TRAINING in EUROPEAN ASSET MANAGEMENT

Statistical Tools for Making Predictions

LCPC Training Week – 5 October 2010

Dr Colin Caprani

www.colincaprani.com

TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

Introduction

About me:

- Did PhD in Bridge Traffic Loading (BTL) in UCD, 2001-2005
- Concentrated on statistical analysis of BTL
- Continuing research in the BTL (and new areas)

About this talk:

- Cover the tools of research rather than particular research results
- Useful basics in statistics for extrapolation
- Identify some potential pitfalls
- Cover some useful aspects of computation

But first a warning…

Warning!

Statistics in the hands of an engineer are like a lamppost to a drunk – they're used more for support than illumination.

A. E. Housman.

Topics

- Extreme Value Statistics

  Because we are mostly interested in the highest, smallest, biggest…

- Statistical Inference

  Because we need to use a model to predict outside the data…

- Prediction

  Because it is the goal of our analysis…

- Computational Tools

  Because the right tool makes the job easier…

TRAINING in EUROPEAN ASSET MANAGEMENT

Extreme Value Statistics

Because we are mostly interested in the highest, smallest, biggest…

Extreme Value Statistics

Take a school with 10 classes of 30 children…

- What is the probability of a child in a class over 180 cm?

Note: heights of all 300 children is the 'Parent' distribution

Can approach this by asking:

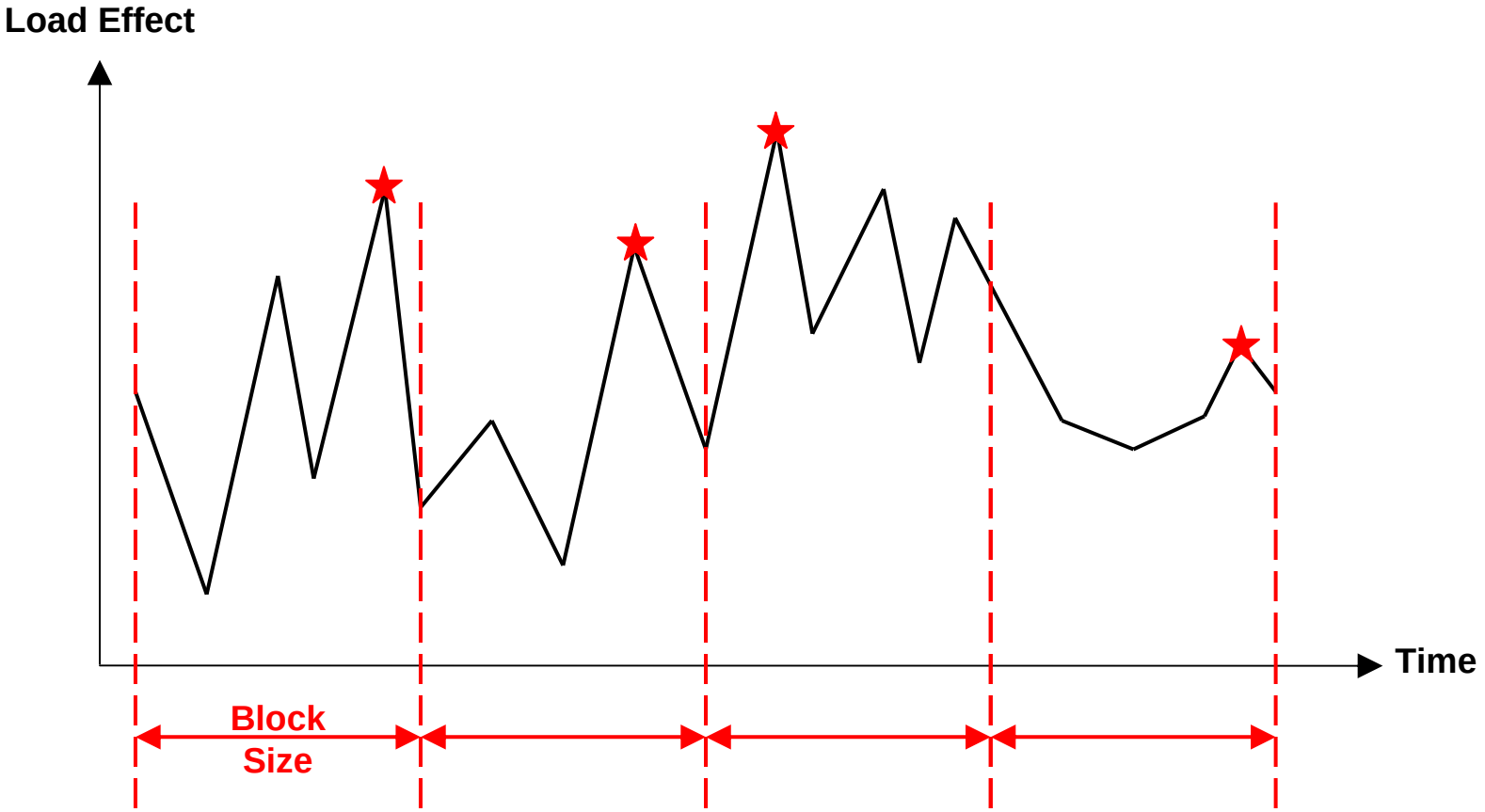
- What is the distribution of tallest child in a class?
- What is distribution of children over 160 cm (say)?

