Machine Learning for Stochastics
Risk

Thorsten Schmidt

Abteilung für Mathematische Stochastik

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

SS 2025

Risk in Finance

What is Risk?

- "Hazard, danger; exposure to mischance or peril. " (Oxford English Dictionary)
- "any event or action that may adversely affect an organization's ability to achieve its objectives and execute its strategies"
- "the quantifiable likelihood of loss or less-than-expected returns"

- The modern way to model and capture risk is to use stochastic methods for describing the uncertain future.
- In this way, past experience can be taken into account but uncertainty and riskiness can be captured in an adequate way.

Monetary measures of risk

- We have only short time to dive into this fascinating subject. For more details see Hans Föllmer and Alexander Schied (2011). **Stochastic finance: an introduction in discrete time**. Walter de Gruyter
- A **financial position** (typically a portfolio) is given by

$$X : \Omega \to \mathbb{R}$$

- We think of a space of financial position $\mathscr{X}$, which is a linear space of bounded random variables containing the constants

Definition

A mapping $\rho : \mathscr{X} \to \mathbb{R}$ is called a **monetary risk measures** if for all $X, Y \in \mathscr{X}$

1. **Monotonicity:** $\rho(X) \leq \rho(Y)$ for $X \geq Y$
2. **Cash invariance:** $\rho(X + c) = \rho(X) - c$ for all $c \in \mathbb{R}$.

1. **Monotonicity:** $\rho(X) \leq \rho(Y)$ for $X \geq Y$
2. **Cash invariance:** $\rho(X + c) = \rho(X) - c$ for all $c \in \mathbb{R}$.
3. Often we add **Normalization:** $\rho(0) = 0$
4. We have the useful property

$$\rho(X + \rho(X)) = 0$$

   such that $\rho(X)$ can be used as reserve to make risks acceptable.
5. **Convexity**: $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$.
6. **Positive homogeneity**: $\rho(cX) = c\rho(X)$.
7. **Subadditivity**: $\rho(X + Y) \leq \rho(X) + \rho(Y)$.
8. **Coherence**: A monetary risk measure is called coherent if it is positive homogeneous and subbadditive (can be replaced by convexity)

VaR and Expected shortfall

- The most common risk measures are value-at-risk and expected shortfall
- $\mathrm{VaR}_\alpha = \inf\{c \in \mathbb{R} \colon P(X + c < 0) \leq \alpha\}$.
- If X is Gaussian, then

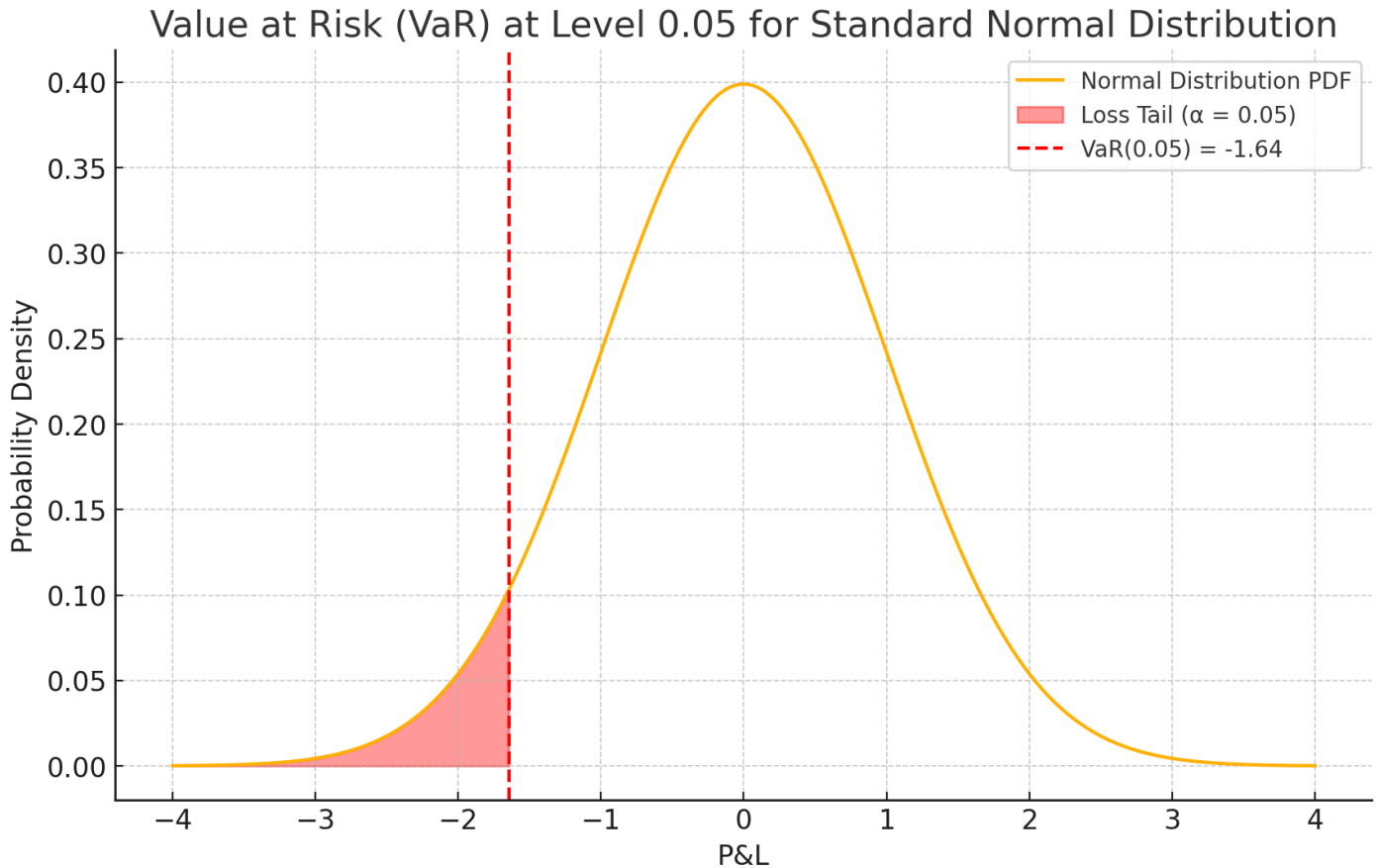
$$\mathrm{VaR}_\alpha = -E[X] + \Phi^{-1}(1-\alpha)\sigma(X).$$

