Data Analysis and Approximate Models

Laurie Davies

Fakultät Mathematik, Universität Duisburg-Essen

CRiSM Workshop: Non-likelihood Based Statistical Modelling, University of Warwick, 7-9 September 2015

Is statistics too difficult?

Cambridge 1963: First course on statistics given by John Kingman based on notes by Dennis Lindley.

LSE 1966-1967: Courses by David Brillinger, Jim Durbin and Alan Stuart.

D. W. Müller Heidelberg (Kiefer-Müller process)

Frank Hampel [Hampel, 1998], title as above.

Two phases of analysis

Phase 1: EDA;
scatter plots, q-q-plots, residual analysis, ...
provides possible models for formal treatment in Phase 2

Phase 2: formal statistical inference;
hypothesis testing, confidence intervals, prior distributions,
posterior distributions, ...

Two phases of analysis

The two phases are often treated separately.

It is possible to write books on Phase 1 without reference to Phase 2 [Tukey, 1977].

It is possible to write books on Phase 2 without reference to Phase 1 [Cox, 2006].

Two phases of analysis

In going from Phase 1 to Phase 2 there is a break in the modus operandi.

Phase 1: probing, experimental, provisional.

Phase 2: Behaving as if true.

Truth in statistics

Phase 2: Parametric family

$$\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$$

Frequentist:
There exists a true $\theta \in \Theta$.
Optimal estimators, or at least asymptotically optimal, maximum likelihood
An α-confidence region for θ is a region which, in the long run, contains the true parameter value with a relative frequency α.

Truth in statistics

Bayesian:
The Bayesian paradigm is completely wedded to truth.

There exists a true $\theta \in \Theta$.

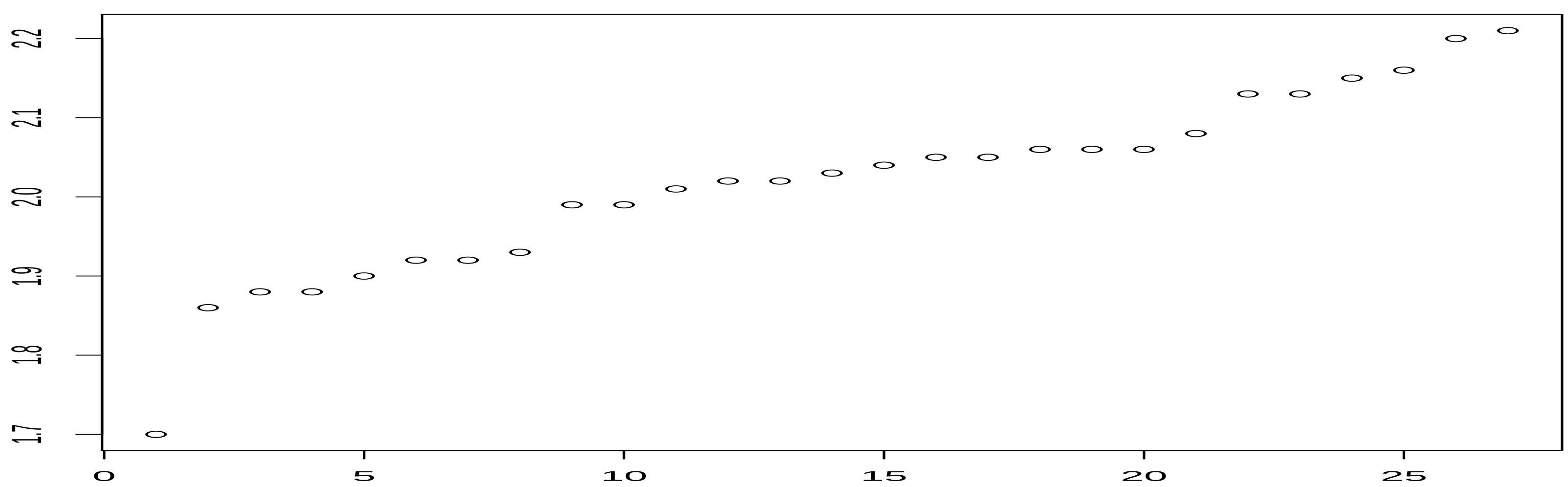
Two different parameter values θ_1, θ_2 with $P_{\theta_1} \neq P_{\theta_2}$, cannot both be true.

A Dutch book argument now leads to the additivity of a Bayesian prior, the requirement of coherence

An example: copper data

27 measurements of amount of copper (milligrammes per litre) in a sample of drinking water.

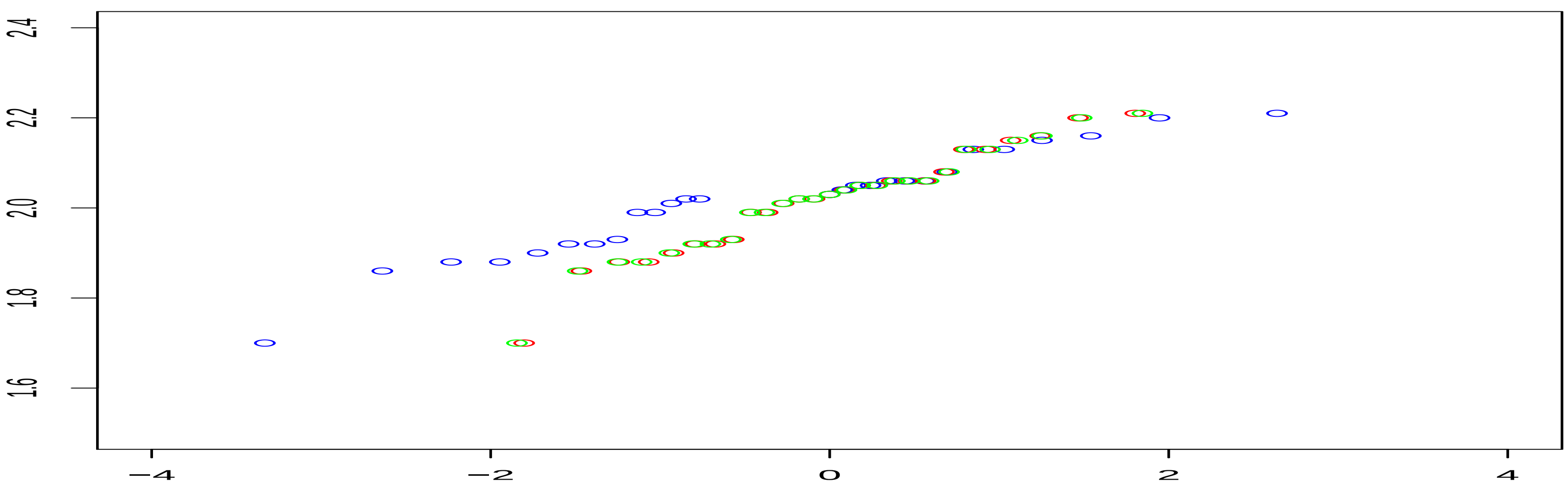
cu=(2.16 2.21 2.15 2.05 2.06 2.04 1.90 2.03 2.06
2.02 2.06 1.92 2.08 2.05 1.88 1.99 2.01 1.86
1.70 1.88 1.99 1.93 2.20 2.02 1.92 2.13 2.13)



An example: copper data

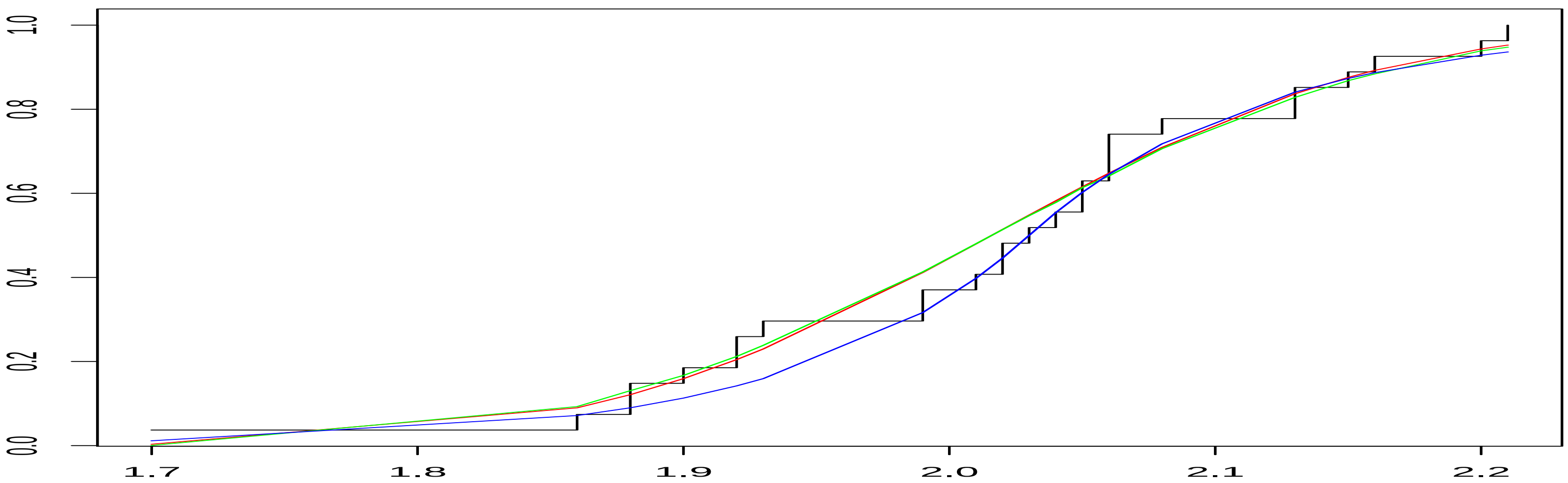
Outliers? Hampel 5.2mad criterion:
$\max|cu - \text{median}(cu)|/\text{mad}(cu) = 3.3 < 5.2$

Three models: (i) the Gaussian (red), (ii) the Laplace (blue), (iii) the comb (green)
q-q-plots



An example: copper data

Distribution functions:



End of phase 1.