These different approaches are:

- Block maximum (i.e. the classroom)
- Peaks over threshold (i.e. those over 160 cm)

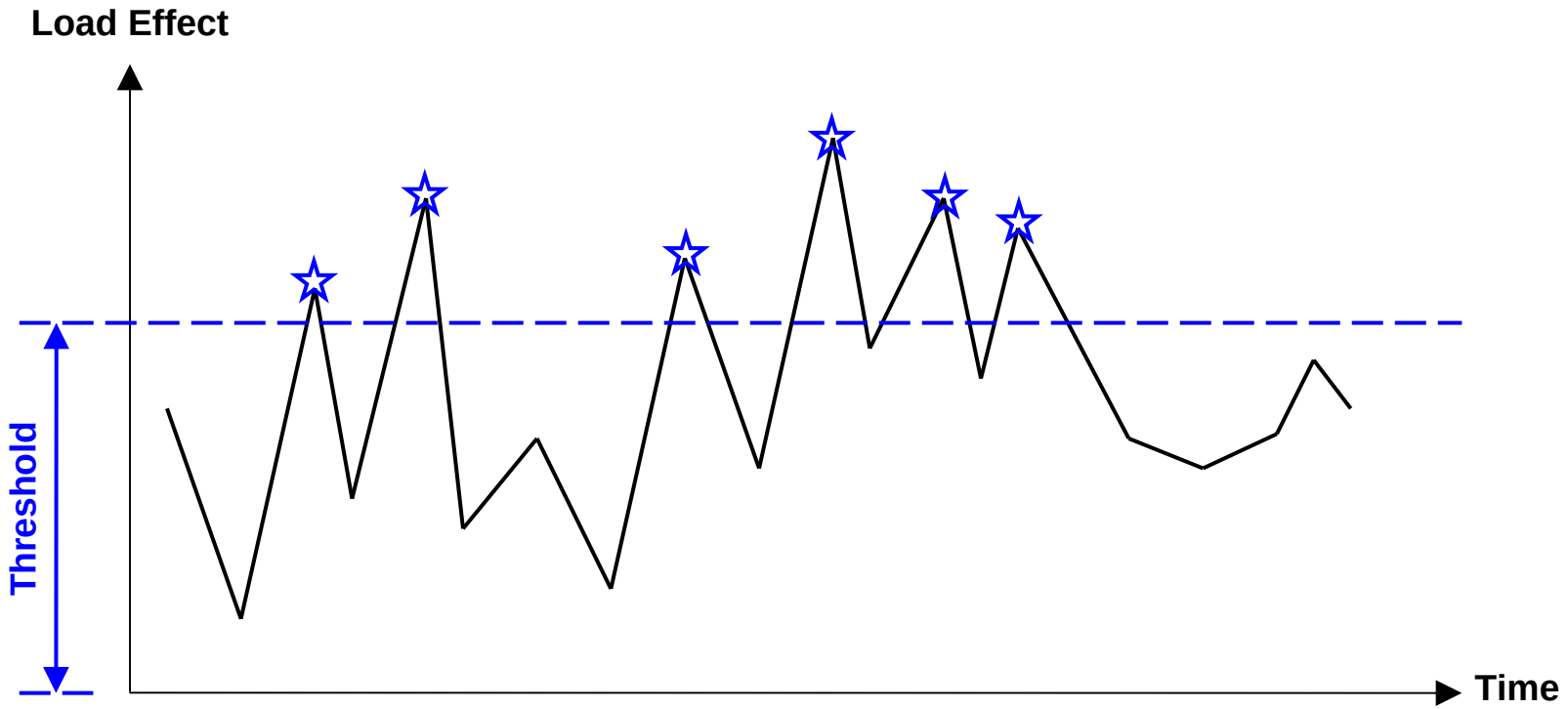
Extreme Value Statistics

Block maxima approach – data modelled using GEV distribution:



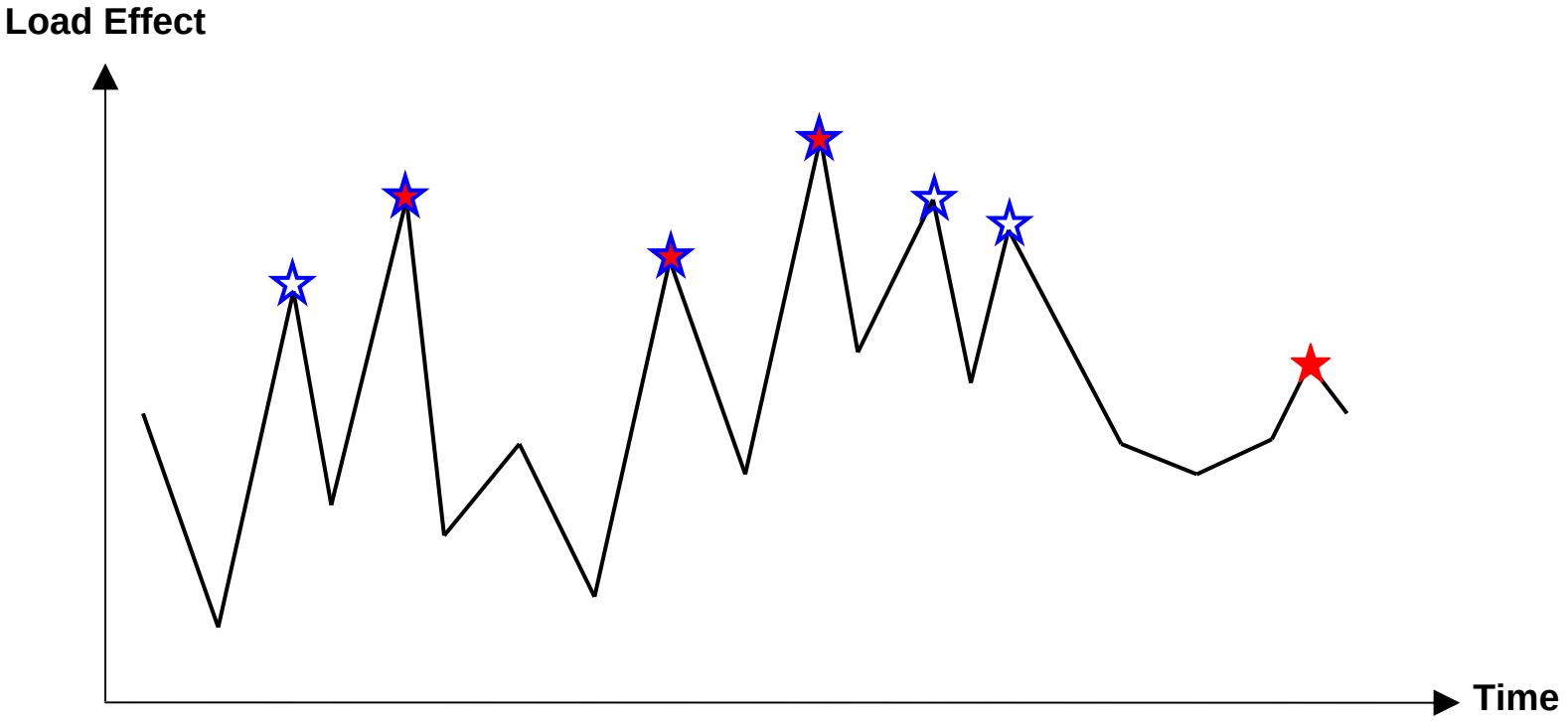
Extreme Value Statistics

Peaks Over Threshold (POT) – data modelled using GPD distribution:



Extreme Value Statistics

Note there can be differences in the approaches:



Extreme Value Statistics

A nice new solution is the Box-Cox-GEV (BCGEV) distribution:

- Introduced by Bali (2003) for use in economic modelling
- Includes both GEV and GPD distributions through a model parameter, λ
- Maintains the usual GEV/GPD parameter set, (μ, σ, ξ):

$$H(x)=\left(\frac{1}{\lambda}\right)\left(\left[\exp\left\{-\left[h(x)\right]_{+}^{1/\xi}\right\}\right]^{\lambda}-1\right)+1 \qquad \textbf{where} \qquad h(x)=1-\xi\left(\frac{x-\mu}{\sigma}\right)$$

Thus, as:

- $\lambda \to 1$, BCGEV $\to$ GEV distribution
- $\lambda \to 0$, BCGEV $\to$ GPD distribution (by L'Hopital's rule)

Extreme Value Statistics – Block Maximum Approach

- Consider *n* random variables: $X_1, \ldots, X_n$
  (e.g. child heights in class of 30 children)

- Take the maximum of these: $Y = \max\left[X_1, \ldots, X_n\right]$

- What is the distribution of *Y* – the maximum height of child?

$$F_Y(y) = P\left[Y \le y\right] = P\left[x_1 \le y; \ldots; x_n \le y\right]$$

- If no correlation (not a basketball class!) and well-mixed (not all boys)
  (requirement of *independent and identically distributed* - iid)

$$F_Y(y) = P\left[Y \le y\right] = \prod_{i=1}^{n} P\left[X_i \le y\right] = \left[F_X(y)\right]^n$$

Extreme Value Statistics – Asymptotic Distributions

What is distribution of maximum: $F_Y(y) = \left[F_X(y)\right]^n$ as *n* gets very large?