- The expected shortfall (or average value at risk) is defined by

$$\mathrm{ES}_\alpha = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\alpha(X)d\alpha$$

It falls in the class of **spectral risk measures**, which are given as integrals over quantiles

VaR in Visual Terms



Expected Shortfall

For **continuous** loss distributions expected shortfall is the expected loss, given that the VaR is exceeded. For any $\alpha \in (0, 1)$ we have

$$\mathrm{ES}_\alpha = E[-X \mid -x \geq \mathrm{VaR}_\alpha]\,. \tag{1}$$

- In general, value-at-risk is not coherent (because it is not always subadditive)
- Expected shortfall is a coherent risk measure
- It is often important to acknowledge that a risk measure is not known and needs to be estimated - which we will discuss in the following

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

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

- **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 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?

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%.

Table: Estimates of $\mathrm{VaR}_{0.05}$ for NASDAQ100 (first column) and for a sample from normally distributed random variable with mean and variance fitted to the NASDAQ data (second column), both for 4.000 data points. **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

- 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.

Measuring risk

We begin with well-known results on the measurement of risk, see A. McNeil, R. Frey, and P. Embrechts (2015). **Quantitative Risk Management: Concepts, Techniques and Tools**. Princeton University Press.

- 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 $X \in L^0$

$$\rho_\theta(X) = R(F_X(\theta)), \tag{3}$$

$F_X(\theta) = P_\theta(X \leq \cdot)$ 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 $\theta \in \Theta$,

$$\rho_\theta(X + \hat{\rho}_n) = 0. \tag{4}$$

- 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 (4) 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.

Unbiased estimation of value-at-risk under normality

- Let $V \sim \mathcal{N}(\theta_1, \theta_2^2)$ and denote $\theta = (\theta_1, \theta_1) \in \Theta = \mathbb{R} \times \mathbb{R}_{>0}$.
- The value-at-risk is, where we again set $\alpha = 1 - \alpha'$

$$\rho_\theta(X) = \inf\{x \in \mathbb{R} \colon P_\theta[V + x < 0] \leq \alpha'\}, \qquad \theta \in \Theta, \tag{5}$$

- Unbiasedness as defined in Equation (4) is equivalent to

$$P_\theta[V + \hat{\rho} < 0] = \alpha', \qquad \text{for all } \theta \in \Theta. \tag{6}$$

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

$$\hat{\rho}(v_1, \ldots, v_n) = -\bar{v} - \bar{\sigma}(\boldsymbol{v}) \sqrt{\frac{n+1}{n}} t_{n-1}^{-1}(\alpha'), \tag{7}$$

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

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

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

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

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

Thus, the random variable T is a pivotal quantity and

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

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

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

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

$$\sqrt{\frac{n+1}{n}}\,\frac{t_{n-1}^{-1}(\alpha)}{\Phi^{-1}(\alpha)}. \tag{8}$$

Machine learning

- If we have a large enough sample, we expect deep neural networks can learn an efficient estimation.
- Yes, but ....