An example: copper data

Phase 2

For each location-scale model $F((\cdot - \mu)/\sigma)$ behave as if were true.

Estimate the parameters μ and σ as efficiently as possible.

Maximum likelihood (at least asymptotically efficient).

Copper data				
Model	Kuiper, p-value	log–lik.	95%–conf. int.	length
Normal	0.204, 0.441	20.31	$[1.970, 2.062]$	0.092
Laplace	0.200, 0.304	20.09	$[1.989, 2.071]$	0.082
Comb	0.248, 0.321	31.37	$[2.0248, 2.0256]$	0.0008

An example: copper data

Bayesian: comb model

Prior for μ uniform over $[1.7835, 2.24832]$, for σ independent of μ and uniform over $[0.042747, 0.315859]$.

Posterior for μ is essentially concentrated on the interval $[2.02122, 2.02922]$ agreeing more or less with the 0.95-confidence interval for μ.

An example: copper data

18 data sets in [Stigler, 1977]

	Normal		Comb	
Data	p-Kuiper	log-lik	p-Kuiper	log-lik
Short 1	0.535	-19.25	0.234	-13.92
Short 2	0.049	-21.27	0.003	-18.17
Short 3	0.314	-16.10	0.132	-8.81
Short 4	0.327	-24.42	0.242	-17.66
Short 5	0.102	-19.20	0.022	-13.91
Short 6	0.392	-28.31	0.238	-25.98
Short 7	0.532	12.41	0.495	22.80
Short 8	0.296	-0.49	0.242	10.19
Newcomb 1	0.004	-85.25	0.000	-73.78
Newcomb 2	0.802	-60.55	0.737	-45.85
Newcomb 3	0.483	-75.97	0.330	-59.71
Michelson 1	0.247	-120.9	0.093	-104.7
Michelson 2	0.667	-111.9	0.520	-93.66
Michelson 3	0.001	-115.3	0.000	-100.0
Michelson 4	0.923	-109.8	0.997	-100.8
Michelson 5	0.338	-107.7	0.338	-97.05
Michelson 6	0.425	-139.6	0.077	-134.6
Cavendish	0.991	3.14	0.187	10.22

An example: copper data

Now use AIC or BIC ([Akaike, 1973] [Akaike, 1974] [Akaike, 1981] [Schwarz, 1978]) to choose the model.
The winner is the comb model.

Conclusion 1:
This shows the power of likelihood methods demonstrated by their ability to give such a precise estimate of the quantity of copper using data using data of such quality.

Conclusion 2:
This is nonsense, something has gone badly wrong.

Two topologies

Generating random variables.

Two distribution functions F and G and a uniform random variable U

$$X = F^{-1}(U) \Rightarrow X \stackrel{D}{=} F, \quad Y = G^{-1}(U) \Rightarrow Y \stackrel{D}{=} G.$$

Suppose F and G close in the Kolmogorov or Kuiper metrics

$$d_{\mathsf{ko}}(F, G) = \max_x |F(x) - G(x)|, \quad d_{\mathsf{ku}}(F, G) = \max_{x<y} |F(y) - F(x) - (G(y) - G(x))|.$$

Then X and Y will in general be close. Taking finite precision into account can result in $X = Y$.

Two topologies

An example: $F = N(0,1)$ and $G = C_{\textsf{comb},(k,ds,p)}$ given by

$$C_{\textsf{comb},(k,ds,p)}(x) = p\left(\frac{1}{k}\sum_{j=1}^{k} F((x-\iota_k(j))/ds)\right) + (1-p)F(x)$$

where

$$\iota_k(j) = F^{-1}(i/(k+1)),\ i = 1,\ldots,k$$

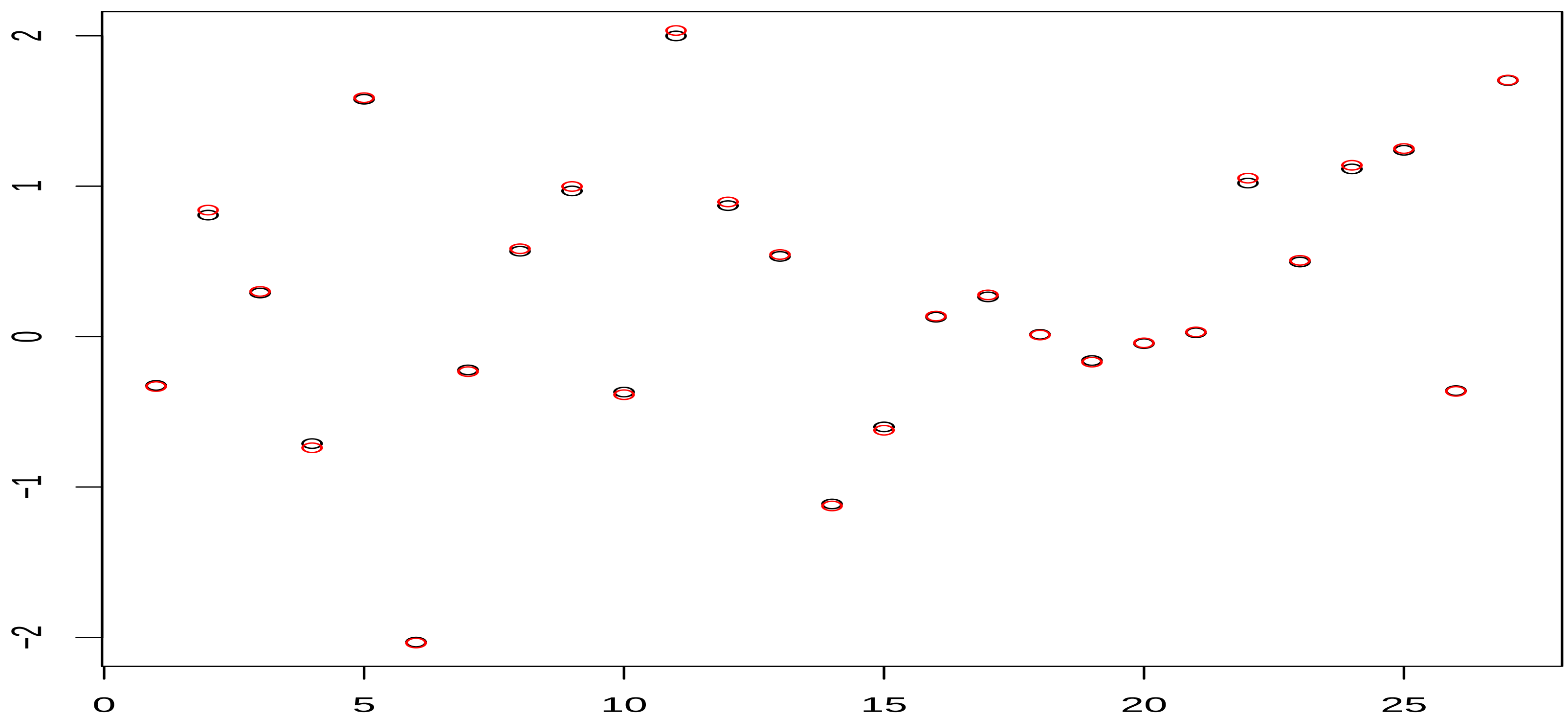
and $(k, ds, p) = (75, 0.005, 0.85)$.

$C_{\textsf{comb},(k,ds,p)}$ is a mixture of normal distributions,
$(k, ds, p) = (75, 0.005, 0.85)$ is fixed

The Kuiper distance is $d_{\textsf{ku}}(N(0,1), C_{\textsf{comb}}) = 0.02$.

Two topologies

Standard normal (black) and comb (red) random variables.



Two topologies

Phase 1 is based on distribution functions.

This is the level at which data distributed according to the model are generated.

The topology of Phase 1 is typified by the Kolmogorov metric d_{ko} or, equivalently, by the Kuiper metric d_{ku}.

Two topologies

Move to Phase 2:

Analyse the copper data using the normal and comb models.

For both models behave as if true, leads to likelihood.

Likelihood is density based $\ell(\theta, \boldsymbol{x}_n) = f(\boldsymbol{x}_n, \theta)$.

Two topologies

Phase 1 based on $F(x, \theta)$, Phase 2 on $f(x, \theta)$, where

$$F(x, \theta) = \int_{-\infty}^{x} f(u, \theta)\, du, \quad f(x, \theta) = D(F(x, \theta))$$

Phase 1 and Phase 2 connected by the linear differential operator D.