- Depends on *F(x)* of course!
- There are three asymptotic families (Gumbel, Frechet, Weibull)
- All 3 included in the Generalized Extreme Value distribution:

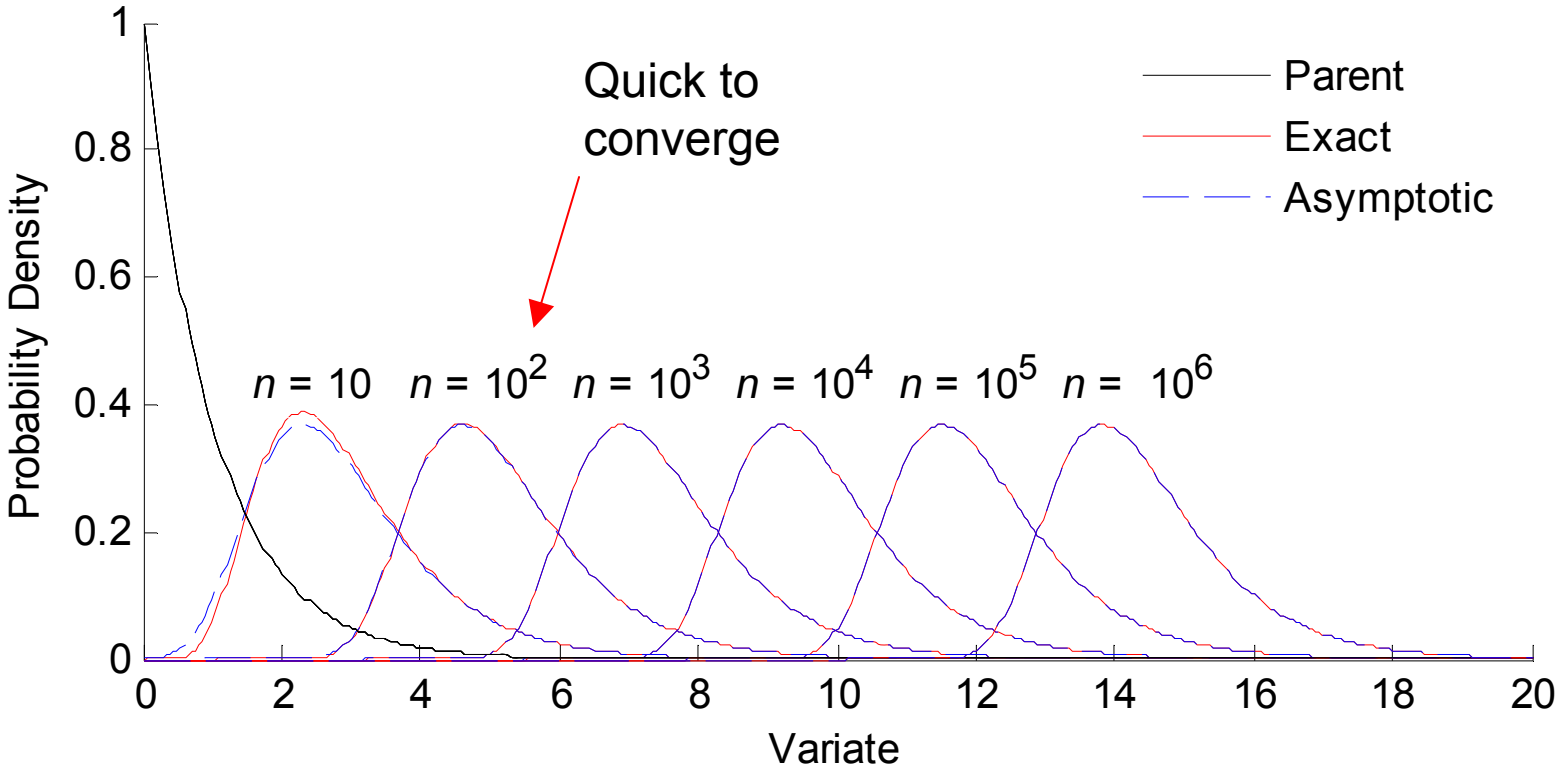
$$G(y) = \exp\left\{-\left[1-\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{1/\xi}\right\}$$