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.

LSTM

- Introduced in Sepp Hochreiter and Jürgen Schmidhuber (1997). „Long short-term memory“. In: **Neural computation** 9.8, pp. 1735–1780
- LSTMs are **recurrent neural networks** (means that they are the right tool for time-series)
- They sequentially take the inputs x_1 (then feed it into the network, and combine the output with) x_2 (feed it into the network ... and so on)
- The LSTM has certain components to capture hidden structures in the time series (right adopted to GARCH) and the time is to short to explain the details....

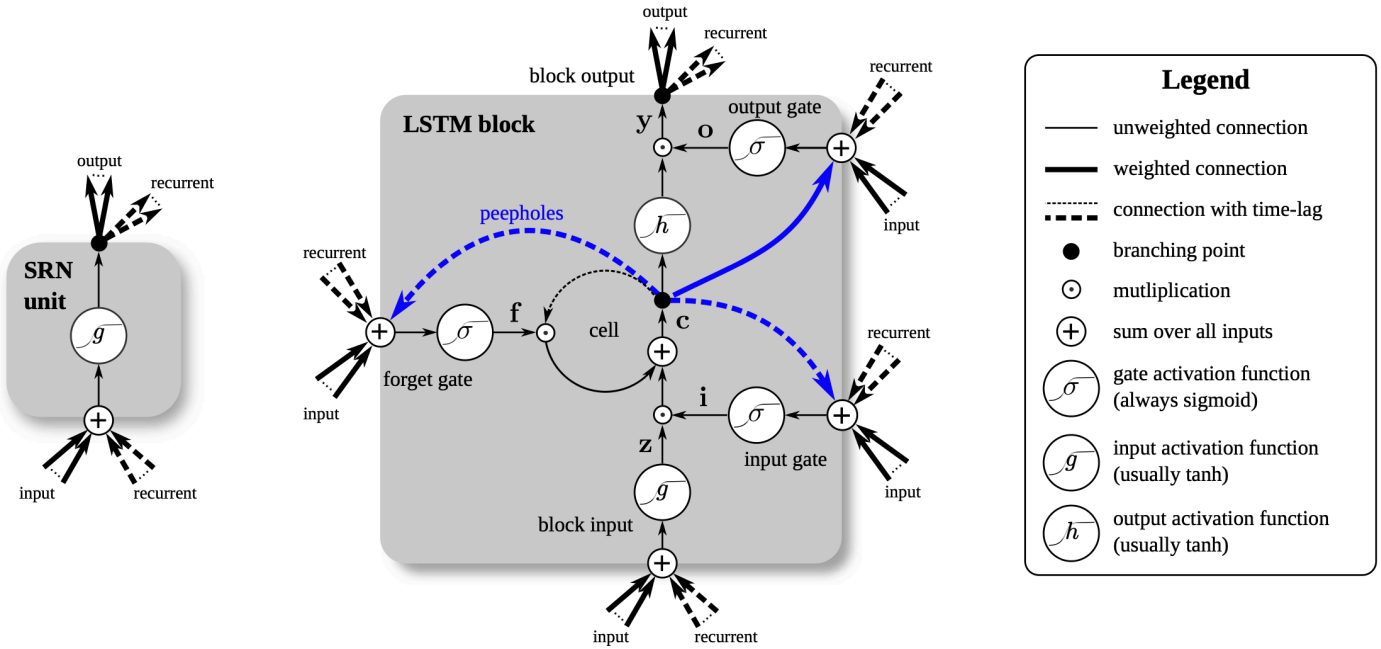


Figure 1. Detailed schematic of the Simple Recurrent Network (SRN) unit (left) and a Long Short-Term Memory block (right) as used in the hidden layers of a recurrent neural network.

- 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

$$\overset{k}{h}(\boldsymbol{x}) = \sigma^k(\boldsymbol{a}_h^k + \boldsymbol{x} A_h^k),$$

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

- Given $\overset{h}{j-1}$, we set

$$\overset{h}{j} := o_j^h \odot \sigma(c_j^h), \quad \text{for } h = 1, 2, \dots, H.$$

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

$$c_j^h := f_j^h \odot c_{j-1}^h + i_j^h \odot g_j^h$$

$$\begin{aligned}
o_j^h &:= \overset{o}{h}([\overset{h}{j-1}, \overset{h-1}{j}]), \\
f_j^h &:= \overset{f}{h}([\overset{h}{j-1}, \overset{h-1}{j}]), \\
i_j^h &:= \overset{i}{h}([\overset{h}{j-1}, \overset{h-1}{j}]), \\
g_j^h &:= \overset{g}{h}([\overset{h}{j-1}, \overset{h-1}{j}]),
\end{aligned}$$