When are two densities f and g close?
Use the L_1 metric

$$d_1(f, g) = \int |f - g|$$

Two topologies

$\mathcal{F} = \{F : \text{absolutely continuous, monotone}, F(-\infty) = 0, F(\infty) = 1\}$

$$D : (\mathcal{F}, d_{\mathsf{ko}}) \to (\mathcal{F}, d_1), \quad D(F) = f$$

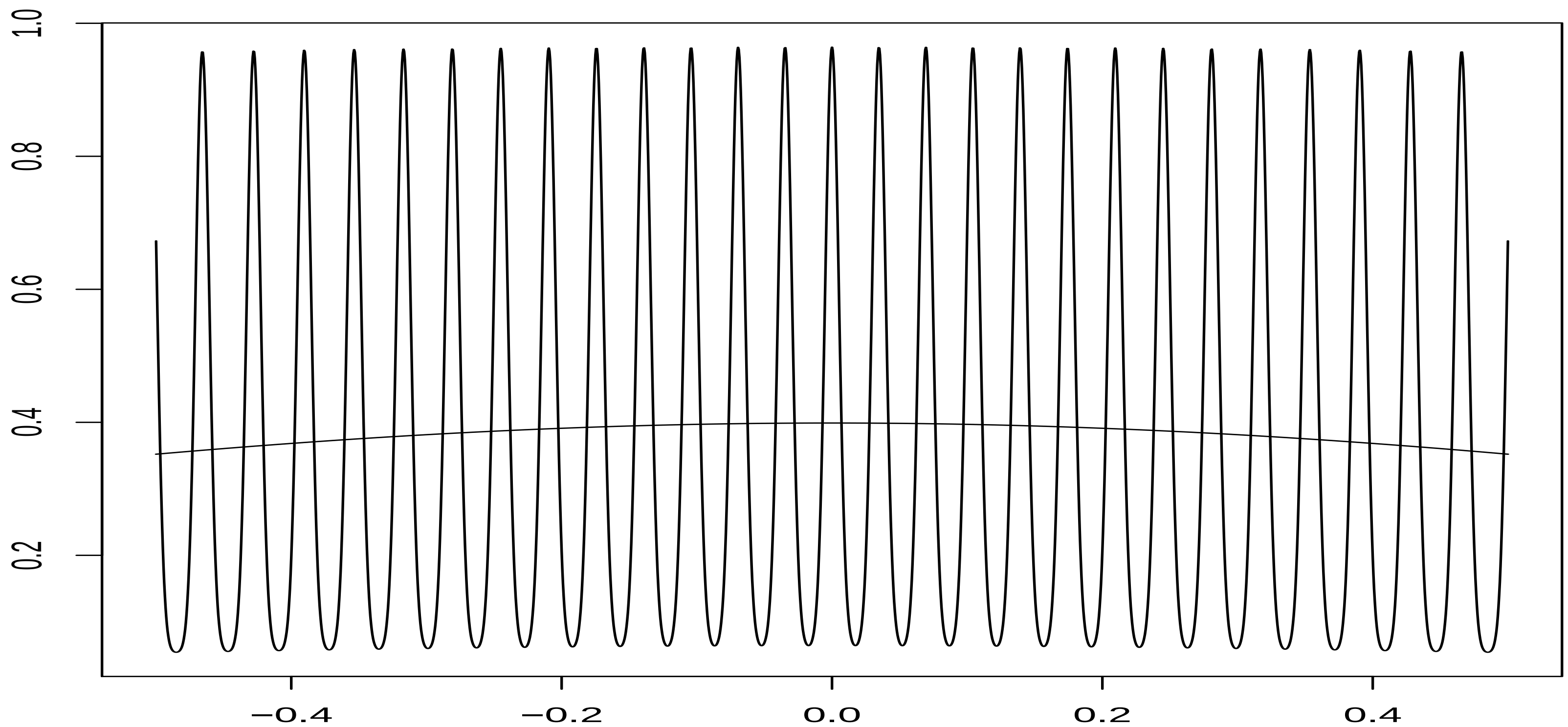
D is an unbounded linear operator and is consequently pathologically discontinuous.

The topology $\mathfrak{O}_{d_{\mathsf{k}o}}$ induced by d_{ko} is weak, few open sets.
The topology $\mathfrak{O}_{d_1}$ induced by d_1 is strong, many open sets.

$$\mathfrak{O}_{d_{\mathsf{k}o}} \quad \subset \quad \mathfrak{O}_{d_1}$$

Two topologies

Standard normal and comb density functions.



$d_1(N(0,1), C_{\mathsf{comb}}) = 0.966.$

Regularization

The location-scale problem $F((\cdot - \mu)/\sigma)$ with choice F is ill-posed and requires regularization.

The results for the copper data show that 'efficiency=small confidence interval' can be imported through the model

Tukey ([Tukey, 1993]) call this a free lunch and states that there is no such thing as a free lunch

TINSTAAFL

He calls models which do not introduce efficiency 'bland' or 'hornless'.

Regularization

Measure of blandness is the Fisher information

Minimum Fisher models: normal and Huber 4.4 of [Huber and Ronchetti, 2009], see also [Uhrmann-Klingen, 1995]

Copper data					
Model	Kuiper, p-value	log–lik.	95%–conf. int.	length	Fisher Inf.
Normal	0.204, 0.441	20.31	$[1.970, 2.062]$	0.092	$2.08 \cdot 10^3$
Laplace	0.200, 0.304	20.09	$[1.989, 2.071]$	0.082	$1.41 \cdot 10^4$
Comb	0.248, 0.321	31.37	$[2.0248, 2.0256]$	0.0008	$3.73 \cdot 10^7$

Regularization

Seems to imply - use minimum Fisher information models

Location and scale are linked in the model

Combined with Bayes or maximum likelihood may be sensitive to outliers

Normal and Huber distributions Section 15.6 of [Huber and Ronchetti, 2009].
Cauchy, t-distributions not sensitive - Fréchet differentiable - Kent-Tyler functionals.

Regularization

Regularize through procedure rather than model

Smooth M-functionals, locally uniformly differentiable.
$(T_L(P), T_S(P))$ solution of

$$\begin{aligned} \int \psi\left(\frac{x - T_L(P)}{T_S(P)}\right) dP(x) &= 0, \\ \int \chi\left(\frac{x - T_L(P)}{T_S(P)}\right) dP(x) &= 0 \end{aligned} \tag{1}$$

Regularization

Possible choice of ψ and χ

$$\begin{aligned}\psi(x) = \psi(x, c) &= \frac{\exp(cx) - 1}{\exp(cx) + 1}, \\ \chi(x) &= \frac{x^4 - 1}{x^4 + 1}.\end{aligned}$$

Solve with $c = 5$, retain $T_S(P)$ and then solve (1) for $T_L(P)$ with $c = 1$ to give a location functional $\tilde{T}_L$.

0.95-approximation interval for copper data $[1.973, 2.065]$, Gaussian model $[1.970, 2.062]$.

A well-posed example

The location-scale problem is ill-posed but likelihood can fail in well-posed problems.
The following example is due to [Gelman, 2003]. Data are distributed as

$$MN(\theta) = 0.5N(\mu_1, \sigma_1^2) + 0.5N(\mu_2, \sigma_2^2)$$

with $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. Maximum likelihood and Bayes fail.

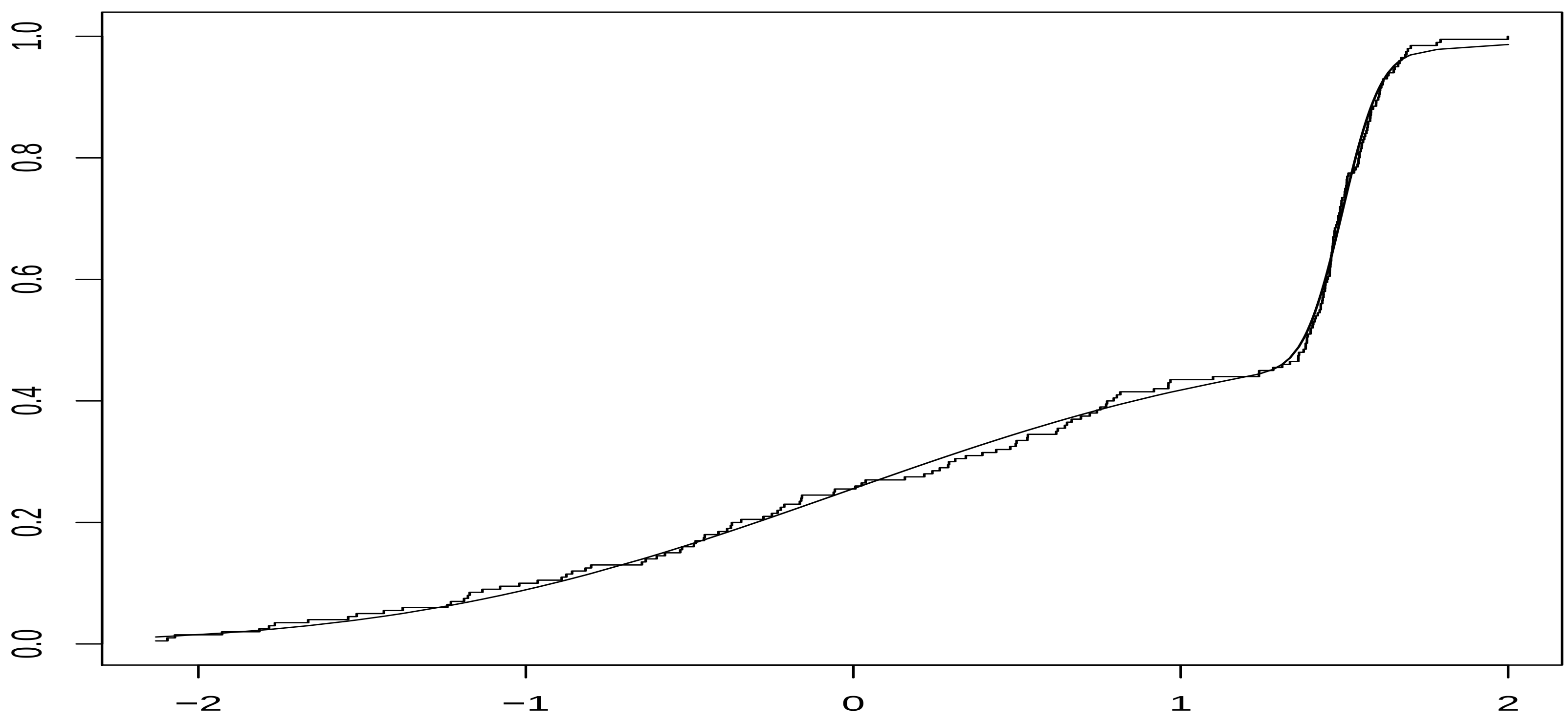
$$\hat{\theta} = \text{argmin}_\theta \, d_{ko}(\mathbb{P}_n, MN(\theta))$$

