








Statistics of Financial Markets

Chapter 6: Model selection and risk management



Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Acronyms used in this material

- **AD**: Anderson–Darling (test)
- **AI**: Artificial Intelligence
- **AIC**: Akaike Information Criterion
- **BET**: Bucharest Exchange Trading (index)
- **BIC**: Bayesian Information Criterion
- **BVB**: Bursa de Valori București (the Bucharest Stock Exchange)
- **CDF**: Cumulative Distribution Function
- **COVID-19**: COronaVIrus Disease 2019
- **EDF**: Empirical Distribution Function
- **EM**: Expectation–Maximisation (algorithm)
- **EMA**: Exponential Moving Average
- **EODHD**: EOD Historical Data (market-data provider)
- **ES**: Expected Shortfall
- **FFT**: Fast Fourier Transform
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GED**: Generalised Error Distribution
- **KS**: Kolmogorov–Smirnov (test)
- **LR**: Likelihood Ratio
- **ML**: Maximum Likelihood
- **NIG**: Normal Inverse Gaussian (distribution)
- **PP**: Probability–Probability (plot)
- **QQ**: Quantile–Quantile (plot)
- **RMA**: Rectangular Moving Average

Today's Question and Route

- **Question**: we have seven candidate distributions for daily returns; which one should a risk manager use, and how sure can we be?
  - Chapters 2, 3 and 5: the Normal distribution fails; Student-t, stable laws and extreme value theory each describe part of the picture
- **Route** of the chapter
  - the candidates and what they must reproduce; maximum likelihood, recalled
  - comparing models: likelihood-ratio tests, AIC and BIC
  - checking one model: goodness-of-fit tests, PP and QQ plots
  - the purpose decides: which model gets VaR 1% right, in sample and out of sample
  - overfitting and model uncertainty

Learning Outcomes

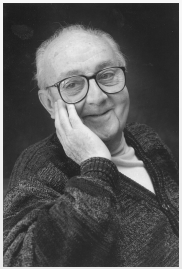
- Write the log-likelihood of a candidate model and explain how maximum likelihood estimates are found
- Test a restricted model against a larger one with the likelihood-ratio test, and know when the test does not apply
- Compute AIC, BIC, ΔAIC and Akaike weights and interpret them
- Run and read KS, Cramér–von Mises and Anderson–Darling tests, PP and QQ plots, and explain why KS is weak in the tails
- Judge a model by its VaR 1% out of sample, recognise overfitting, and report model uncertainty

Reading and Tools

- Textbook: Franke, Härdle & Hafner (2019), Ch. 16 (Value at Risk and backtesting) and Ch. 18 (statistics of extreme risks); exercises: Borak, Härdle & López-Cabrera (2013)
  - distributional properties of returns: Tsay (2010), Ch. 1
- Model selection: Burnham & Anderson (2002); short version Burnham & Anderson (2004)
- Goodness of fit: Stephens (1974); the original papers Anderson & Darling (1952), Lilliefors (1967)
- Python Quantlets of this chapter: Quantlets/Ch_06
  - ported from the Quantlets SFESumm and SFEVaRqqplot; each chart links to the Quantlet that draws it
- Lecture notebook: open in Google Colab
- Video course: Measuring Statistical Risk

"All Models Are Wrong"

- George Box: a statistician should look for a model that is useful, not one that is true Box (1976)
  - "all models are wrong, but some are useful" is the short form of his argument
- For returns, "useful" depends on the purpose
  - pricing and portfolio choice: the body of the distribution (typical days)
  - risk management: the left tail (the worst 1% of days)
- This chapter: a toolbox to compare models, and the habit of asking "useful for what?"



George E. P. Box (1919–2013)

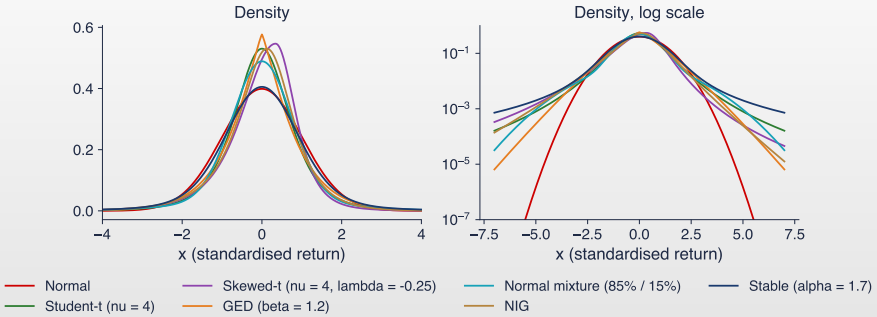
Photo: DavidMCEddy (2011); CC BY-SA 3.0; Wikimedia Commons

Seven Candidate Distributions

Model	k	Tails	Skewness	Reference
Normal	2	thin, $e^{-x^2/2}$	no	Chapter 2
Student-t	3	power, $x^{-\nu}$	no	Chapter 2
skewed-t	4	power, $x^{-\nu}$	yes (λ)	Hansen (1994)
GED	3	$e^{-\lvert x\rvert^{\beta}}$, heavy if $\beta < 2$	no	Nelson (1991)
Normal mixture	5	thin, but a wide component	yes	Kon (1984)
NIG	4	semi-heavy, $\lvert x\rvert^{-3/2}e^{-c\lvert x\rvert}$	yes	Barndorff-Nielsen (1997)
stable	4	power, $x^{-\alpha}$, $\alpha < 2$	yes (β)	Chapter 3

- k: number of parameters; GED: generalised error distribution; NIG: normal inverse Gaussian, a member of the generalised hyperbolic family Eberlein & Keller (1995)
- Normal mixture: with probability w a calm-day $N(\mu_1, \sigma_1^2)$, otherwise a turbulent-day $N(\mu_2, \sigma_2^2)$

Comparison of the Candidate Distributions



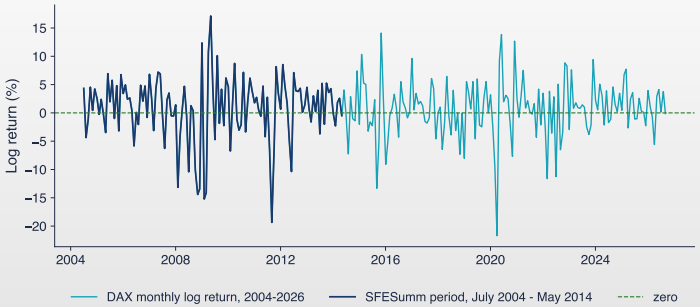
- All with mean 0 and variance 1 (stable: the same interquartile range as the Normal law); right panel on a log scale
- Near the centre they look alike; in the tails they differ by orders of magnitude

SFM_ch6_candidates

Tail Probabilities at Equal Variance

- $P(X < -4)$, a move of four standard deviations, under each candidate of the chart:
  - Normal 3.2×10^{-5}; GED 9.5×10^{-4}; Student-t 2.4×10^{-3}; Normal mixture 2.5×10^{-3}
  - NIG 3.2×10^{-3}; skewed-t 4.5×10^{-3}; stable 8.1×10^{-3}
- The Student-t gives this loss about 76 times the Normal probability
- A risk measure such as VaR 1% lives exactly where the candidates disagree
- **Question for the room**: if all fit the centre well, why not keep the simplest one?
- **Answer**
  - the simplest one (Normal) is wrong precisely in the region that drives capital and margin requirements

Benchmark Statistics for the Models: SFESumm



- Quantlet SFESumm: DAX monthly log returns between first trading days, July 2004 – May 2014 ($n = 119$), extended here to 1 Sep 2026
- Even monthly returns are skewed and heavy-tailed: the next slide gives the numbers