There are different rates of convergence for different parent distributions

- This can have huge impact on results…

Extreme Value Statistics – Asymptotic Distributions

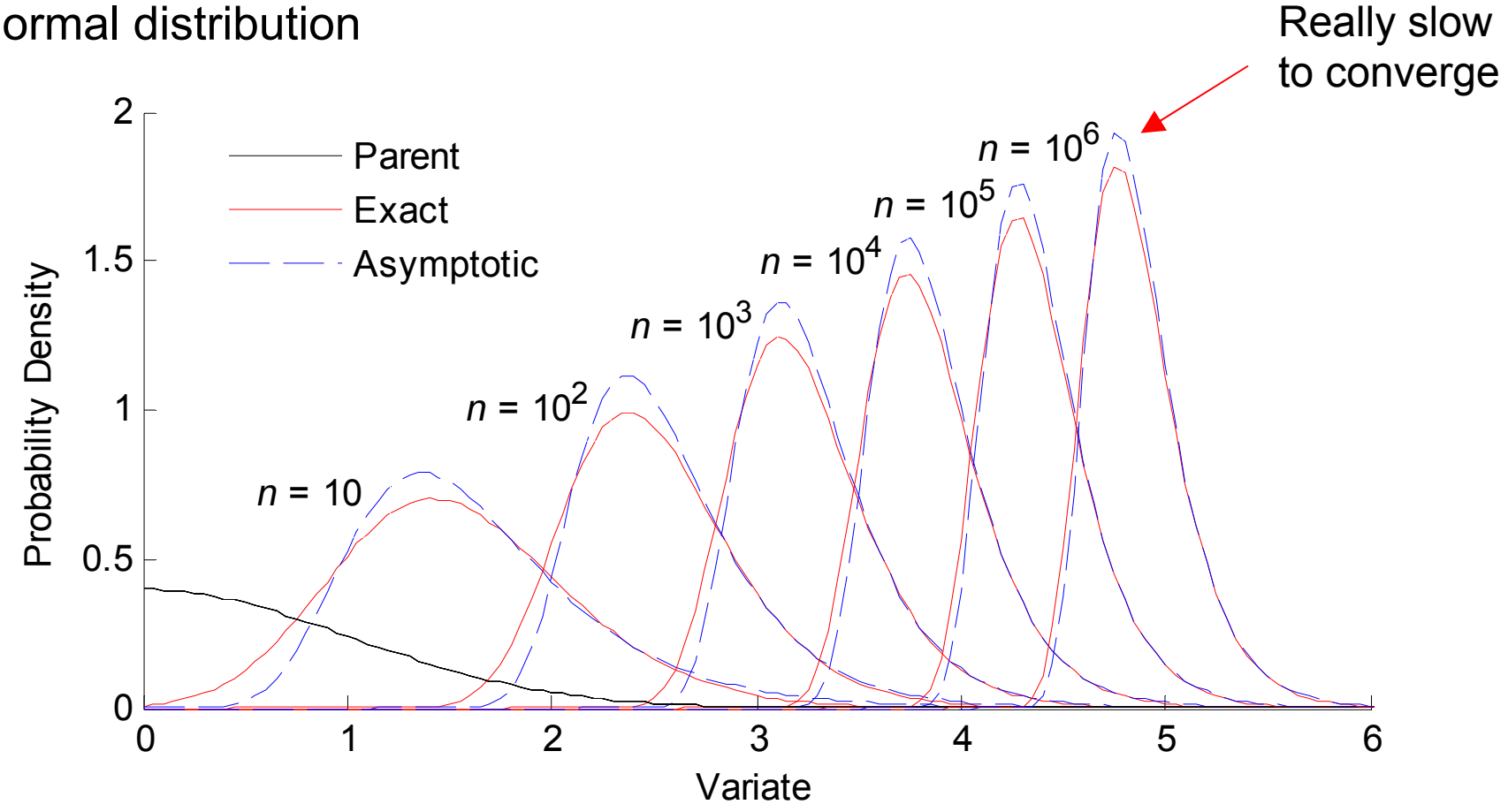
Exponential distribution



TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

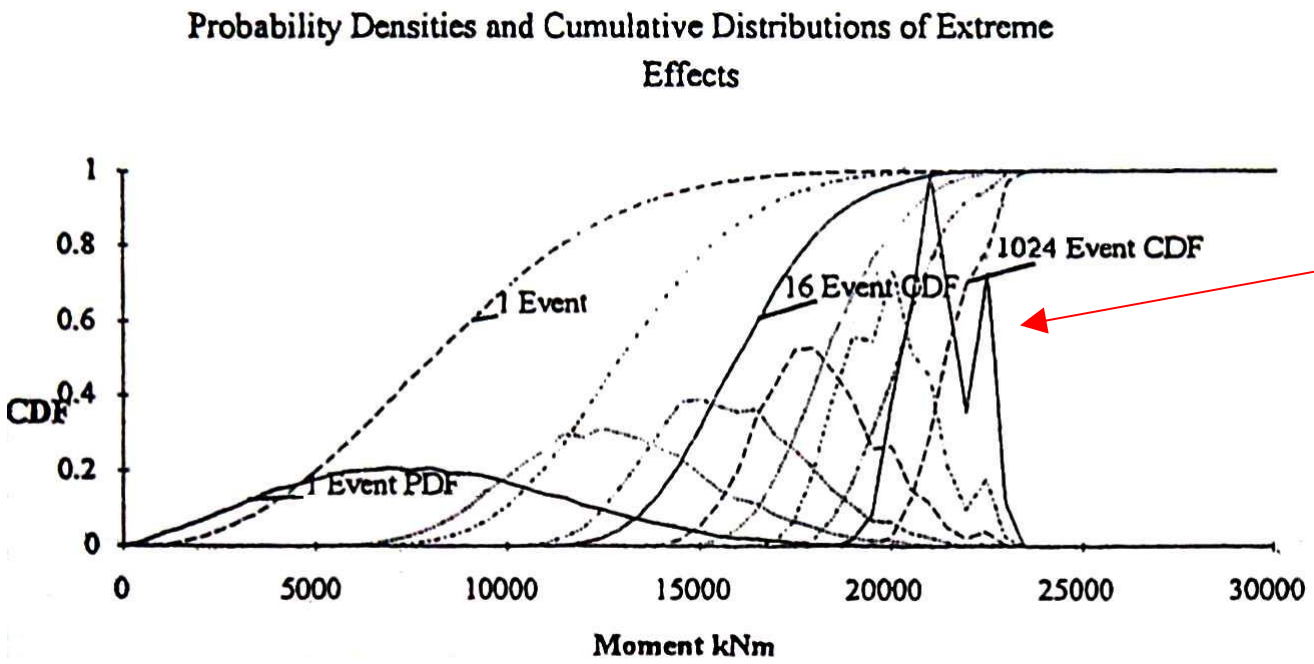
Extreme Value Statistics – Asymptotic Distributions