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

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

Elicitability

Gives the right answer which statistic to choose:

- Within the VaR context, the typical choice of a consistent scoring function is the **quantile score** given by

$$S_\alpha(x, y) := (\alpha - \mathbb{1}_{\{y \leq x\}})(y - x). \tag{9}$$

- With the projected risk values at hand, we therefore specify as the objective function the average score with respect to the scoring function S_α, i.e.

$$\begin{aligned} \bar{S}_\alpha(\hat{\rho}(X, A)) &:= \frac{1}{m} \sum_{i=1}^{m} S_\alpha(-\hat{\rho}(\tilde{X}^i, A), Y^i) \\ &= \frac{1}{m} \sum_{i=1}^{m} (\alpha - \mathbb{1}_{\{Y^i + \hat{\rho}(\tilde{X}^i, A) \leq 0\}}) \, (Y^i + \hat{\rho}(\tilde{X}^i, A)). \end{aligned} \tag{10}$$

- The goal of the technical implementation is now to efficiently search for an optimal set or network parameters that produces a sequence of projected risk which minimizes the average score $\bar{S}_\alpha$.

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

$$\tilde{X}^i = \left(\begin{bmatrix} \bar{X}^i \\ \bar{X}^i + \bar{\sigma}^i \Phi^{-1}(\alpha) \\ X^i_{(n\alpha)+1} \\ f_1(X_i - \bar{X}^i) \\ f_2(X_i - \bar{X}^i) \\ f_3(X_i - \bar{X}^i) \\ f_4(X_i - \bar{X}^i) \end{bmatrix}, \begin{bmatrix} \bar{X}^i \\ \bar{X}^i + \bar{\sigma}^i \Phi^{-1}(\alpha) \\ X^i_{(n\alpha)+1} \\ f_1(X_{i+1} - \bar{X}^i) \\ f_2(X_{i+1} - \bar{X}^i) \\ f_3(X_{i+1} - \bar{X}^i) \\ f_4(X_{i+1} - \bar{X}^i) \end{bmatrix}, \ldots, \begin{bmatrix} \bar{X}^i \\ \bar{X}^i + \bar{\sigma}^i \Phi^{-1}(\alpha) \\ X^i_{(n\alpha)+1} \\ f_1(X_{i+n-1} - \bar{X}^i) \\ f_2(X_{i+n-1} - \bar{X}^i) \\ f_3(X_{i+n-1} - \bar{X}^i) \\ f_4(X_{i+n-1} - \bar{X}^i) \end{bmatrix} \right), \quad (11)$$

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

$$\bar{S}_\alpha = \frac{1}{m}\sum_{i=1}^{m}(\alpha - \mathbb{1}_{\{Y^i+\hat{\rho}(X^i)\leq 0\}})\,(Y^i + \hat{\rho}(X^i)), \tag{12}$$