SFM_ch6_summary

Summary Statistics: Monthly DAX and Daily Returns

	n	min	max	st. dev.	ann. vol.	skew.	kurt.
DAX 2004–2014	119	−19.35	17.12	5.80	20.08	−0.83	4.60
DAX 2004–2026	267	−21.70	17.12	5.39	18.66	−0.72	5.06

- Monthly, in %; st. dev.: standard deviation σ; ann. vol.: $\sigma\sqrt{12}$
  - skewness $E(X-\mu)^3/\sigma^3$ and kurtosis $E(X-\mu)^4/\sigma^4$, with μ the mean; 0 and 3 for the Normal distribution
  - Jarque–Bera test of Normality (Chapter 2): p-value < 0.001
- Daily log returns since 2000 (Bitcoin since 2014, BVB stocks since 2010): kurtosis
  - BET 14.1, S&P 500 13.7, DAX 9.1, Bitcoin 15.0, TLV 16.5, SNP 11.9, BRD 10.3
  - skewness: S&P 500 −0.35, BET −0.65, Bitcoin −0.70

SFM_ch6_summary

Data Errors: Two Days in the Banca Transilvania Series

- Banca Transilvania (TLV), adjusted close from EODHD: log returns of -15.0% on 30 May 2016 and 20.5% on 31 May 2016
  - the adjustment for the bonus shares is applied one day late (Chapter 2): a fake crash followed by a fake rally
- Effect of the two days on the fitted model
  - kurtosis 20.5 with them, 16.5 without; Student-t $\hat{\nu}$: 3.33 vs 3.44
  - the AIC ranking does not change here (best: Student-t), but the moments do
- Rule: check the largest returns against the close and the news before any model selection; the two days are removed in the rest of the chapter

Likelihood and Log-Likelihood

- Data $r_1, \ldots, r_n$, assumed i.i.d. (independent, identically distributed) with density $f(r;\theta)$
  - θ: the vector of model parameters (for example μ, σ); f: the density of the model
  - **likelihood**: $L(\theta) = \prod_{t=1}^{n} f(r_t;\theta)$, the probability density of the observed sample as a function of θ
  - **log-likelihood**: $\ell(\theta) = \sum_{t=1}^{n} \ln f(r_t;\theta)$; sums are easier than products
- **Maximum likelihood (ML) estimate**: $\hat{\theta} = \arg\max_\theta \ell(\theta)$
  - $\ell(\hat{\theta})$, the maximised log-likelihood, is the raw material of every comparison in this chapter
- Large n: $\hat{\theta}$ is approximately Normal around the true θ
  - covariance $\approx [-\partial^2 \ell / \partial\theta\, \partial\theta^\top]^{-1}$, the inverse of the observed information
  - a sharply curved ℓ (large second derivative) means a precise estimate

Worked Example: the Normal Model, Step by Step

- $\ell(\mu, \sigma) = -\dfrac{n}{2}\ln(2\pi\sigma^2) - \dfrac{1}{2\sigma^2}\sum_t (r_t - \mu)^2$
  - setting the derivatives to zero: $\hat{\mu} = \bar{r}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_t (r_t - \bar{r})^2$ (divisor n, not $n-1$)
  - plugging back: $\ell(\hat{\theta}) = -\dfrac{n}{2}\big[\ln(2\pi\hat{\sigma}^2) + 1\big]$
- S&P 500, daily log returns in %, 2000–2026: $n = 6\,714$, $\hat{\mu} = 0.025$, $\hat{\sigma} = 1.213$
  - $\ln(2\pi\hat{\sigma}^2) = 2.2240$, so $\ell(\hat{\theta}) = -\frac{6\,714}{2}(2.2240 + 1) = -10\,823$
- Only the Normal model has a closed form; the others are maximised numerically

Numerical ML for the Other Candidates (S&P 500)

Model	Estimates	$\ell(\hat{\theta})$
Normal	$\hat{\mu} = 0.025$, $\hat{\sigma} = 1.213$	$-10\,823$
Student-t	$\hat{\nu} = 2.67$, location 0.068, scale 0.703	$-9\,848$
skewed-t	$\hat{\nu} = 2.69$, $\hat{\lambda} = -0.067$, $\hat{\mu} = 0.024$, $\hat{\sigma} = 1.39$	$-9\,838$
GED	$\hat{\beta} = 0.88$, location 0.068, scale 0.653	$-9\,854$
Normal mixture	$\hat{w} = 79\%$: $N(0.089, 0.71^2)$; 21%: $N(-0.223, 2.26^2)$	$-9\,925$
NIG	$\hat{a} = 0.403$, $\hat{b} = -0.048$, location 0.115, scale 0.755	$-9\,824$
stable	$\hat{\alpha} = 1.540$, $\hat{\beta} = -0.181$, $\hat{\gamma} = 0.582$, $\hat{\delta}_0 = 0.093$	$-9\,893$

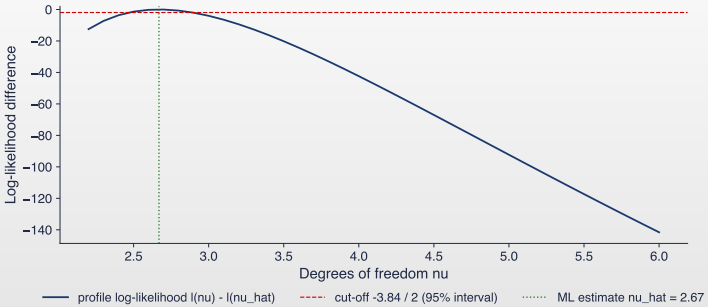
- Optimiser: Nelder–Mead on transformed parameters (for example $\ln\sigma$, $\ln(\nu - 2)$); the Normal mixture by the EM (expectation–maximisation) algorithm; the stable density by FFT (Chapter 3)
- Skewed-t: μ and σ are the mean and the standard deviation (Hansen's parameterisation)

SFM_ch6_ml_fits

Reading the Estimates

- Student-t $\hat{\nu} = 2.67$: tails like $x^{-2.67}$; the variance exists ($\nu > 2$), the kurtosis does not ($\nu < 4$)
  - the sample kurtosis 13.7 estimates a quantity that the fitted model says is infinite
- Skewed-t $\hat{\lambda} = -0.067 < 0$: a slightly longer left tail (crashes)
- GED $\hat{\beta} = 0.88$: below 2 (Normal) and even below 1 (Laplace): a sharp peak
- Normal mixture: 79% calm days with standard deviation 0.71%, 21% turbulent days with 2.26% and a negative mean
- Stable $\hat{\alpha} = 1.540$ (Chapter 3): heavier tails than any other candidate

Precision of $\hat{\nu}$: the Profile Log-Likelihood



- For each ν, maximise ℓ over location and scale; at $\nu = 5$ the curve is 92 below its maximum
- 95% interval: all ν with $2[\ell(\hat{\nu}) - \ell(\nu)] \leq 3.84$, here $[2.47, 2.89]$ around $\hat{\nu} = 2.67$: the tails are pinned down well

SFM_ch6_ml_fits

Numerical Problems in Maximum Likelihood

- **Local maxima**: the optimiser can stop on a hill that is not the highest
  - remedy: several starting values (the mixtures in this chapter use 4 to 12 starts)
- **Boundaries**: $\nu \to 2$, a mixture weight $\to 0$, $|\lambda| \to 1$
  - the usual standard errors and tests are not valid at a boundary (next section)
- **Parameterisation**: the same law, different parameters
  - scipy's Student-t scale is not the standard deviation: $\mathrm{sd} = \mathrm{scale}\sqrt{\nu/(\nu-2)}$
- **Check**: compare $\ell(\hat{\theta})$ of nested models; a larger model can never have a lower maximum