with the added bonus that you may decide that the data are not distributed as $MN(\theta)$ for any θ.

A well-posed example

$$\theta = (0, 1, 1.5, 0.01)$$

$$\hat{\theta} = (-0.029, 1.053, 1.494, 0.0912)$$



Likelihood

(a) Likelihood reduces the measure of fit between a data set $\boldsymbol{x}_n$ and a statistical model P_θ to a single number irrespective of the complexity of both.

(b) Likelihood is dimensionless and imparts no information about closeness.

(c) Likelihood is blind. Given the data and the model or models, it is not possible to deduce from the values of the likelihood whether the models are close to the data or hopelessly wrong.

(d) Likelihood does not order models with respect to their fit to the data.

Likelihood

(e) Likelihood based procedures for model choice (AIC, BIC, MDL, Bayes) give no reason for being satisfied or dissatisfied with the models on offer.

(f) Likelihood does not contain all the relevant information in the data $\boldsymbol{x}_n$ about the values of the parameter θ.

(g) Given the model, the sample cannot be reduced to the sufficient statistics without loss of information.

(h) Likelihood is based on the differential operator and is consequently pathologically discontinuous.

(i) Likelihood is evanescent: a slight perturbation of the model P_θ to a model P_θ^* can cause it to vanish.

Likelihood

On the positive side:

(j) Likelihood delimits the possible.

The likelihood principle:
Pointless and a waste of intellectual effort.
Birnbaum [Birnbaum, 1962]: likelhood principle when model is 'adequate'.
Adequate never spelt out, constitutes an intellectual failure.
Chasm between 'adequate' and 'true'.
There are many adequate likelihoods, which one and why?

Approximate models

Project:
Give an account of data analysis which consistently treats models as approximations.

A model P is an adequate approximation to a data set $\boldsymbol{x}_n$ if 'typical' data sets $\boldsymbol{X}_n(P)$ generated under P 'look like' $\boldsymbol{x}_n$.

[Neyman et al., 1953], [Neyman et al., 1954], [Donoho, 1988], [Davies, 1995], [Davies, 2008], [Buja et al., 2009], [Xia and Tong, 2011], [Berk et al., 2013], [Huber, 2011], [Davies, 2014].

Approximate models

'Approximation' is a measure of closeness and this requires a topology.

The topology is a weak topology characterized by the Kolmogorov metric.

Approximate the data set as given. Non-frequentist.

No true parameter so no confidence intervals in the frequentist sense - no true value to be covered.

Approximate models

Bayesian approximation?

Parametric family $\mathcal{P}_\Theta$ and prior Π over Θ. No two different P_θ can both be true but two different P_θ can both be approximations. No exclusion, no Dutch book, no coherence.

Within the standard Bayesian set-up there can be no concept of approximation.

More generally there can be no likelihood based concept of approximation.
In particular, no Kullback-Leibler, no AIC, no BIC

Approximate models

Data $\boldsymbol{x}_n$, family of models $N(\mu, 1)$, 'typical'=0.95, (95% of the data generated under the model are classified as typical) 'looks like' $=$ mean.

Under $N(\mu, 1)$ typical means lie in

$$(\mu - 1.96/\sqrt{n}, \mu + 1.96/\sqrt{n}).$$

The mean $\bar{\boldsymbol{x}}_n$ of the data looks like a typical mean of an $N(\mu, 1)$ sample, that is, $N(\mu, 1)$ is an adequate approximation, if

$$\mu - 1.96/\sqrt{n} \leq \bar{\boldsymbol{x}}_n \leq \mu + 1.96/\sqrt{n}$$

Approximate models

Approximation region

$$\mathcal{A}(\boldsymbol{x}_n, 0.95, \mathbb{R}) = \left\{\mu : |\mu - \bar{\boldsymbol{x}}_n| \le \frac{1.96}{\sqrt{n}}\right\}$$

Note there is no assumption that the $\boldsymbol{x}_n$ are a realization of $\boldsymbol{X}_n(\mu)$ for some 'true' μ.

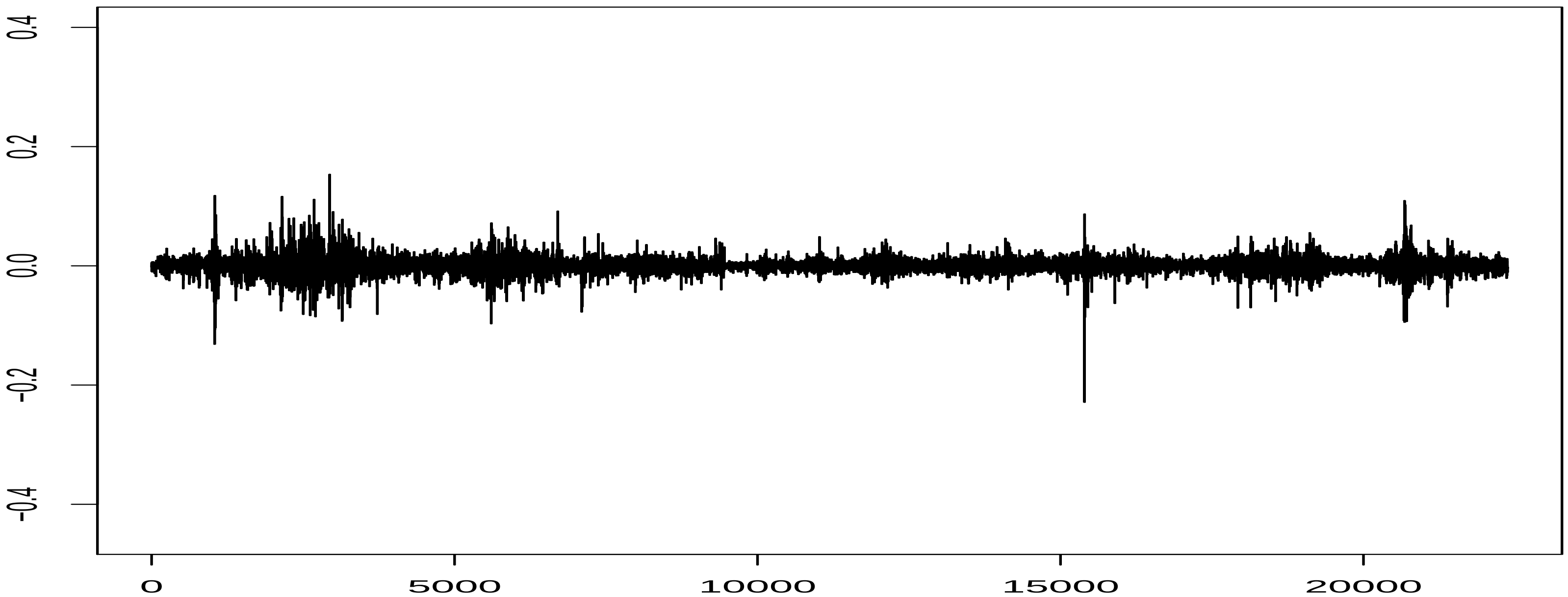
A more complicated approximation region

$$\begin{aligned}\mathcal{A}(\boldsymbol{x}_n, \alpha, \mathcal{N}) \;=\; & \Big\{(\mu,\sigma) : d_{\mathsf{ku}}(\mathbb{P}_n, N(\mu,\sigma^2)) \le \mathsf{qku}(\alpha_1, n), \\ & \max_i |x_i - \mu|/\sigma \le \mathsf{qout}(\alpha_2, n), |T_{\mathsf{skew}}(\mathbb{P}_n)| \le \mathsf{qskew}(\alpha_3, n), \\ & \sqrt{n}|\bar{\boldsymbol{x}}_n - \mu|/\sigma \le \mathsf{qnorm}(\alpha_4), \mathsf{qchisq}((1-\alpha_5)/2, n) \le \\ & \sum_{i=1}^n (x_i - \mu)^2/\sigma^2 \le \mathsf{qchisq}((1+\alpha_5)/2, n)\Big\}\end{aligned}$$