Normal distribution



Extreme Value Statistics – Fitting EV Distributions

Do not raise measured PDF to a power!



Peaks due to one or two events will dominate extrapolation

Extreme Value Statistics – Stability Postulate

If the parent distribution has the form of *G(x)* then the exact and asymptotic distributions are the same for any *n.*

This is really powerful!

- GEV is a very flexible distribution (3 parameters), so…
- Use GEV to fit your parent data: $G(x;\mu,\sigma,\xi)$
- Then the distribution of extreme is known for any *n*:

$$G_{\max}(x;\mu_n,\sigma_n,\xi_n) = G^n(x;\mu,\sigma,\xi)$$

$$\mu_n = \frac{\sigma}{\xi}\left(1-\frac{1}{n^{\xi}}\right)+\mu \qquad \sigma_n = \frac{\sigma}{n^{\xi}} \qquad \xi_n = \xi$$

Thus fit is done to much more data and is more reliable

Statistical Inference

Because we need to use a model to predict outside the data…

Statistical Inference (or Fitting Distributions)

Typical (bad) approach:

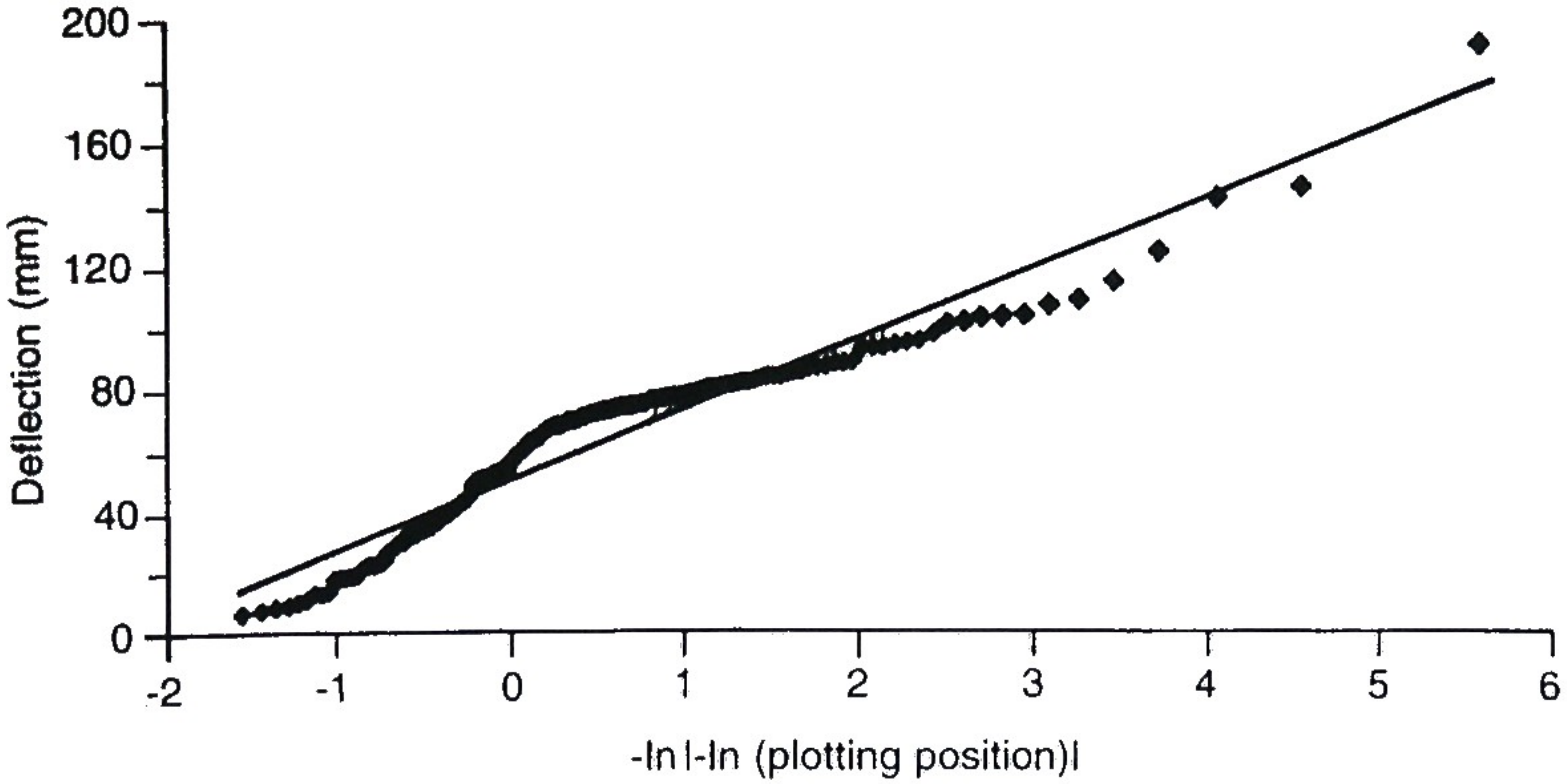
- get data vector *x*
- Use some formula to get corresponding *y*-values
- Fit (*x*,*y*) pairs using least squares