Nested Models

- Model 0 is **nested** in model 1 if model 0 is model 1 with some parameters fixed
  - Normal inside GED: $\beta = 2$
  - Student-t inside skewed-t: $\lambda = 0$
  - Normal inside Student-t: $\nu = \infty$, i.e. $1/\nu = 0$, on the **edge** of the parameter space
- Not nested: Student-t and GED, Student-t and stable, skewed-t and NIG
  - neither is a special case of the other: they need other tools (Vuong test, AIC)
- Nested: $\ell_1(\hat{\theta}_1) \geq \ell_0(\hat{\theta}_0)$ always; the question is whether the gain is larger than chance

The Likelihood-Ratio Test

- H_0: model 0 (restricted) is true; H_1: model 1 (larger)
  - $LR = 2\left[\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)\right] \geq 0$: twice the gain in maximised log-likelihood
  - ℓ_0, ℓ_1: the log-likelihoods of the two models; k_0, k_1: their numbers of parameters
- **Wilks' theorem** (Wilks (1938)): under H_0, $LR \approx \chi^2(d)$, with $d = k_1 - k_0$ restrictions
  - $\chi^2(d)$: the chi-square distribution with d degrees of freedom
  - valid if the restricted values are inside the parameter space
- Decision at 5%: reject H_0 if $LR > \chi^2_{0.95}(d)$
  - $\chi^2_{0.95}(1) = 3.84$, $\chi^2_{0.95}(2) = 5.99$
- Intuition: twice the log of how many times more likely the data are under the larger model

Worked Example: Testing the Skewness Parameter

- S&P 500: $\ell(\text{Student-t}) = -9\,848$, $\ell(\text{skewed-t}) = -9\,838$
  - $LR = 2(-9\,838 - (-9\,848)) = 19.63 > 3.84$; $p < 0.001$: reject $\lambda = 0$
- The same test elsewhere
  - DAX: $LR = 20.92$, $p < 0.001$: skewness needed
  - BET: $LR = 0.35$, $p = 0.555$; Bitcoin: $LR = 0.42$, $p = 0.515$: no evidence of skewness
- Normal inside GED ($\beta = 2$): LR between 1 431 (Bitcoin) and 2 298 (BET); the Normal model is rejected everywhere
- Interpretation: the two developed indices have a longer left tail; the BET and Bitcoin do not

SFM_ch6_lr_aic

A Boundary Case: Normal inside Student-t

- H_0: $1/\nu = 0$ is on the edge of the space $1/\nu \geq 0$: Wilks' theorem does not apply
  - Self & Liang (1987): the null law is a 50:50 mixture of $\chi^2(0)$ (a point mass at 0) and $\chi^2(1)$
  - 5% critical value: $\chi^2_{0.90}(1) = 2.71$ instead of 3.84; the naive p-value is twice the correct one
- Real data: $LR = 1\,949$ (S&P 500), $1\,337$ (Bitcoin): the Normal model is rejected with either critical value
- The same issue appears for a mixture weight $w = 0$ and for variance parameters equal to 0
- **What do you think?** Does the boundary matter when LR is in the thousands?
- **Answer**
  - no for the decision; yes for small samples and borderline LR values, where it halves the p-value

Non-Nested Models: the Vuong Test

	Student-t vs GED	Student-t vs stable	skewed-t vs NIG
BET	3.70 (< 0.001)	9.33 (< 0.001)	-1.34 (0.179)
S&P 500	0.40 (0.686)	6.52 (< 0.001)	-2.20 (0.028)
DAX	-0.23 (0.817)	7.66 (< 0.001)	-2.97 (0.003)
Bitcoin	-3.80 (< 0.001)	16.47 (< 0.001)	-6.76 (< 0.001)

- Vuong (1989): $m_t = \ln f_1(r_t) - \ln f_2(r_t)$, $V = \sqrt{n}\,\bar{m}/s_m \approx N(0,1)$ if both models are equally close to the truth
  - f_1, f_2: the two fitted densities; m_t: by how much model 1 explains day t better; $\bar{m}$, s_m: mean and standard deviation of the m_t
  - $V > 1.96$: the first model is closer; $V < -1.96$: the second; (p-value) in brackets
- Student-t beats the stable law everywhere; NIG beats the skewed-t on the S&P 500 and the DAX; Student-t vs GED splits by market

SFM_ch6_lr_aic

Akaike's Approach: the Kullback–Leibler Divergence

- Hirotugu Akaike asked: which model will predict **new** data best?
  - the distance from the truth g to a model f: the Kullback–Leibler divergence Kullback & Leibler (1951)
  - $KL(g, f) = E_g[\ln g(R) - \ln f(R)] \geq 0$, zero only if $f = g$
  - R: a new return drawn from the truth g; E_g: the mean under g
- $\ell(\hat{\theta})$ is too optimistic: the same data choose $\hat{\theta}$ and judge it
  - the optimism is about k, the number of parameters Akaike (1974)
- Result: a penalty of 2 per parameter corrects the optimism; the criterion, AIC, is on the next slide



Hirotugu Akaike (1927–2009)

Photo: The Institute of Statistical Mathematics; CC BY-SA 4.0; Wikimedia Commons

AIC and BIC

- **AIC** (Akaike information criterion): $\mathrm{AIC} = -2\ell(\hat{\theta}) + 2k$
  - aim: the best predictive model; asymptotically equivalent to leave-one-out cross-validation Stone (1977)
  - small samples: $\mathrm{AICc} = \mathrm{AIC} + 2k(k+1)/(n-k-1)$ Hurvich & Tsai (1989)
- **BIC** (Bayesian information criterion) Schwarz (1978): $\mathrm{BIC} = -2\ell(\hat{\theta}) + k \ln n$
  - aim: the true model, if it is among the candidates (consistency)
  - penalty per parameter $\ln n$: 8.81 for $n = 6\,714$, more than four times the AIC penalty 2
- Both: lower is better; only differences between models on the **same data** mean anything

ΔAIC and Akaike Weights

- $\Delta_i = \mathrm{AIC}_i - \min_j \mathrm{AIC}_j$; rules of thumb of Burnham & Anderson (2004):
  - $\Delta \leq 2$: substantial support; $4 \leq \Delta \leq 7$: considerably less; $\Delta > 10$: essentially none
- **Akaike weight**: $w_i = \dfrac{e^{-\Delta_i/2}}{\sum_j e^{-\Delta_j/2}}$
  - the weight of evidence for model i among the candidates; the weights sum to 1
  - two models with $\Delta = 0$ and $\Delta = 2$: $w = 1/(1 + e^{-1}) \approx 0.73$ and 0.27
- **What do you think?** Is $\Delta = 2$ a "significant" difference?
- **Answer**
  - no: AIC is not a test; a gap of 2 means both models remain plausible (the weights above)