T_{skew} measure of skewness. You have to pay for everything

Simulating long range financial data

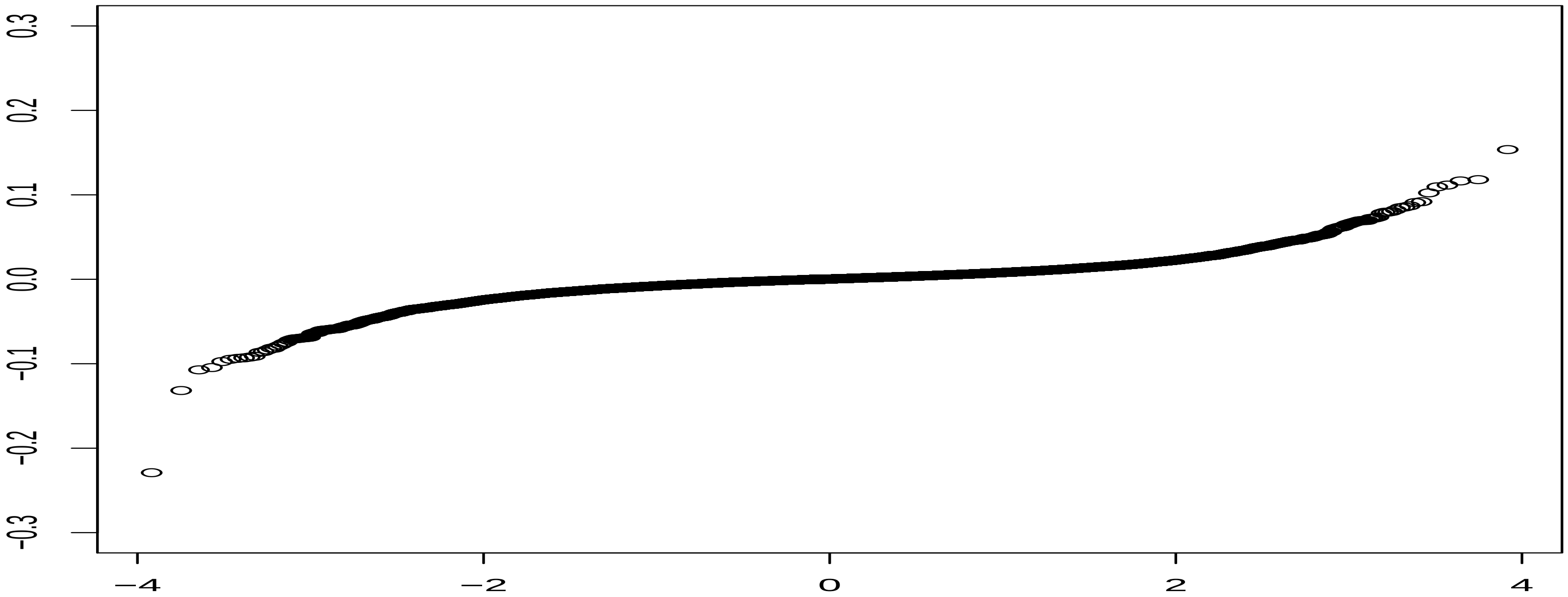
Daily returns of Standard and Poor's, 22381 observations over about 90 years.



Stylized facts 1: volatility clustering

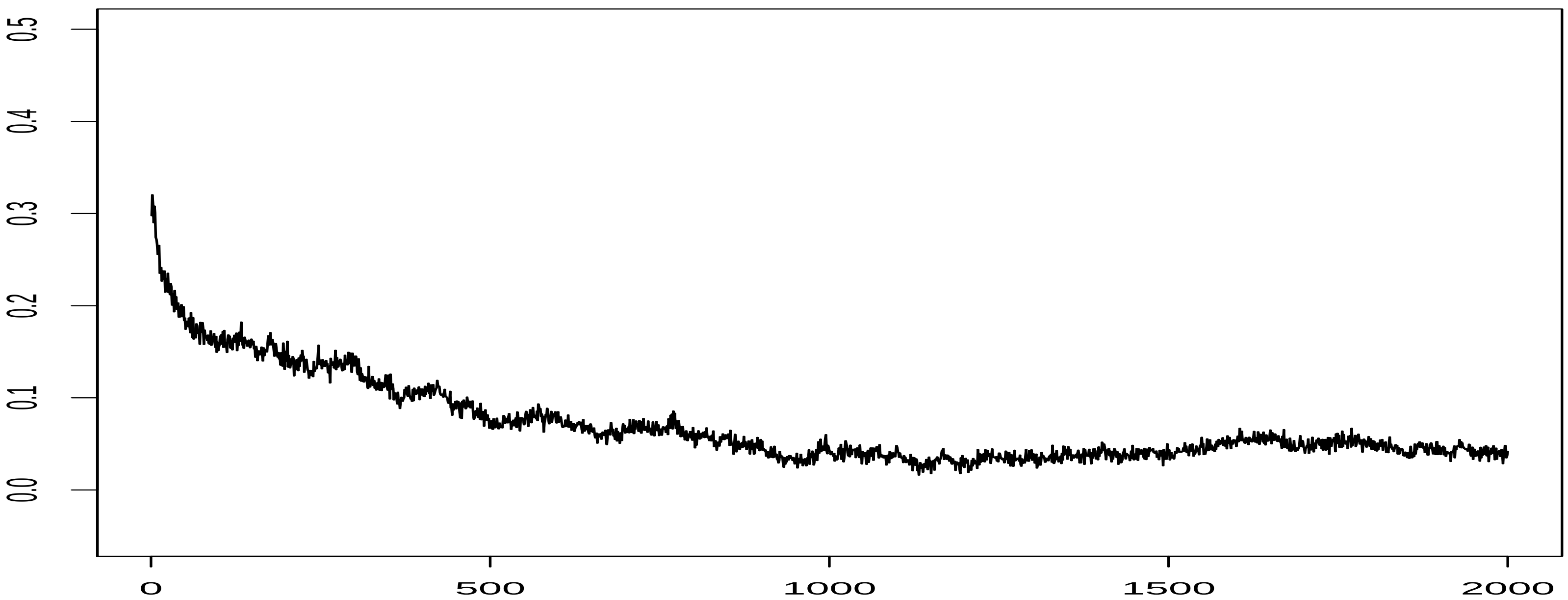
Simulating long range financial data

Stylized facts 2: heavy tails, q-q-plot



Simulating long range financial data

Stylized facts 3: slow decay of correlations of absolute values (long term memory (?))

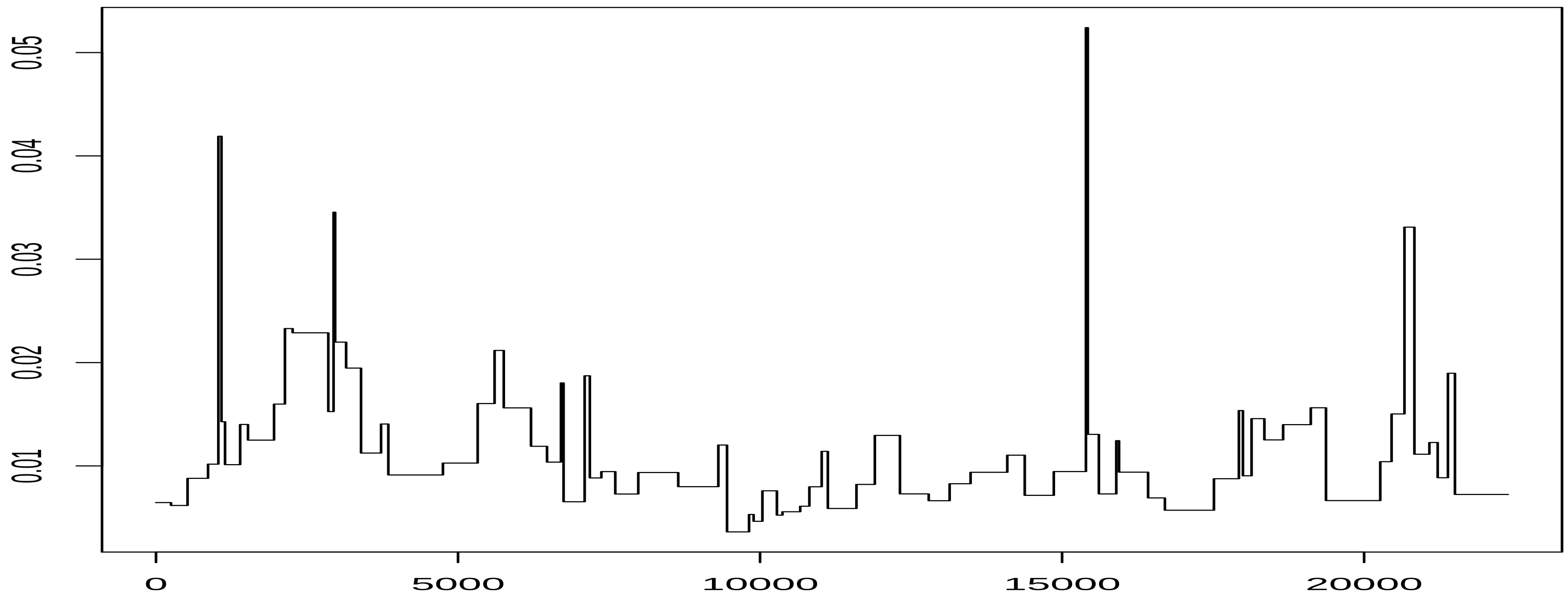


Simulating long range financial data

Quantifying stylized facts:

Piecewise constant volatility with 76 intervals

[Davies et al., 2012]



Simulating long range financial data

Also take the unconditional volatility

$$\frac{1}{n}\sum_{t=1}^{n}|r(t)|, \quad \frac{1}{n}\sum_{t=1}^{n}r(t)^2$$

and the long range return

$$\exp\left(\sum_{t=1}^{n}r(t)\right)$$

into account.