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

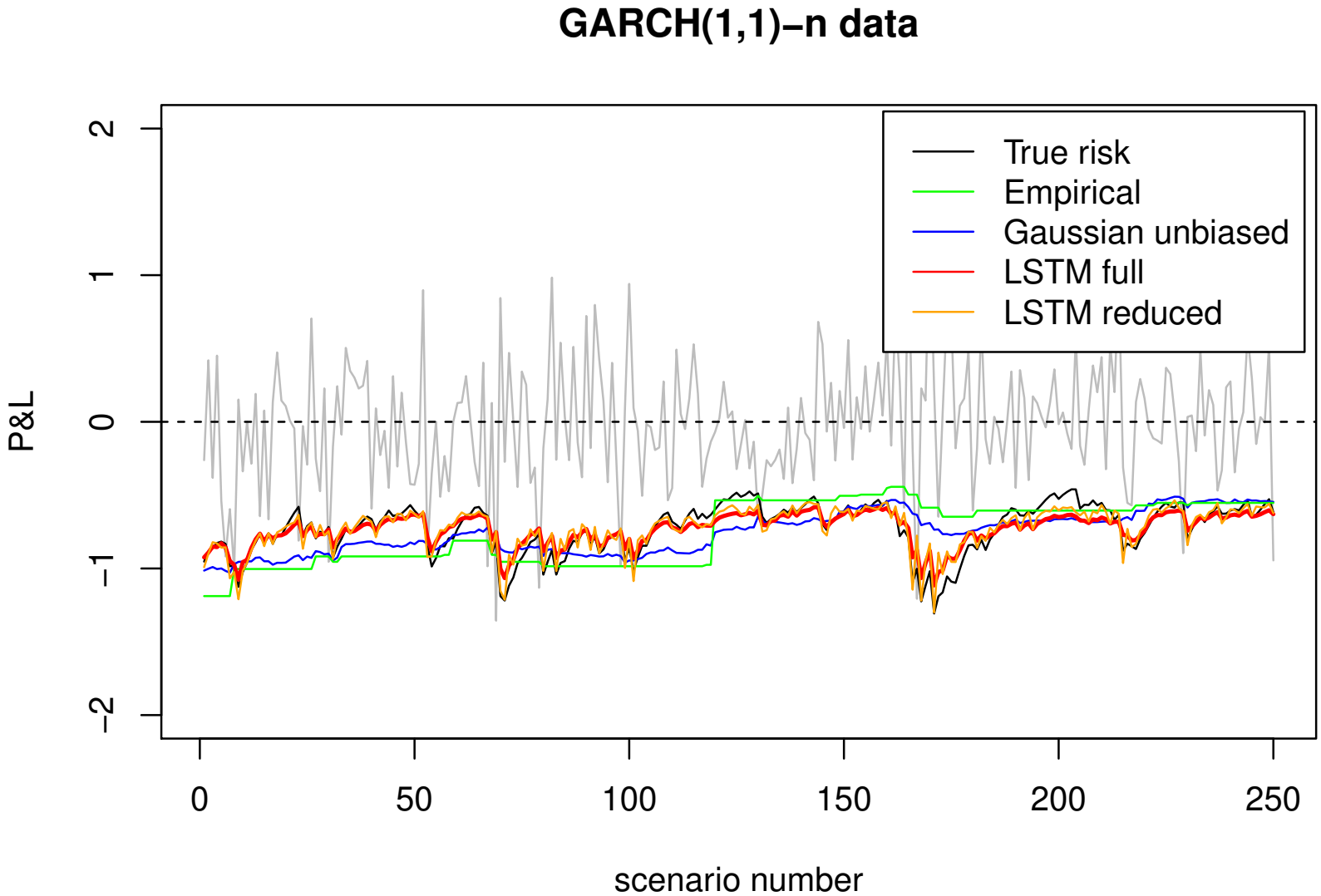
- Probably much more suited for financial time series is a GARCH model

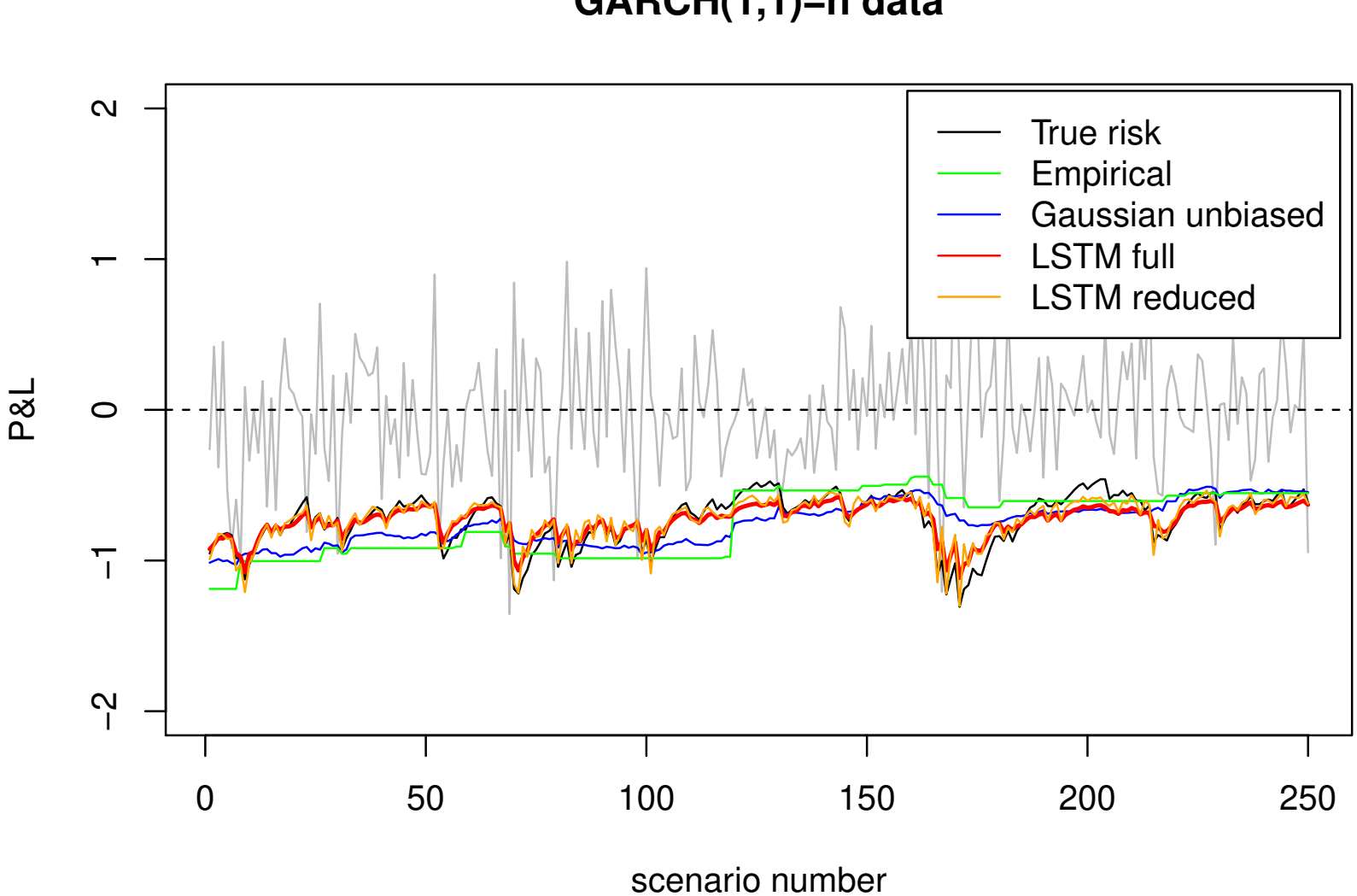
$$\begin{cases} r_t & = \sigma_t \epsilon_t, \\ \sigma_t^2 & = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \end{cases}, \tag{13}$$