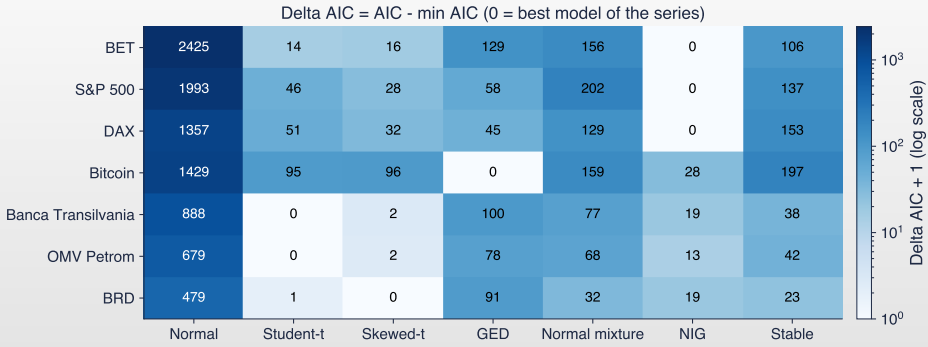
Worked Example: Seven Models for the S&P 500

	k	$\ell(\hat\theta)$	AIC	BIC	ΔAIC	ΔBIC	w
Normal	2	−10 823	21 650	21 663	1993.2	1979.6	0.000
Student-t	3	−9 848	19 702	19 723	45.7	38.9	0.000
skewed-t	4	−9 838	19 685	19 712	28.1	28.1	0.000
GED	3	−9 854	19 714	19 735	57.9	51.1	0.000
Normal mixture	5	−9 925	19 859	19 893	202.5	209.3	0.000
NIG	4	−9 824	19 657	19 684	0.0	0.0	1.000
stable	4	−9 893	19 794	19 821	137.3	137.3	0.000

- Example: NIG, $\mathrm{AIC} = -2(-9\,824) + 2 \cdot 4 = 19\,657$
- NIG wins on both criteria; the next model, skewed-t, has $\Delta = 28.1$, so $e^{-\Delta/2} \approx 8.1 \times 10^{-7}$: weight $w \approx 1$ for NIG
- Normal: ΔAIC of 1 993; the stable law: 137

SFM_ch6_lr_aic

ΔAIC across Seven Series



- Each row: one series; 0 marks the best model; darker cells: worse models (log colour scale)

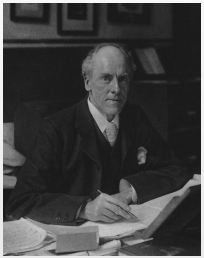
SFM_ch6_lr_aic

Reading the ΔAIC Map

- No single winner: the best model depends on the market
  - indices (BET, S&P 500, DAX): NIG; Bitcoin: GED; BVB stocks: Student-t (TLV, SNP) and skewed-t (BRD)
- Robust conclusions
  - the Normal model is last everywhere, by 479 (BRD) to 2 425 (BET) points
  - the stable law (Δ from 23 to 197) and the Normal mixture are never competitive
- **Question for the room**: for BRD, AIC prefers the skewed-t (Δ of the Student-t: 1.0); which model does BIC prefer?
- **Answer**
  - the Student-t: one parameter less saves $\ln n \approx 8.2$ BIC points, more than the gain in fit; with $\Delta \approx 1$ the two are practically tied

Pearson's χ^2 Test and the Empirical Distribution Function

- Pearson (1900): the first goodness-of-fit test, the χ^2 test
  - group the data into bins and compare observed with expected counts
  - the result depends on the bins, and the tails hold few observations
- **Empirical distribution function** (EDF): $F_n(x) = \frac{1}{n}\#\{t : r_t \leq x\}$
  - no bins: compare F_n with the model distribution function F
- Three EDF tests: Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling Stephens (1974)

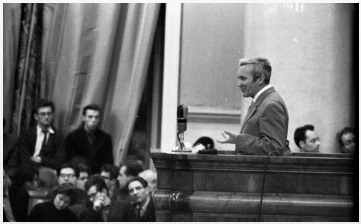


Karl Pearson (1857–1936)

Photogravure (1910); public domain; Wikimedia Commons

The Kolmogorov–Smirnov Test

- $D_n = \sup_x |F_n(x) - F(x)|$, the largest vertical gap between the EDF and the model CDF (sup: the largest value over all x)
  - $r_{(i)}$: the i-th smallest return; $u_{(i)} = F(r_{(i)})$: its model probability
  - computing formula: $D_n = \max_i \max\{i/n - u_{(i)},\ u_{(i)} - (i-1)/n\}$
- F fully known: the law of D_n does not depend on F (Kolmogorov, 1933)
  - large n: reject at 5% if $D_n > 1.358/\sqrt{n}$; tables for small n: Smirnov (1948)
  - $n = 6\,714$: critical value 0.017
- Andrey Kolmogorov also gave probability its axioms (Chapter 4)



Andrey Kolmogorov (1903–1987) lecturing, 1963

Photo: Vsevolod Tarasevich (1963–1964); CC BY 4.0; Wikimedia Commons

Worked Example: KS on Five Returns

- Returns (%): $-2.1, -0.4, 0.3, 0.9, 1.6$; H_0: $N(0,1)$, fully specified
  - $u_{(i)} = \Phi(r_{(i)})$: 0.018, 0.345, 0.618, 0.816, 0.945 (Φ: the standard Normal distribution function)
  - i/n: $0.2, 0.4, 0.6, 0.8, 1.0$; the largest gap is $u_{(3)} - 2/5 = 0.218$
- $D_5 = 0.218 < 0.563$, the exact 5% critical value for $n = 5$: do not reject
- With five observations, almost any law passes: the test has little power in small samples

SFM_ch6_gof_tests

Estimated Parameters: the Lilliefors Problem

- In practice F has estimated parameters: $F(x;\hat{\theta})$ is fitted to the same data
  - the fit pulls F towards F_n: D_n is smaller than under a known F
  - the usual critical values are too large: the test almost never rejects
- Lilliefors (1967): critical values by simulation for the Normal law with $\hat{\mu}$, $\hat{\sigma}$ (5% level)
  - this chapter's simulation: $n = 5$: 0.351 instead of 0.563; $n = 1000$: 0.029 instead of 0.043
- Any model: **parametric bootstrap** Stute et al. (1993)
  - simulate n draws from $F(\cdot;\hat{\theta})$, refit, recompute the statistic; repeat B times
  - p-value $= (1 + \#\{\text{bootstrap statistics} \geq \text{observed}\})/(B+1)$: the share of simulated statistics at least as large as the observed one

Cramér–von Mises and Anderson–Darling

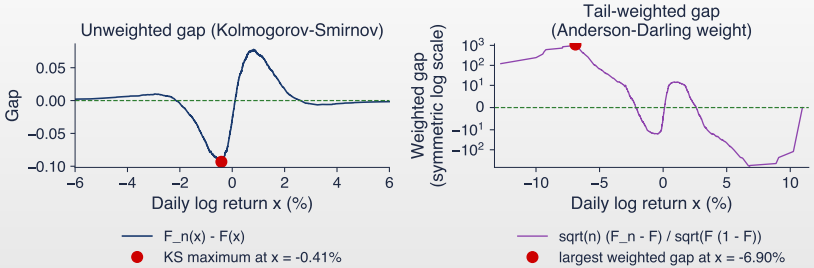
- Both integrate the squared gap $[F_n(x) - F(x)]^2$ over all x instead of taking its maximum
  - **Cramér–von Mises**: $W^2 = n\int [F_n - F]^2\, dF = \frac{1}{12n} + \sum_i \left(u_{(i)} - \frac{2i-1}{2n}\right)^2$ Cramér (1928)
  - **Anderson–Darling**: $A^2 = n\int \frac{[F_n - F]^2}{F(1-F)}\, dF$ Anderson & Darling (1952)
- Computing formula Anderson & Darling (1954):
  $A^2 = -n - \frac{1}{n}\sum_i (2i-1)\left[\ln u_{(i)} + \ln(1 - u_{(n+1-i)})\right]$
  - $u_{(i)} = F(r_{(i)})$ as in KS; large W^2 or A^2: a poor fit
  - the weight $1/[F(1-F)]$ is huge where F is near 0 or 1: the tails



Harald Cramér (1893–1985)

