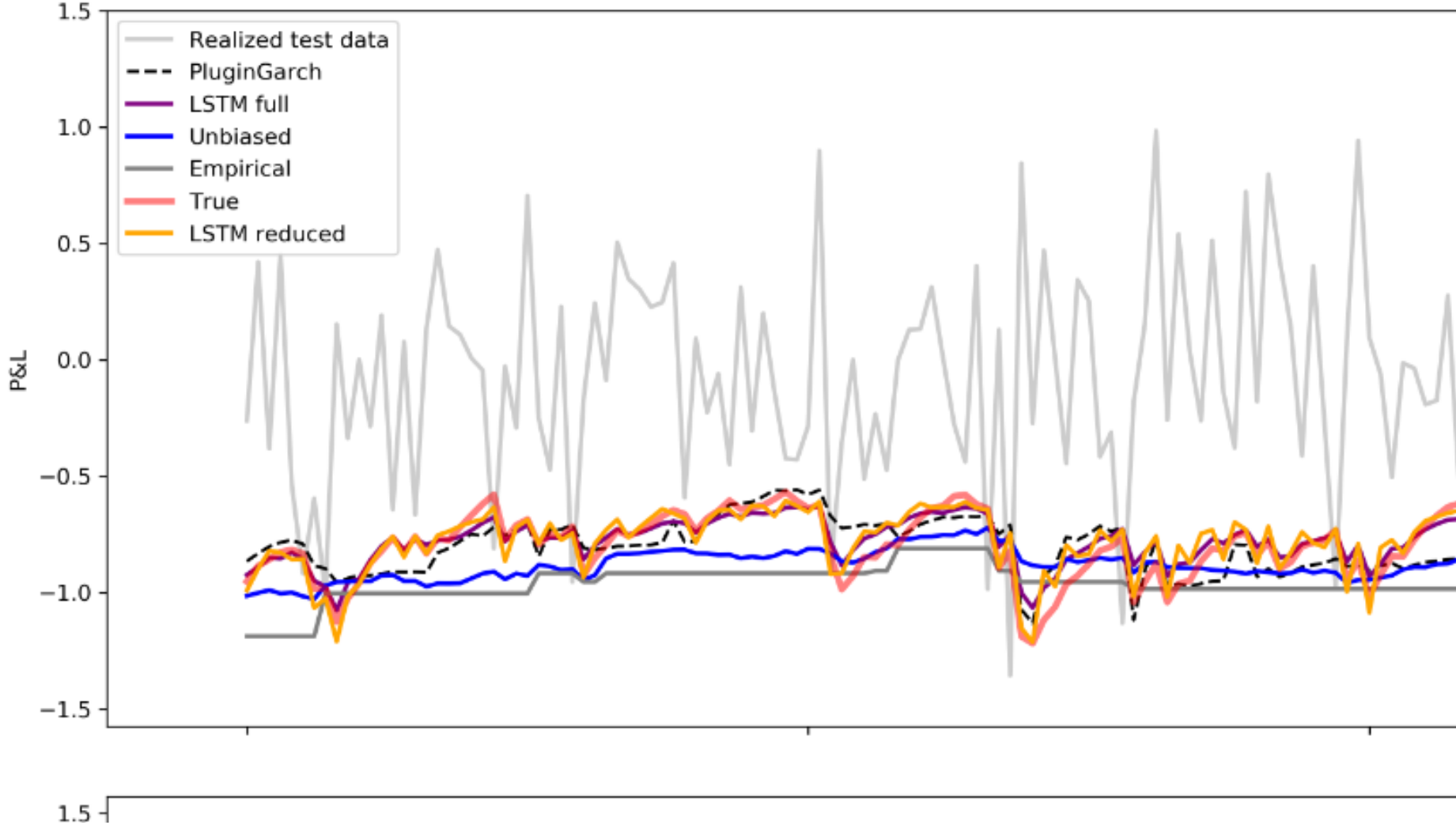
$$\overset{\text{GARCH, u}}{\underset{\alpha}{\nwarrow}}(X^i) := \hat{\sigma}^i \sqrt{\frac{n+1}{n}} t_{n-1}^{-1}(\alpha). \tag{16}$$


GARCH(1, 1)
P&L
Realized test data
PluginGarch
LSTM full
Unbiased
Empirical
True
LSTM reduced


GARCH(2, 1)
P&L
Realized test data
PluginGarch
LSTM full
Unbiased
Empirical
True
LSTM reduced


GARCH(3, 1)
P&L
Realized test data
PluginGarch
LSTM full
Unbiased
Empirical
True
LSTM reduced


GARCH(4, 1)
P&L
Realized test data
PluginGarch
LSTM full


Realized test data
PluginGarch
LSTM full
Unbiased
Empirical
True
LSTM reduced
P&L
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Model	(true)	$\bar{S}_\alpha$ lstm	emp	u	plugin$\mathscr{G}$
$\mathscr{G}$(1,1)-n	(0.051)	**0.052**	0.056	0.056	0.053
$\mathscr{G}$(1,1)-t	(0.050)	**0.051**	0.059	0.056	0.055
$\mathscr{G}$(2,1)-n	(0.052)	**0.053**	0.059	0.057	0.054
$\mathscr{G}$(2,1)-t	(0.056)	**0.056**	0.064	0.062	0.059
$\mathscr{G}$(3,1)-n	(0.047)	**0.047**	0.054	0.053	0.049
$\mathscr{G}$(3,1)-t	(0.049)	**0.052**	0.061	0.058	0.053
$\mathscr{G}$(4,1)-n	(0.048)	**0.049**	0.054	0.053	0.051
$\mathscr{G}$(4,1)-t	(0.053)	**0.057**	0.062	0.060	0.058

- Based on 100.000 simulations the LSTM is able to sufficiently learn the structure of the time series
- And comes up with an estimator which is very close to the true risk.
- Heavier tails and more complex structure makes it more difficult to learn the model (as expected).

Conclusion

- We studied the estimation of risk, with a particular view on **unbiased** estimators and backtesting.
- Unfortunately, in many situations unbiased estimators can not be computed but need to be targeted numerically.
- By using a supervised learning approach we are able to construct an estimator generically in a wide range of time series scenarios
- Our results show that the estimator outperforms all existing approaches in the GARCH setting.
- On data examples the LSTM even outperforms the GARCH estimators


Artzner, Philippe et al. (1999). „Coherent measures of risk“. In: **Mathematical finance** 9.3, pp. 203–228.


Föllmer, Hans and Alexander Schied (2011). **Stochastic finance: an introduction in discrete time**. Walter de Gruyter.


Pitera, Marcin and Thorsten Schmidt (2018). „Unbiased estimation of risk“. In: **Journal of Banking & Finance** 91, pp. 133–145.