Avoid subjectivity:

- As engineers we love straight lines and least squares – be wary!
- To fit (*x*,*y*) pairs we need a *y* value – where does this come from?

Statistical Inference (or Fitting Distributions)

Example dodgy practice:



TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

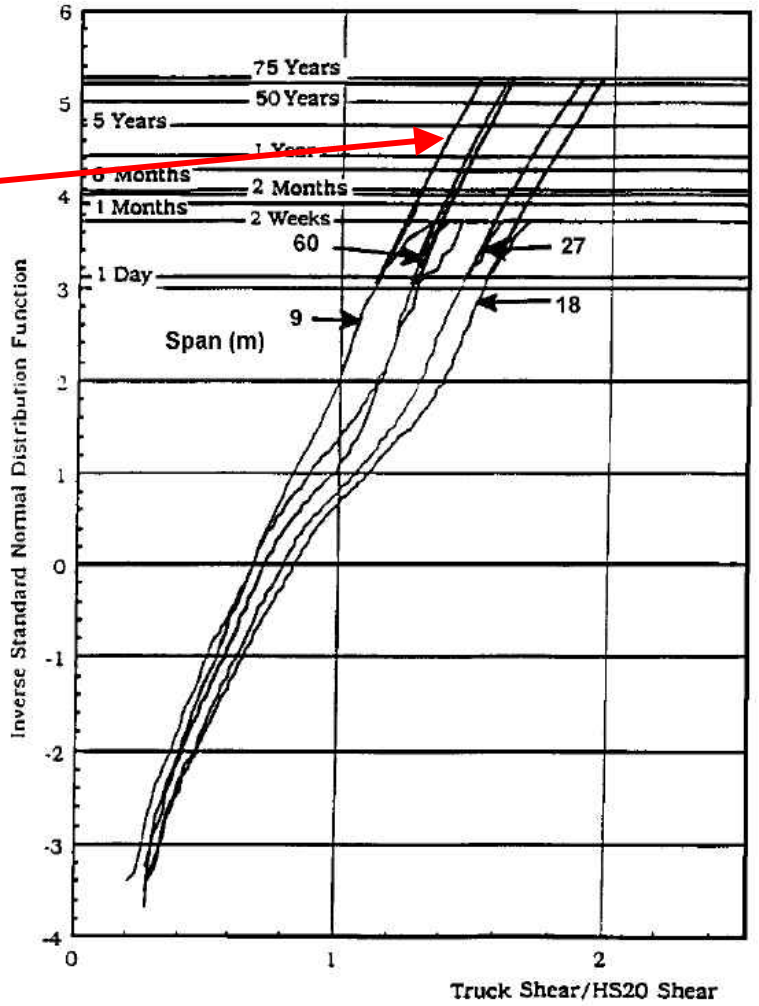
Statistical Inference (or Fitting Distributions)

Another example of dodgy practice:

Arbitrary straight lines used for extrapolation

The goal is to avoid subjectivity so that it does not matter who 'draws the line' – the data should be the only thing that determines the prediction

Maximum Likelihood gives objective & minimum-variance estimates in general



Statistical Inference – Maximum Likelihood

RA Fisher's Idea:

• examine the probability of having observed the data that was observed, given the proposed probability model

Consider GEV as example:

CDF: $$G(x) = \exp\left\{-\left[1-\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{1/\xi}\right\}$$

PDF: $$g\left(x;\theta\right) = G\left(x;\theta\right)\cdot\sigma^{-1}\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi-1}$$