Photo: unknown photographer (1951); public domain; Wikimedia Commons

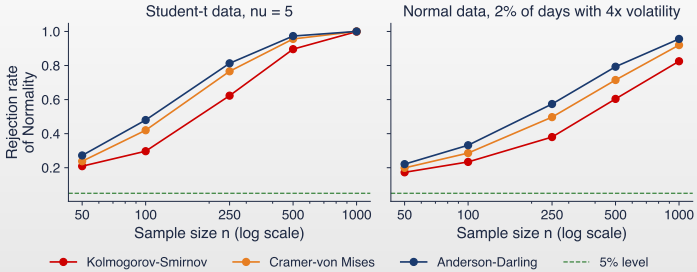
Low Power of KS in the Tails



- S&P 500 against its Normal fit. Left: the KS gap peaks at $x = -0.41\%$, near the centre ($F = 36\%$), with $D_n = 0.093$
- In the tails F_n and F are both close to 0 or 1, so the raw gap is small even when the model is badly wrong
  - the Anderson–Darling weight (right) puts the largest gap at $x = -6.9\%$

SFM_ch6_gof_tests

Power of the Three Tests



- Share of 1000 simulated samples in which Normality (estimated parameters) is rejected at 5%; power = probability of rejecting a false H_0
- Student-t data, $n = 250$: KS 62%, Cramér–von Mises 77%, Anderson–Darling 81%
  - tail-only departures ($n = 250$): 38%, 50%, 57%

SFM_ch6_gof_tests

EDF Tests on the S&P 500

	KS	CvM	AD
Normal	0.093	22.84	127.35
Student-t	0.018	0.57	4.69
skewed-t	0.017	0.41	2.25
GED	0.019	0.38	3.94
Normal mixture	0.025	1.09	5.80
NIG	0.010	0.13	0.88
stable	0.022	0.81	4.61

- Smaller statistics mean a better fit

- Parametric bootstrap, $B = 199$, 95% critical values
  - Student-t: KS 0.009, AD 0.71
  - skewed-t: KS 0.008, AD 0.35
- Normal, Student-t, skewed-t: bootstrap $p = 0.005$, the smallest possible: all rejected
- Student-t, naive KS p-value (as if θ were known): 0.021

SFM_ch6_gof_tests

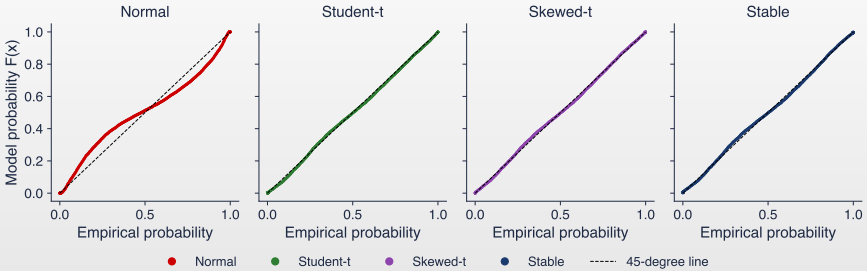
Statistical vs Practical Significance

- With $n = 6\,714$ every model is rejected: tests detect tiny departures
  - and real returns are not i.i.d. (volatility clustering), so no i.i.d. law can pass
- Use the statistics to **rank** the models
  - the AD column of the previous table ranks the models almost as AIC does: NIG first, skewed-t second, the Normal model far behind (AD puts GED and the stable law ahead of the Student-t)
- **Question for the room**: a KS test on 50 daily returns does not reject the Normal law; is the Normal law then fine for VaR 1%?
- **Answer**
  - no: with $n = 50$ the test rarely rejects even Student-t data (30% of the time at $n = 100$), and KS barely looks at the tail

Definitions of PP and QQ Plots

- **PP plot** (probability–probability): points $\left(p_i, F(r_{(i)})\right)$ with $p_i = (i - 0.5)/n$
  - both axes in $[0, 1]$: differences in the centre show; the tails are squeezed into the corners
  - the visual twin of KS: the largest vertical distance to the diagonal is close to D_n
- **QQ plot** (quantile–quantile): points $\left(F^{-1}(p_i), r_{(i)}\right)$
  - in units of returns: the tails are stretched out, each extreme day is visible
  - ends above/below the line: the model tails are too short; ends flattened: the model tails are too long
- Risk management reads QQ plots; PP plots and KS share the same blind spot

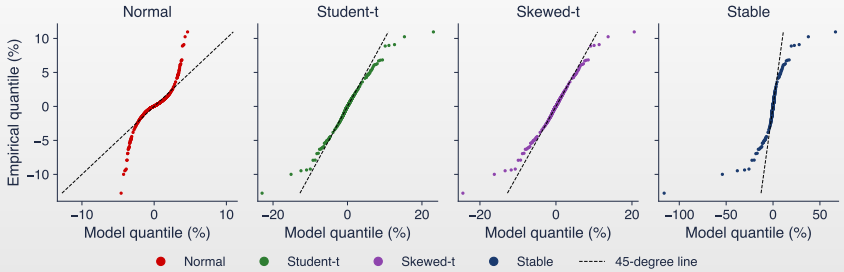
PP Plots: S&P 500



- Normal: an S-shape in the centre (too few small returns, too many medium ones), largest gap 0.093
- Student-t, skewed-t, stable: on the diagonal; in a PP plot they are indistinguishable

SFM_ch6_pp_qq

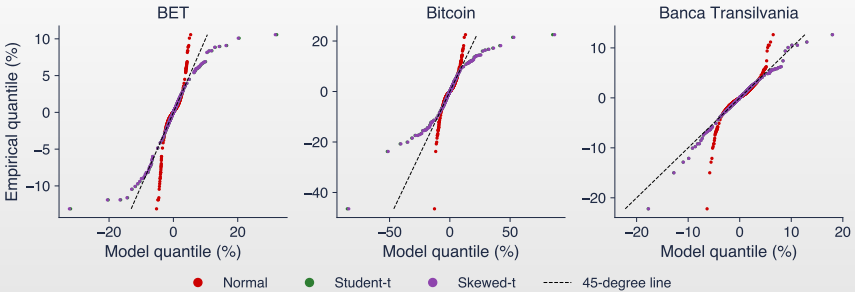
QQ Plots: S&P 500



- The worst day was −12.8%; the model quantile at the same probability:
  - Normal −4.6%, Student-t −22.9%, skewed-t −24.5%, stable −116.1%
- The models that looked identical in the PP plot are far apart in the far tail

SFM_ch6_pp_qq

QQ Plots for BET, Bitcoin and Banca Transilvania



- Normal (red): the steep ends of all three series; the Student-t and skewed-t follow the data far into the tails
- The last few points are single days (the largest crashes): there the fitted tails disagree most and the uncertainty is widest

SFM_ch6_pp_qq

VaR and ES

- Return X, level α (here $\alpha = 1\%$ or 2.5%, the probability of the tail)
  - **VaR** (Value at Risk): $\mathrm{VaR}_\alpha = -q_\alpha(X)$, the loss exceeded with probability α
  - **ES** (expected shortfall): $\mathrm{ES}_\alpha = -\frac{1}{\alpha}\int_0^\alpha q_u\,du = -E[X \mid X \le q_\alpha]$, the average loss in the tail
- Normal model: $\mathrm{VaR}_\alpha = -(\mu + \sigma z_\alpha)$, $z_{0.01} = -2.326$
  - S&P 500: $-(0.025 - 2.326 \times 1.213) = 2.80\%$, against the empirical VaR 1% of 3.45%
- Basel rules: ES 2.5% for market risk, VaR 1% for backtesting (Chapter 10)

VaR 1% and ES 2.5% under Each Model (In Sample)

VaR 1% (%)