In all 6 features of the data will be taken into account, all quantified.

Simulating long range financial data

Basic model

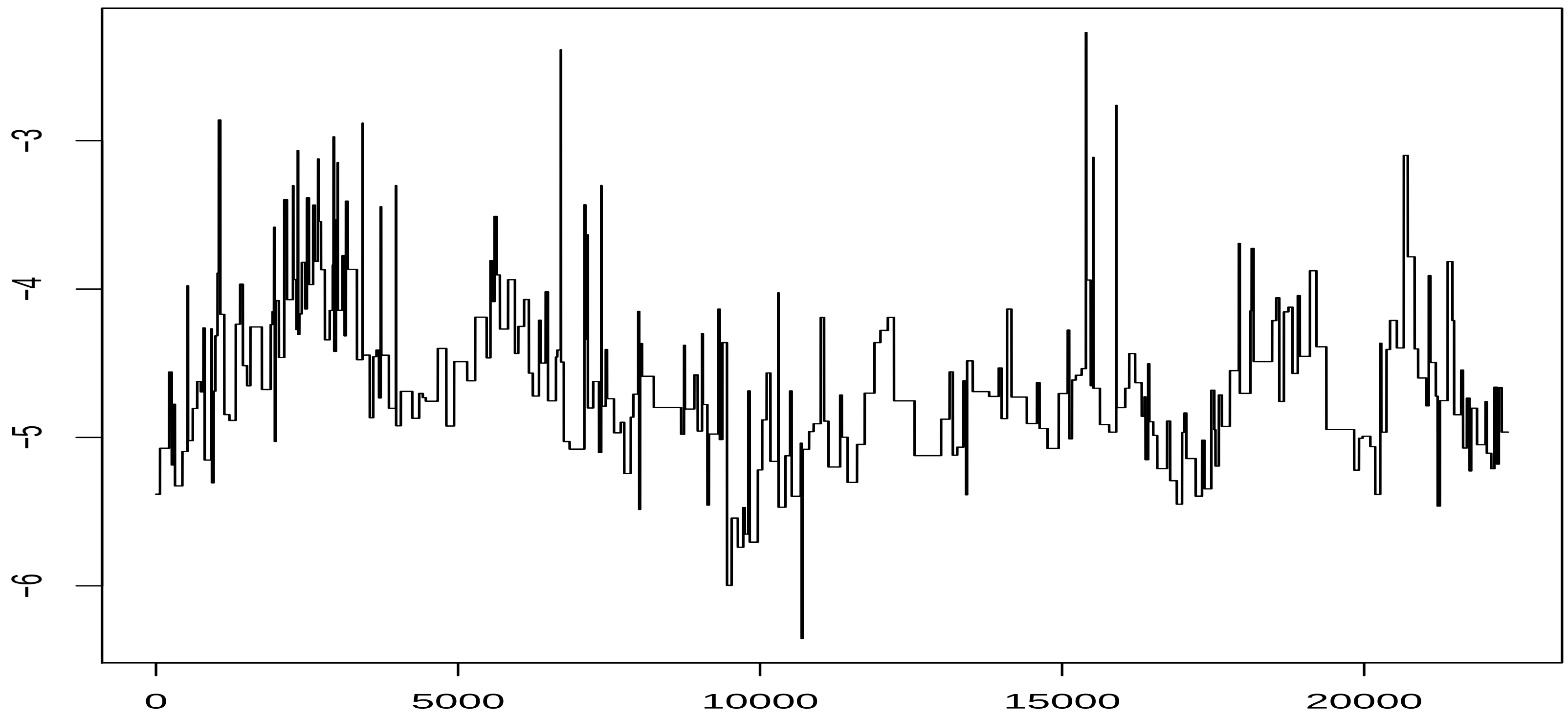
$$R(t) = \Sigma(t)Z(t)$$

Model for $\Sigma(t)$ is the main problem.

Default for $Z(t)$ is i.i.d. $N(0,1)$ but allow for heavier or lighter tails, correlations and dependency of sign of $R(t)$ on $|R(t)|$.

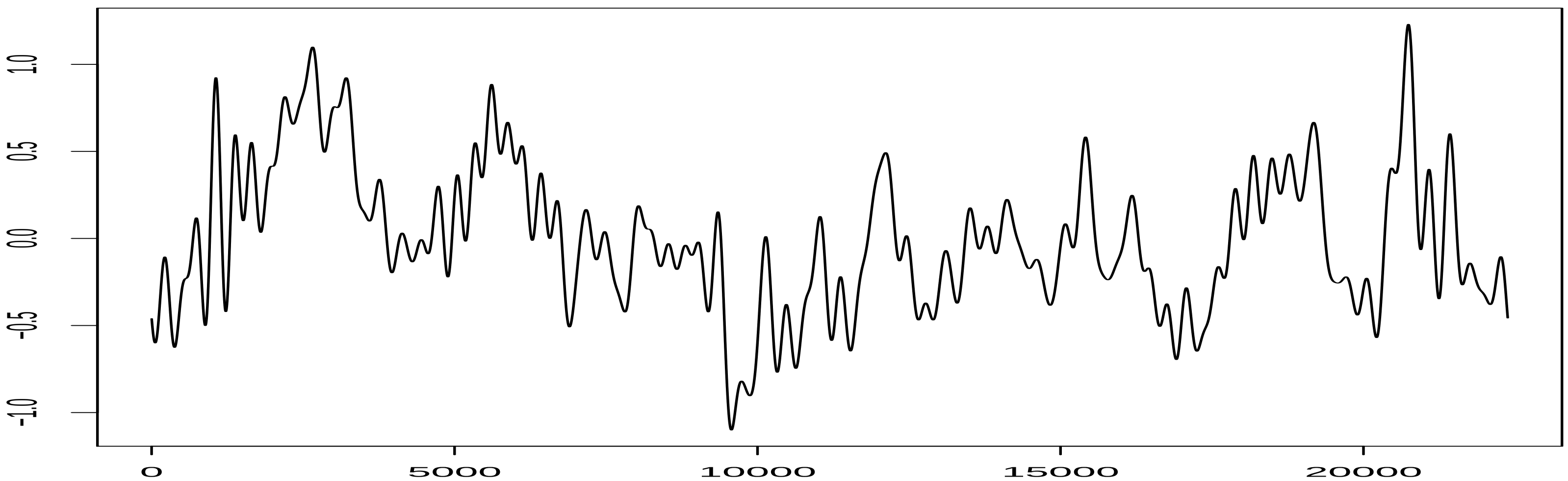
Simulating long range financial data

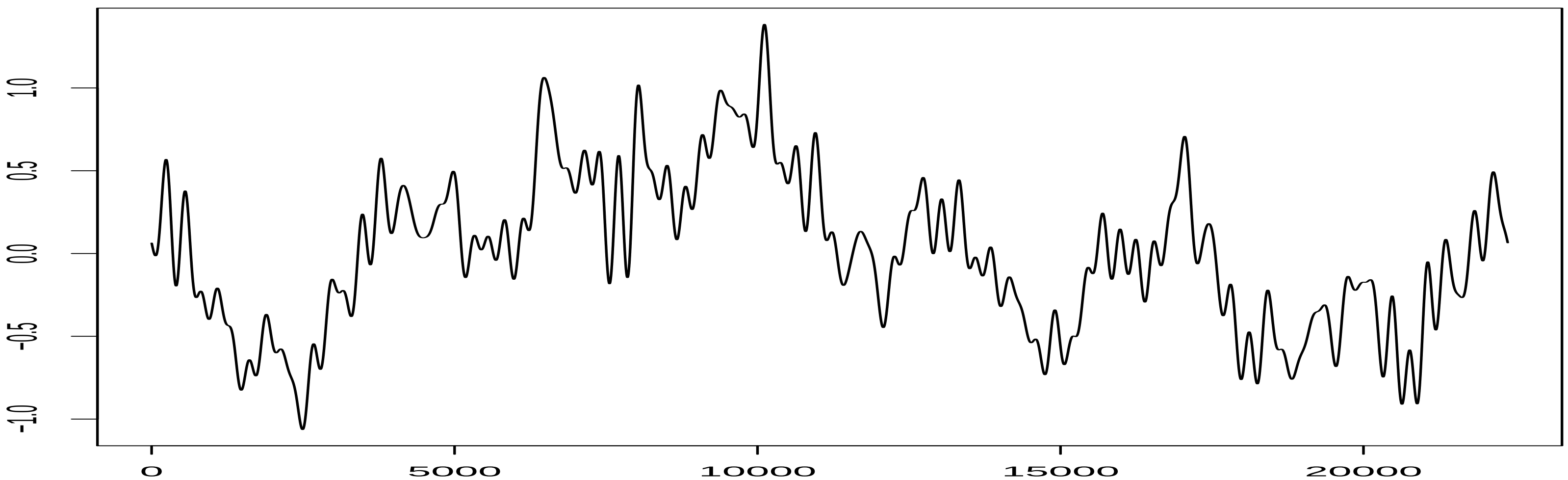
Piecewise constant log-volatility with 283 intervals (1st screw)