- 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	$t(0,1,5)$

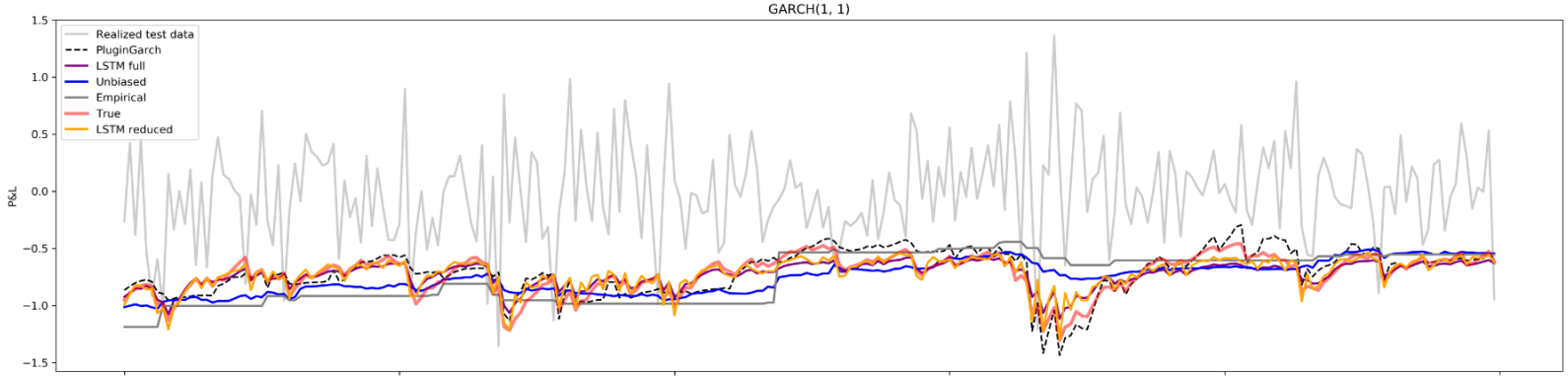
The chosen specifications of the GARCH models used for simulation. We fitted the parameter from a dataset of S&P 500 data.