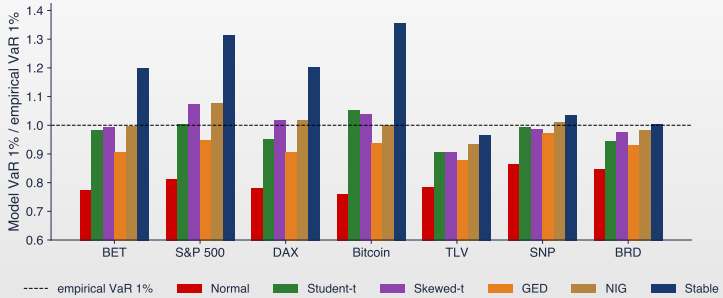
	data	Normal	Student-t	skewed-t	GED	mixture	NIG	stable
BET	4.13	3.19	4.06	4.10	3.74	4.53	4.11	4.94
S&P 500	3.45	2.80	3.46	3.71	3.27	3.97	3.72	4.54
DAX	4.16	3.25	3.96	4.24	3.77	4.35	4.24	5.01
Bitcoin	10.51	7.99	11.06	10.91	9.85	10.40	10.52	14.26
TLV	5.10	4.01	4.63	4.61	4.49	5.41	4.76	4.92
SNP	4.29	3.70	4.26	4.22	4.16	4.70	4.32	4.43
BRD	4.50	3.81	4.24	4.39	4.18	4.74	4.43	4.51

ES 2.5% (%)

	data	Normal	Student-t	skewed-t	GED	mixture	NIG	stable
BET	4.57	3.21	4.76	4.81	3.85	4.53	4.31	7.58
S&P 500	3.76	2.81	3.95	4.23	3.36	3.97	3.89	6.70
Bitcoin	10.84	8.03	13.37	13.18	10.17	10.44	11.06	22.87

SFM_ch6_tail_var

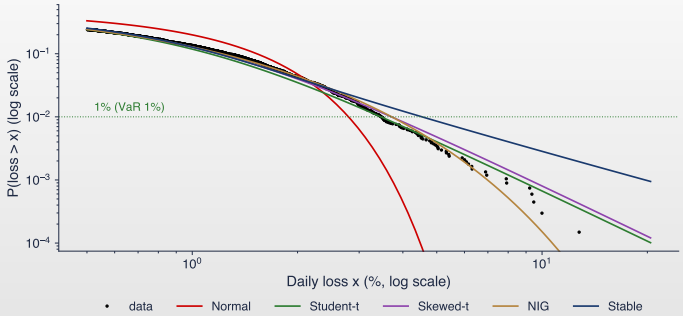
Model VaR 1% Relative to the Data



- Normal: 14–24% too low everywhere; stable: up to 36% too high
- Student-t, skewed-t and NIG stay within 10% of the empirical VaR 1%; GED is a little low

SFM_ch6_tail_var

The Left Tail on Log-Log Axes (S&P 500)



- Points: share of days with a loss above x; lines: the same probability under each fitted model
- At 1% all heavy-tailed models are close; beyond 5% losses the stable line is far too high and the Normal one collapses

SFM_ch6_tail_var

Counting Exceedances: the Kupiec Test

	expected	Normal	Student-t	skewed-t	GED	NIG	stable
BET	67	138	71	69	91	68	43
S&P 500	67	131	67	52	80	52	30
DAX	68	145	78	66	91	65	39
Bitcoin	44	86	34	36	56	44	16

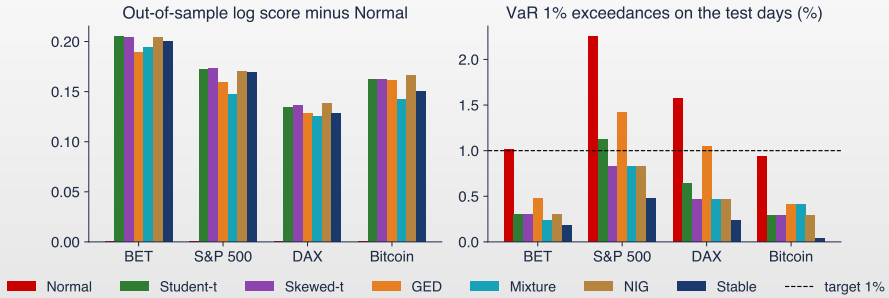
- Exceedance: a day with $r_t < -\mathrm{VaR}_{1\%}$; under a correct model their number is $\mathrm{Binomial}(n, 0.01)$
- Kupiec (1995): $LR_{uc} = -2 \ln \dfrac{(1-\alpha)^{n-x}\alpha^x}{(1-x/n)^{n-x}(x/n)^x} \approx \chi^2(1)$, x = number of exceedances
  - an LR test comparing the target rate α with the observed rate x/n; large LR_{uc}: wrong number of exceedances
- S&P 500: Student-t 67 exceedances for 67 expected ($p = 0.986$)
  - NIG, the AIC winner, 52 ($p = 0.053$)
- The best model for the whole density is not always the best for one quantile; tail-focused scores exist Diks, Panchenko & van Dijk (2011)

SFM_ch6_tail_var

The Role of Out-of-Sample Validation

- In sample, a larger model always fits at least as well: ℓ never decreases when parameters are added
  - **overfitting**: the extra parameters fit noise of the estimation sample, not features of future returns
- The honest check: estimate until 31 December 2019, evaluate on 2020–2026 (COVID-19 crash, 2022 inflation shock, 2025 tariffs)
  - **log score**: the mean of $\ln f(r_t; \hat{\theta})$ over the test days; higher is better, a proper scoring rule Gneiting & Raftery (2007)
  - **exceedances** of VaR 1% (Kupiec) and the **pinball loss** $\frac{1}{n}\sum_t(\alpha - 1\{r_t < q\})(r_t - q)$ of the 1% quantile q
  - pinball: $1\{r_t < q\} = 1$ on a day below q, 0 otherwise; it penalises misses asymmetrically, lower is better
- A score is **proper** if the true distribution gets the best expected score: forecasters cannot gain by distorting

Out-of-Sample Results, 2020–2026



- Left: every heavy-tailed model beats the Normal log score by 0.13 to 0.21 per day; the differences among them are small
- Right: VaR 1% exceedance rates; the target is the dashed line at 1%

SFM_ch6_out_of_sample

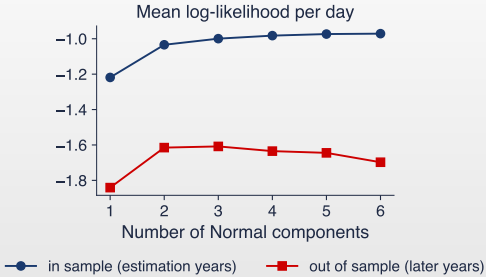
Reading the Out-of-Sample Table

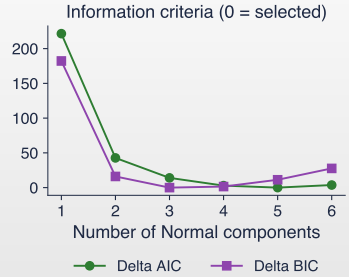
	exp.	Normal	t	skew-t	NIG	stable	hist.	best AIC	best log score
BET	16.8	17	5	5	5	3	5	NIG	Student-t
S&P 500	16.9	38	19	14	14	8	23	NIG	skewed-t
DAX	17.1	27	11	8	8	4	8	NIG	NIG
Bitcoin	24.5	23	7	7	7	1	10	GED	NIG

- VaR 1% exceedances on the test days; exp.: expected number; hist.: historical simulation (the 1% quantile of the estimation sample)
- The AIC winner of 2000–2019 has the best log score only for the DAX; elsewhere the best log score belongs to a close neighbour
- BET: the Normal model has 17 exceedances for 16.8 expected, the heavy-tailed models only 5
  - 2020–2026 was calmer for the BET than 2000–2019
- The right count for the wrong reason: a static model cannot follow changes in volatility