Simulating long range financial data

Low frequency trigonometric approximation (2nd screw) and randomized version





Simulating long range financial data

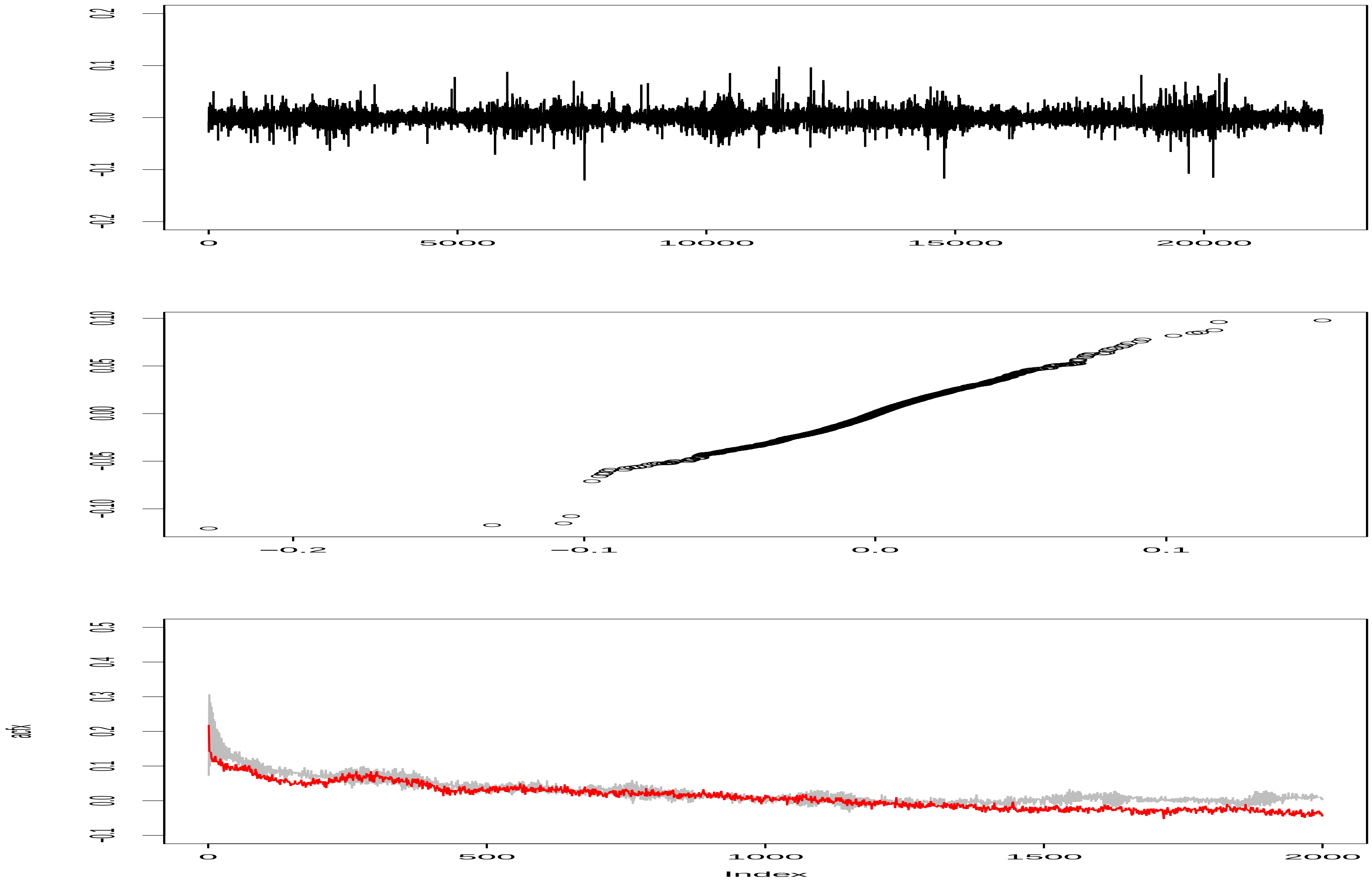
Add high frequency component (3rd screw) and noise (4th screw) to the log-volatility.

Multiply volatility by $Z(t)$ with screws for
(i) heaviness of tails (5th screw)
(ii) short term correlations (6th screw)
(iii) dependence of sign$(R(t))$ on $|R(t)|$ (7th screw)

Adjust the screws if possible so that all six features have high p-values, at least 0.1. Form of feature matching as in [Xia and Tong, 2011].

Simulating long range financial data

A simulated data set



Simulating long range financial data

Statistics for the simulated data set:
Intervals: 84 as against 76 for S+P
Mean absolute deviation of quantiles: 0.00067
Mean absolute deviation of acf: 0.020
Mean absolute volatility: 0.00773 as against 0.00766
Mean squared volatility: 0.000116 as against 0.000137
Returns: 37.93 as against 27.06

p-values based on 1000 simulations

returns	intervals	mean abs. vol.	mean squ.vol.	quantiles	acf
0.934	0.531	0.292	0.305	0.977	0.532

Simulating long range financial data

How can one simulate a non-repeatable data set?

What can actually be estimated?

Title of **8.2d** of [Hampel et al., 1986]

Copper data. What do we want to estimate?

The amount of copper in the sample of water, say q_{cu}.

To do this the statistician often formulates a parametric model P_θ, estimates θ based on the data and then identifies q_{cu} with some function of θ, say $h(\theta)$.

What can actually be estimated?

Models: Gaussian, Laplace and comb.

Symmetric, identify the quantity of copper with the point of symmetry, namely μ.

Gives a consistent interpretation over the different models.

[Tukey, 1993] data in analytical chemistry are often not symmetric.

What can actually be estimated?

Log-normal model $LN(\mu, \sigma^2)$.

Identify amount of copper with $h(\mu, \sigma)$, h?

Consistency of interpretation across all four models?

Model P, identify quantity of water with $T(P)$
T mean, median, M-functional, ...

No explicit parametric model.

Choice of regression functional

Dependent variable y, covariates $x = (x_1, \ldots, x_k)$
Linear regression model

$$Y = x^t \beta + \varepsilon$$

Which covariates to include is a question of model choice

$$Y = x(\mathcal{S})^t \beta(\mathcal{S}) + \varepsilon, \quad \mathcal{S} \subset \{1, \ldots, k\}$$

Assumptions about the distribution of ε.

Methods AIC, BIC, FIC ([Claeskens and Hjort, 2003]), full Bayesian etc.

Choice of regression functional

Distribution P

$$T_{\ell_1,\mathcal{S}}(P) = \mathsf{argmin}_{\beta(\mathcal{S})} \int |y - x(\mathcal{S})^t \beta(\mathcal{S})| \, dP(y,x)$$

$$T_{\ell_2,\mathcal{S}}(P) = \mathsf{argmin}_{\beta(\mathcal{S})} \int (y - x(\mathcal{S})^t \beta(\mathcal{S}))^2 \, dP(y,x)$$

Discrete y

$$T_{D_{\mathsf{kl}},\mathcal{S}}(P) = \mathsf{argmin}_{\beta(\mathcal{S})} - \int q(y) \log(p(x(\mathcal{S})^t \beta(\mathcal{S}))/q(y)) \, dP(y,x)$$

with for example

$$p(u) = \frac{\exp(u)}{1 + \exp(u)}$$

Choice of regression functional

Quantile regression. Stack loss data of [Brownlee, 1960], data set provided by [R Core Team, 2013], example in [Koenker, 2010].

R output for 95% confidence interval based on rank inversion

	coefficients	lower bd	upper bd
(Intercept)	-39.68985507	-53.7946377	-24.49145429
stack.xAir.Flow	0.83188406	0.5090902	1.16750874
stack.xWater.Temp	0.57391304	0.2715066	3.03725908
stack.xAcid.Conc	-0.06086957	-0.2777188	0.01533628

Assume a linear regression model with i.i.d. error term ε

$$Y = x\beta + \varepsilon$$

Choice of regression functional

The sum of the absolute residuals without Acid.Conc is 43.694.
Sum with Acid.Conc is 42.081, reduction is 1.613.
Highest daily temperatures in Berlin from 01/01/2015-21/01/2015

$$6, 8, 6, 5, 4, 3, 6, 7, 9, 13, 5, 8, 12, 8, 10, 10, 5, 4, 1, 2, 2\,.$$

Replace Acid.Conc by Cos.Temp.Berlin
Inclusion of Cos.Temp.Berlin reduces sum of absolute residuals by 1.162

Not much worse than Acid.Conc.

Choice of regression functional

Replace Acid.Conc by 21 i.i.d. $N(0,1)$ random variables.