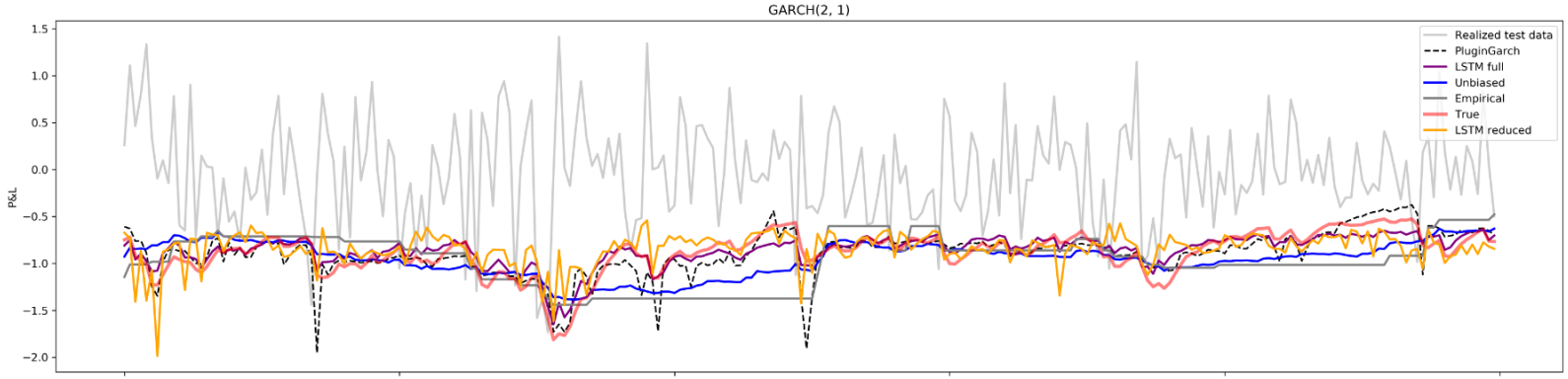

GARCH(1,1)–n data
True risk
Empirical
Gaussian unbiased
LSTM full
LSTM reduced
P&L
2
1
0
−1
−2
0
50
100
150
200
250
scenario number

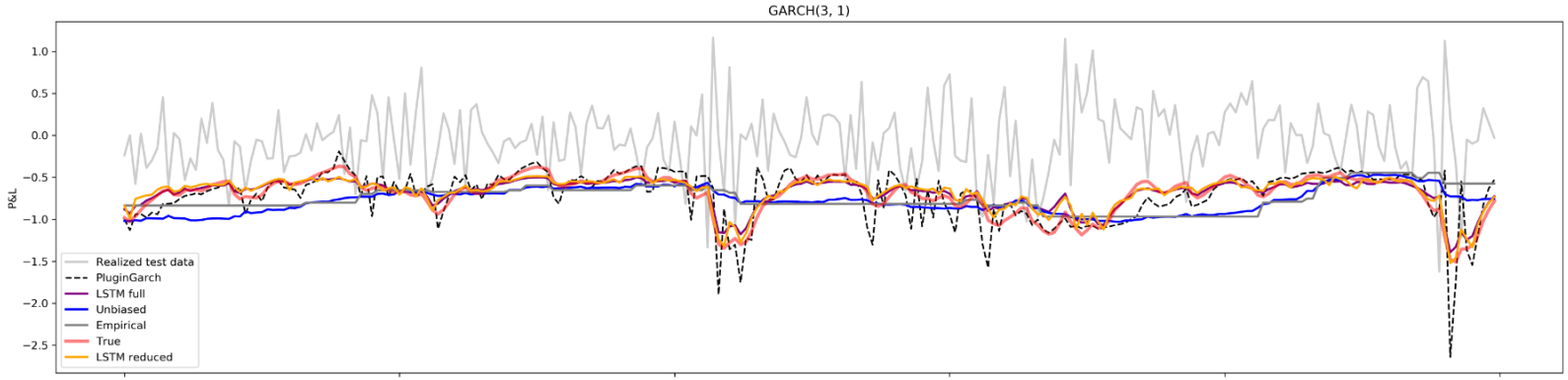

GARCH(1,1)–ll data
True risk
Empirical
Gaussian unbiased
LSTM full
LSTM reduced
P&L
2
1
0
−1
−2
0
50
100
150
200
250
scenario number

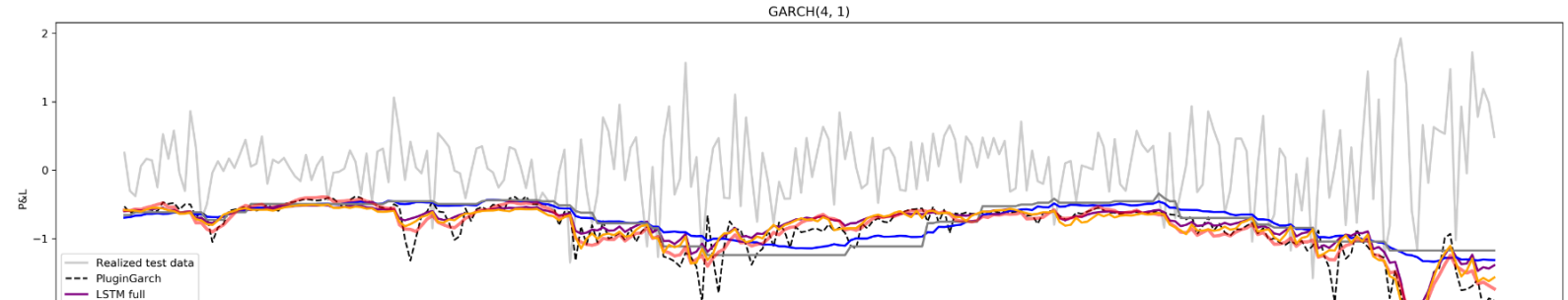
- This is actually a remarkable result:
- Note that the usual procedure is a numerical quasi-maximum-likelihood estimation
- and then the plug-in estimation of the value-at-risk.
- We call this the plugin-GARCH estimator and use an approximately unbiased version of it:

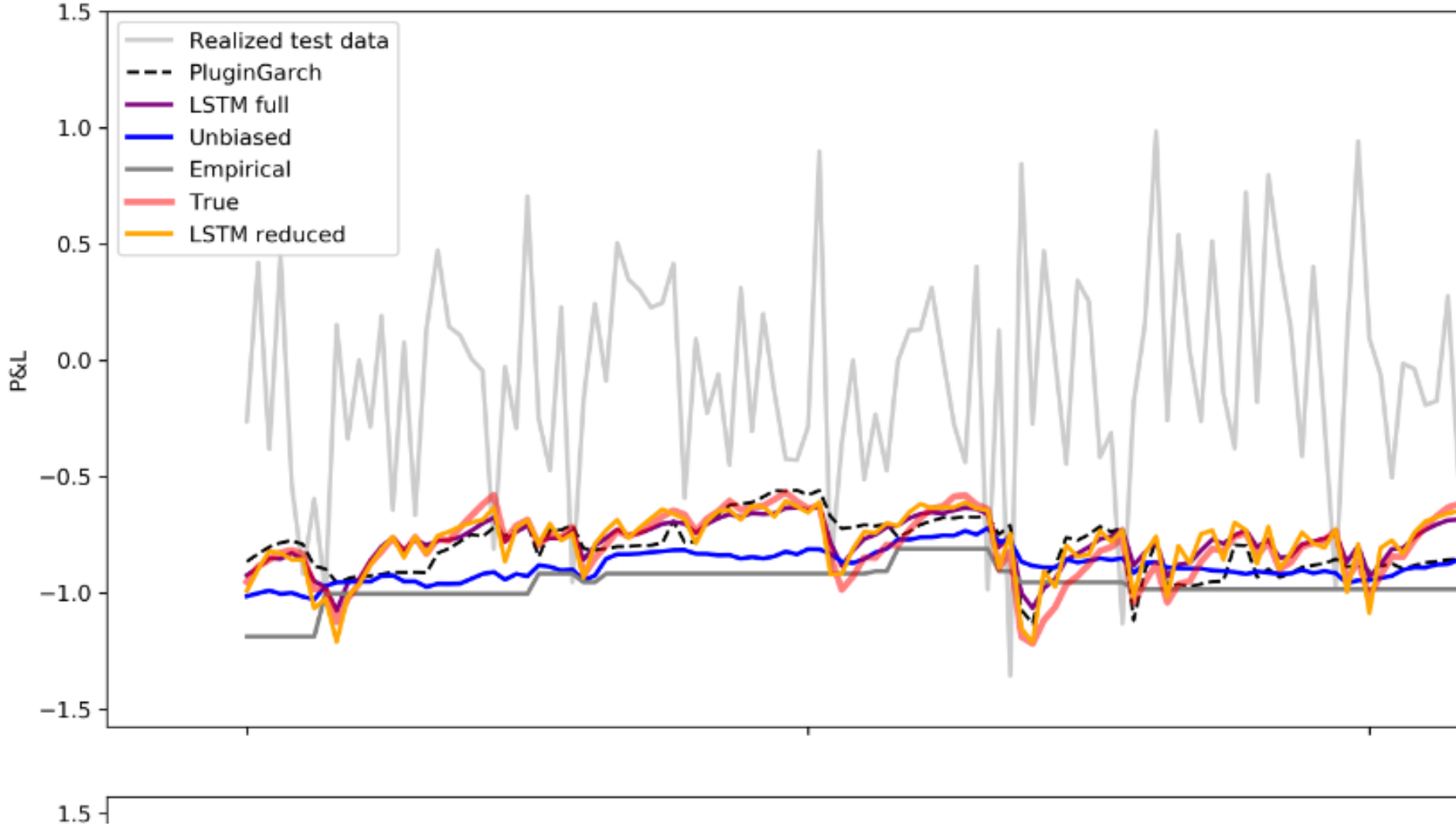
$$\overset{\curvearrowleft}{\alpha}^{\mathrm{GARCH,\ u}}(X^i) := \hat{\sigma}^i \sqrt{\frac{n+1}{n}} t_{n-1}^{-1}(\alpha). \tag{14}$$


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


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


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


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


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
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).
- Data examples are currently being developed.

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.
- Future work: application to real data (which will be small).