SFM_ch6_out_of_sample

Overfitting: Normal Mixtures with 1 to 6 Components





- S&P 500, estimation 2017–2018 ($n = 501$), test 2019–2026 ($n = 1\,939$); c components have $3c - 1$ parameters
- In sample (blue) the fit improves with every component; out of sample (red) it peaks at $c = 3$ and then declines

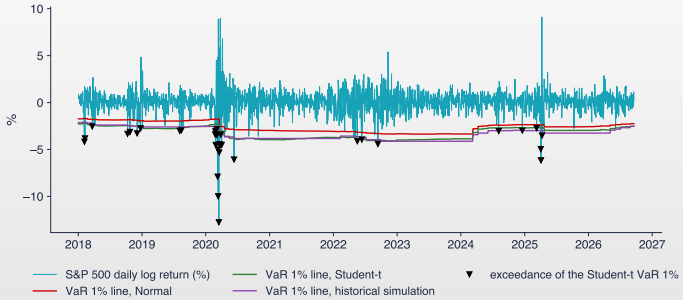
SFM_ch6_out_of_sample

Information Criteria and Overfitting

- Mean log-likelihood per day, in sample → out of sample
  - $c = 1$ (Normal): $-1.219 \to -1.841$; $c = 3$: $-0.999 \to -1.608$; $c = 6$ (17 parameters): $-0.971 \to -1.698$
- BIC chose $c = 3$, the best model out of sample; AIC chose $c = 5$
  - with $n = 501$, AIC's penalty of 2 per parameter is too weak to stop extra components
- The gap between in-sample and out-of-sample fit grows with the number of parameters: that gap is the overfitting
- Rule: short samples, many parameters ⇒ prefer BIC or AICc, and always keep a test period

SFM_ch6_out_of_sample

Rolling VaR 1%: Refit Every Month



- S&P 500: each model refitted every 20 trading days on the previous 1000 days; the VaR 1% line is drawn as a loss threshold $-\mathrm{VaR}$
- The chart shows 2018–2026 only; the backtest of the next slide covers 2003–2026. Black triangles: exceedances of the Student-t VaR; they come in clusters (March 2020)

SFM_ch6_out_of_sample

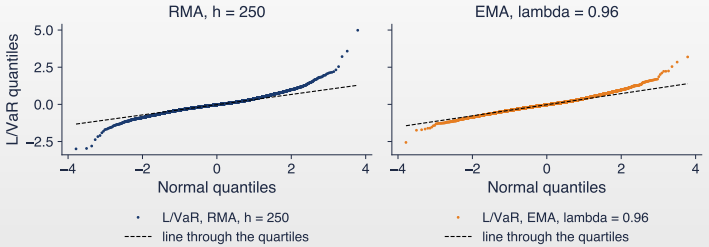
Rolling Backtest: the Results

	Normal	Student-t	skewed-t	historical
S&P 500, rate (%)	2.35	1.56	1.47	1.58
S&P 500, Kupiec p	< 0.001	< 0.001	< 0.001	< 0.001
S&P 500, most in one year	36	31	28	29
BET, rate (%)	1.98	1.52	1.38	1.25
BET, Kupiec p	< 0.001	< 0.001	0.007	0.066

- Backtest 2003–2026, $n = 5\,714$ days for the S&P 500; target rate 1%; the worst year is 2008 for every model and both series
- Heavy tails halve the excess of exceedances, but no unconditional model reaches 1%
  - the volatility changes faster than a 1000-day window can adapt
- The remedy is a model for the volatility itself: GARCH, Chapters 8 and 9

SFM_ch6_out_of_sample

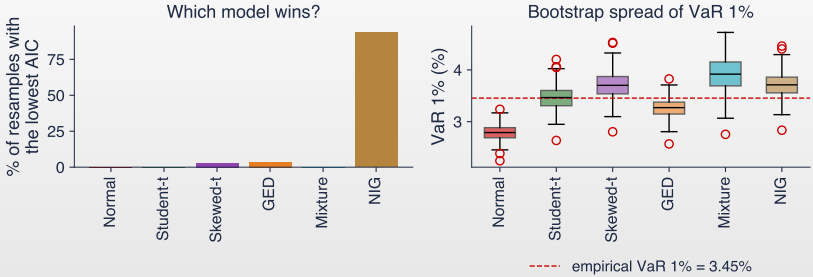
The VaR Reliability Plot (SFEVaRqqplot)



- DAX since 2000: Normal VaR 1% $= -z_{\mathbf{0.01}}\hat{\sigma}_t = 2.326\,\hat{\sigma}_{t_1}$ with $\hat{\sigma}_t$ from the previous 250 days
  - RMA (rectangular moving average): $\hat{\sigma}_t^{\mathbf{2}} = \frac{\mathbf{1}}{\mathbf{250}} \sum_{j=\mathbf{1}}^{\mathbf{250}} r_{t-j}^{\mathbf{2}}$, equal weights
  - EMA (exponential moving average): $\hat{\sigma}_t^{\mathbf{2}} = (\mathbf{1}-\lambda) \sum_{j\geq \mathbf{0}} \lambda^j r_{t-\mathbf{1}-j}^{\mathbf{2}}$, $\lambda = 0.96$: each older day weighs 4% less
  - QQ plot of L/VaR (loss divided by VaR) against Normal quantiles
- A straight line would mean a reliable VaR; both curve upwards at the right end
  - the losses beyond VaR are too large and too frequent (2.13% and 2.07% exceedances)

SFM_ch6_var_qqplot

Uncertainty of the Model Ranking



- S&P 500, 200 moving-block bootstrap resamples (blocks of 20 days keep the volatility clusters); in each, the six models other than the stable law (Normal, Student-t, skewed-t, GED, Normal mixture, NIG) are refitted
- Left: NIG has the lowest AIC in 94% of the resamples; right: the bootstrap spread of VaR 1% under each model

SFM_ch6_model_uncertainty

Two Sources of Uncertainty

- **Parameter uncertainty**: the same model, re-estimated on resampled data
  - Student-t VaR 1%: 90% bootstrap band $[3.14, 3.87]$%; NIG: $[3.34, 4.18]$%
- **Model uncertainty**: different models on the same data
  - median VaR 1%: GED 3.27%, Student-t 3.47%, NIG 3.71%, Normal mixture 3.92%
  - the spread between models is comparable to the spread within one model
- Even when the AIC winner is clear, the number that matters for risk is not
- Entropy-based tail measures are another way to summarise this uncertainty: Pele, Lazar & Dufour (2017), Pele, Lazar & Mazurencu-Marinescu-Pele (2019)

Model Averaging and Model Risk

	data	Normal	heavy-tailed range	averaged	range (%)
BET	4.13	3.19	[3.74, 4.94]	4.11	32
S&P 500	3.45	2.80	[3.27, 4.54]	3.72	39
DAX	4.16	3.25	[3.77, 5.01]	4.24	33
Bitcoin	10.51	7.99	[9.85, 14.26]	9.85	45
TLV	5.10	4.01	[4.49, 5.41]	4.62	21
SNP	4.29	3.70	[4.16, 4.70]	4.25	13
BRD	4.50	3.81	[4.18, 4.74]	4.33	14

- VaR 1% in %; heavy-tailed range: the smallest and largest VaR 1% of the six non-Normal models, the Normal mixture included
  - range (%): max/min -1, the relative distance between the extreme models
- Averaged: $\sum_i w_i \, \mathrm{VaR}_i$ with the Akaike weights w_i
  - a frequentist cousin of Bayesian model averaging Hoeting et al. (1999)
- **Model risk**: the range across plausible models; regulators ask banks to measure it Kerkhof, Melenberg & Schumacher (2010), Danielsson et al. (2016)