Repeat say 1000 times. In 21.2% of the cases greater decrease in the sum of the absolute residuals than that due to covariate Acid.Conc.

21.2% will be referred to as a p-value, $p = 0.212$.

Replace all three covariates by i.i.d. $N(0,1)$ $p = 1.93e-7$

Choice of regression functional

p-values for the $2^3 = 8$ possibilities

functional	0	1	2	3	4	5	6	7
p-value	1.93e-7	1.41e-2	4.90e-4	2.32e-1	5.02e-9	7.43e-3	2.57e-4	1.00

where $j = j(\mathcal{S}) = \sum_{i \in \mathcal{S}} 2^{i-1}$

A small p-value indicates that the omitted covariates have some influence on the value of the dependent variate, at least one is significant

Choose functionals with high p-values such that all included covariates are significant
Choice is $j = 3$ corresponds to $\mathcal{S} = \{1, 2\}$.

Choice of regression functional

For ℓ_2 regression simple asymptotic approximations for p-values

$$p(\mathcal{S}) \approx 1-\mathsf{pchisq}\left(\frac{n(\|\boldsymbol{y}_n - \boldsymbol{x}_n(\mathcal{S})\beta_{\mathsf{l}sq}(\mathcal{S})\|_2^2 - \|\boldsymbol{y}_n - \boldsymbol{x}_n\beta_{\mathsf{l}sq}\|_2^2)}{\|\boldsymbol{y}_n - \boldsymbol{x}_n(\mathcal{S})\beta_{\mathsf{l}sq}(\mathcal{S})\|_2^2}, k - k(\mathcal{S})\right) .$$

where $\beta_{\mathsf{l}sq}(\mathcal{S}) = T_{\ell_2,\mathcal{S}}(\mathbb{P}_n)$ and $\beta_{\mathsf{l}sq} = T_{\ell_2,\mathcal{S}_f}(\mathbb{P}_n)$ with $\mathcal{S}_f = \{1, \ldots, k\}$.

Non-significance regions

The 'Stack-Loss' data are

$$42, 37, 37, 28, 18, 18, 19, 20, 15, 14, 14, 13, 11, 12, 8, 7, 8, 8, 9, 15, 15$$

with median 15. The sum of the absolute deviations from the median is 145.

The non-significance region is defined as those m such that the difference between $\sum_{i=1}^{21} |\text{stack.loss}_i - m|$ and 145 is of the same order as that which can be obtained by regressing the dependent variable on random noise, that is, the difference is not significant

Non-significance regions

Let $\mathsf{ql1}(\alpha, m)$ denote the α-quantile of the random variable

$$\sum_{i=1}^{21} |\mathsf{stack.loss}_i - m| - \inf_b \sum_{i=1}^{21} |\mathsf{stack.loss}_i - m - bZ_i|.$$

The non-significance region is defined as

$$\begin{aligned}
&\mathcal{NS}(\mathsf{stack.loss}, \mathsf{median}, \alpha) \\
&= \left\{ m : \sum_{i=1}^{21} |\mathsf{stack.loss}_i - m| - \sum_{i=1}^{21} |\mathsf{stack.loss}_i - 145| \leq \mathsf{ql1}(\alpha, m) \right\}.
\end{aligned}$$

Non-significance regions

This can be calculated using simulations and gives

$$\mathcal{NS}(\text{stack.loss}, \text{median}, 0.95) = (11.94, 18.47) \tag{2}$$

which may be compared with the 0.95-confidence interval $[11, 18]$ based on the order statistics.

Covering properties? $\alpha = 0.95$

	n	10	20	50	100
$N(0,1)$	in.reg.	0.940 1.512	0.954 1.040	0.948 0.648	0.942 0.464
	rank	0.968 2.046	0.968 1.198	0.970 0.767	0.964 0.530
$C(0,1)$	in.reg.	0.960 3.318	0.956 1.670	0.960 0.958	0.952 0.629
	rank	0.978 5.791	0.950 1.850	0.968 1.069	0.964 0.700
χ^2_1	in.reg.	0.944 1.368	0.936 0.877	0.932 0.550	0.942 0.396
	rank	0.982 2.064	0.958 1.086	0.970 0.675	0.968 0.452
Pois(4)	in.reg.	0.934 1.918	0.925 0.993	0.926 0.288	0.938 0.071
	rank	0.996 3.948	0.964 2.342	0.997 1.573	1.000 1.085

Non-significance regions

Asymptotics. Y_i i.i.d. with median m and density f

$$\left[\mathsf{med}(\boldsymbol{Y}_n) - \sqrt{\frac{\mathsf{qchisq}(\alpha, 1)}{4f(m)^2 n}},\ \mathsf{med}(\boldsymbol{Y}_n) + \sqrt{\frac{\mathsf{qchisq}(\alpha, 1)}{4f(m)^2 n}}\right]$$

Method does not require an estimate of $f(m)$.

Non-significance regions

Requires linear regression model with true parameter values. Covering frequencies and average interval lengths for data generated according to

$$Y = \beta_1 + \beta_2 \cdot Air.Flow + \beta_3 \cdot Water.Temp + \beta_4 \cdot Acid.Conc + \varepsilon$$

with $\beta_i, i = 1, \ldots, 4$ the ℓ_1 estimates and different distributions for the error term: $\alpha = 0.95$.

		β_2		β_3		β_4	
residuals	in.reg.	0.944	0.265	0.982	0.682	0.998	0.248
	rank	0.976	0.390	0.970	1.205	0.970	0.273
Normal	in.reg.	0.954	0.381	0.946	1.042	0.964	0.442
	rank	0.974	0.435	0.956	1.208	0.962	0.542
Laplace	in.reg.	0.953	0.501	0.959	1.375	0.952	0.580
	rank	0.966	0.594	0.959	1.697	0.960	0.761
Cauchy	in.reg.	0.928	1.467	0.942	4.052	0.936	1.731
	rank	0.936	1.948	0.946	5.676	0.942	2.984

An attitude of mind

D. W. Müller, Heidelberg

> *... distanced rationality. By this we mean an attitude to the given, which is not governed by any possible or imputed immanent laws but which confronts it with draft constructs of the mind in the form of models, hypotheses, working hypotheses, definitions, conclusions, alternatives, analogies, so to speak from a distance, in the manner of partial, provisional, approximate knowledge.*

(Thesen zur Didaktik der Mathematik)

References

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second international symposium on information theory*, pages 267–281, Budapest. Acadamiai Kiado.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

[Akaike, 1981] Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3–14.

[Berk et al., 2013] Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41(2):802–837.

[Birnbaum, 1962] Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–326.

[Brownlee, 1960] Brownlee, K. A. (1960). *Statistical Theory and Methodology in Science and Engineering*. Wiley, New York, 2nd edition.

[Buja et al., 2009] Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A*, 367:4361–4383.

[Claeskens and Hjort, 2003] Claeskens, G. and Hjort, N. L. (2003). Focused information criterion. *Journal of the American Statistical Association*, 98:900–916.

[Cox, 2006] Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

[Davies, 2014] Davies, L. (2014). *Data Analysis and Approximate Models*. Monographs on Statistics and Applied Probability 133. CRC Press.

[Davies, 1995] Davies, P. L. (1995). Data features. *Statistica Neerlandica*, 49:185–245.

[Davies, 2008] Davies, P. L. (2008). Approximating data (with discussion). *Journal of the Korean Statistical Society*, 37:191–240.

[Davies et al., 2012] Davies, P. L., Höhenrieder, C., and Krämer, W. (2012). Recursive estimation of piecewise constant volatilities. *Computational Statistics and Data Analysis*, 56(11):3623–3631.

[Donoho, 1988] Donoho, D. L. (1988). One-sided inferences about functionals of a density. *Annals of Statistics*.

[Gelman, 2003] Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, 71(2):369–382.

[Hampel, 1998] Hampel, F. R. (1998). Is statistics too difficult? *Canadian Journal of Statistics*, 26(3):497–513.

[Hampel et al., 1986] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[Huber, 2011] Huber, P. J. (2011). *Data Analysis*. Wiley, New Jersey.

[Huber and Ronchetti, 2009] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, New Jersey, second edition.

[Koenker, 2010] Koenker, R. (2010). quantreg: Quantile regression. http://CRAN.R-project.org/package=quantreg. R package version 4.53.

[Neyman et al., 1953] Neyman, J., Scott, E. L., and Shane, C. D. (1953). On the spatial distribution of galaxies a specific model. *Astrophysical Journal*, 117:92–133.

[Neyman et al., 1954] Neyman, J., Scott, E. L., and Shane, C. D. (1954). The index of clumpiness of the distribution of images of galaxies. *Astrophysical Journal Supplement*, 8:269–294.

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Schwarz, 1978] Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

[Stigler, 1977] Stigler, S. M. (1977). Do robust estimators work with real data? (with discussion). *Annals of Statistics*, 5(6):1055–1098.

[Tukey, 1977] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Massachusetts.

[Tukey, 1993] Tukey, J. W. (1993). Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper. Princeton University, Princeton.

[Uhrmann-Klingen, 1995] Uhrmann-Klingen, E. (1995). Minimal Fisher information distributions with compact supports. *Sankhya Series A*, 57:360–374.

[Xia and Tong, 2011] Xia, Y. and Tong, H. (2011). Feature matching in time series modelling. *Statistical Science*, 26(1):21–46.