SFM_ch6_model_uncertainty

A Model-Selection Workflow

- **1. Data**: check the largest returns against prices and news; remove only proven errors
- **2. Candidates**: choose them for the stylised facts (tails, skewness), not for convenience
- **3. Fit and rank**: ML, then AIC and BIC; LR tests for nested pairs, Vuong for the others
- **4. Check**: AD statistic and QQ plot of the leaders; the region you care about must fit
- **5. Validate**: out of sample, with the loss that matches the purpose (exceedances for VaR)
- **6. Report**: the winner, the runners-up and the range of the risk number across them

An Open Research Question

- **Does NIG still win after we remove volatility clustering?**
  - part of the heavy tails of daily returns comes from changing volatility (Chapter 2)
  - returns divided by a volatility forecast (for example EMA, $\lambda = 0.96$) have lighter tails; which candidate fits them best?
- Why it is open: the answer differs across markets, and the volatility model itself is a choice
- Starting point: on raw returns the winner depends on the market (the ΔAIC map of Section 4)
- AI tools can speed up such a study; they do not replace checking it Wang et al. (2023)

Possible Uses of AI Tools

- **Literature**: list studies that compare return distributions after GARCH filtering, with their samples and criteria
- **Code**: draft the loop "standardise, fit seven models, rank by AIC and BIC, test out of sample"
- **Robustness**: propose other volatility filters, windows and markets
- **Explanation**: a first draft of why the ranking changes after filtering
- Example prompt
  - `Write Python code that divides daily log returns by an EMA volatility (lambda = 0.96), fits Normal, Student-t, skewed-t, GED and NIG by maximum likelihood, and reports AIC, BIC and Akaike weights.`

What to Check

- Look-ahead: the volatility of day t must use only returns up to $t-1$
- Parameter counts: k in AIC must include every estimated parameter, also those of the volatility filter if they were estimated
- Same data for all models: AIC values computed on different samples cannot be compared
- Parameterisations: scale vs standard deviation in the Student-t; S0 vs S1 for the stable law
- References: every cited paper must exist; check the DOI

Project Idea

- **Question**: does the best distribution for standardised returns differ from the best one for raw returns?
  - data: BET, S&P 500, DAX, Bitcoin and BVB stocks from the course data (EODHD)
- Steps
  - standardise by EMA volatility; fit the seven candidates; table of ΔAIC and Akaike weights
  - repeat the out-of-sample VaR 1% backtest of this chapter with the standardised model
  - compare the exceedance clusters of 2008 and 2020 with those of the unconditional models
- Deliverable: one table, one chart and a paragraph on what changes and why
- Declare any AI use, and list the errors of the AI that you corrected

Key Takeaways

- Model choice starts with clean data and a purpose: the body or the tail?
- ML fits; LR tests compare nested models (watch boundaries); Vuong compares non-nested ones
- AIC and BIC rank all candidates; read Δ and the Akaike weights
- EDF tests: AD looks at the tails, KS does not; with large n all are rejected, so rank and plot
- Real returns: NIG, GED and Student-t win; the Normal and stable laws lose everywhere
- Validate out of sample and report the range of VaR across plausible models

Key Formulas

Quantity	Formula
Log-likelihood	$\ell(\theta) = \sum_t \ln f(r_t;\theta)$
Likelihood ratio	$LR = 2(\ell_{\mathbf{1}} - \ell_{\mathbf{0}}) \approx \chi^{\mathbf{2}}(k_{\mathbf{1}} - k_{\mathbf{0}})$
AIC, BIC	$-2\ell(\hat\theta) + 2k, \quad -2\ell(\hat\theta) + k\ln n$
Akaike weight	$w_i = e^{-\Delta_i/\mathbf{2}} / \sum_j e^{-\Delta_j/\mathbf{2}}$
Vuong	$V = \sqrt{n}\,\bar m / s_m, \quad m_t = \ln f_{\mathbf{1}}(r_t) - \ln f_{\mathbf{2}}(r_t)$
KS	$D_n = \sup_x \lvert F_n(x) - F(x)\rvert$
Anderson–Darling	$A^{\mathbf{2}} = n\int [F_n - F]^{\mathbf{2}}/[F(1-F)]\,dF$
Kupiec	$LR_{uc} = -2\ln \frac{(\mathbf{1}-\alpha)^{n-x}\alpha^{x}}{(\mathbf{1}-x/n)^{n-x}(x/n)^{x}}$
Pinball loss	$\frac{\mathbf{1}}{n}\sum_t(\alpha - 1\{r_t < q\})(r_t - q)$

Self-Assessment

- **Question**: model A has $\ell = -1000$ with $k = 3$, model B has $\ell = -998$ with $k = 5$; which has the lower AIC?
  - **Answer**: A: 2006; B: 2006; a tie ($\Delta = 0$): the two extra parameters buy exactly their penalty
- **Question**: A is nested in B; is B significantly better at 5%?
  - **Answer**: $LR = 2 \cdot 2 = 4 < \chi^2_{0.95}(2) = 5.99$: no
- **Question**: why can a KS test with estimated parameters not use the textbook KS table?
  - **Answer**: the fit pulls F towards the data, so D_n is too small; use Lilliefors or a parametric bootstrap
- Next: Chapter 7, efficient markets, the random walk and variance-ratio tests

References (1/3)

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Amisano, G., Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2), 177–190.
- Anderson, T.W., Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2), 193–212.
- Anderson, T.W., Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765–769.
- Barndorff-Nielsen, O.E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24(1), 1–13.
- Borak, S., Härdle, W.K., López-Cabrera, B. (2013). *Statistics of Financial Markets: Exercises and Solutions*, 2nd ed. Springer.
- Box, G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.
- Burnham, K.P., Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer.
- Burnham, K.P., Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261–304.
- Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236.
- Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1), 13–74.
- Danielsson, J., James, K.R., Valenzuela, M., Zer, I. (2016). Model risk of risk models. *Journal of Financial Stability*, 23, 79–91.
- Diks, C., Panchenko, V., van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2), 215–230.

References (2/3)

- Eberlein, E., Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli*, 1(3), 281–299.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). *Statistics of Financial Markets: An Introduction*, 5th ed. Springer.
- Gneiting, T., Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
- Hansen, B.E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35(3), 705–730.
- Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4), 382–417.
- Hurvich, C.M., Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307.
- Kerkhof, J., Melenberg, B., Schumacher, H. (2010). Model risk and capital reserves. *Journal of Banking & Finance*, 34(1), 267–279.
- Kon, S.J. (1984). Models of stock returns – a comparison. *Journal of Finance*, 39(1), 147–165.
- Kullback, S., Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 79–86.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- Lilliefors, H.W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), 399–402.
- Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, 59(2), 347–370.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302), 157–175.

References (3/3)

- Pele, D.T., Lazar, E., Dufour, A. (2017). Information entropy and measures of market risk. *Entropy*, 19(5), 226.
- Pele, D.T., Lazar, E., Mazurencu-Marinescu-Pele, M. (2019). Modeling expected shortfall using tail entropy. *Entropy*, 21(12), 1204.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Self, S.G., Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398), 605–610.
- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2), 279–281.
- Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347), 730–737.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B*, 39(1), 44–47.
- Stute, W., González Manteiga, W., Presedo Quindimil, M. (1993). Bootstrap based goodness-of-fit-tests. *Metrika*, 40(1), 243–256.
- Tsay, R.S. (2010). *Analysis of Financial Time Series*, 3rd ed. Wiley.
- Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307–333.
- Wang, H., Fu, T., Du, Y., Gao, W., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620, 47–60.
- Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9(1), 60–62.