Statistical Inference – Maximum Likelihood

Probability of single data point being realized

≈ PDF value at data point, given trial parameter set

$$L(\theta) = L(\theta; x) \propto g(x; \theta)$$

Probability of exact data set being realized

≈ Multiply individual data probabilities

$$L(\theta) \propto \prod_{i=1}^{n} g(x_i; \theta)$$

Statistical Inference – Maximum Likelihood

This leads to very small numbers so deal with logs and add…

$$\log L(\theta) = l(\theta) = \log\left[\prod_{i=1}^{n} g(x_i;\theta)\right] = \sum_{i=1}^{n} \log\left[g(x_i;\theta)\right]$$

So for the GEV distribution:

$$l(\theta;x) = -n\log\sigma - \left(1-\frac{1}{\xi}\right)\sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n} x_i^{-1/\xi}$$

Maximization of log-likelihood optimizes the parameters
(i.e. above log-likelihood is the 'objective function' of an optimization)

Statistical Inference – Likelihood Surface

The likelihood function gives a lot of information about the parameter fit

Example:

- 8% of balls are black – the rest are white
- Model with 1-parameter distribution (binomial)

- Take 50 samples and find 4 are black
- Take 100 samples and find 8 are black

- Both cases give the Maximum Likelihood Estimate (MLE) as 8%
- But which gives more ‘information’?

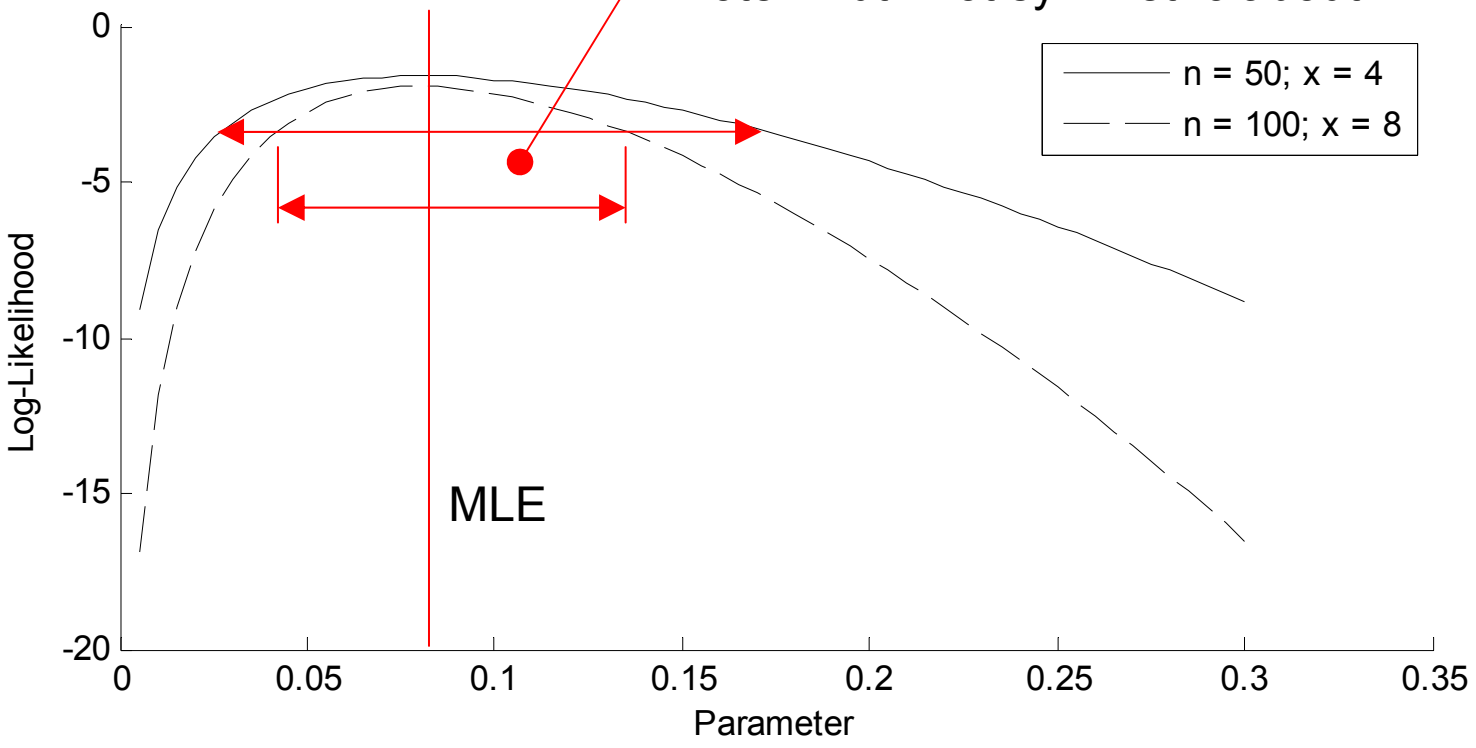
Statistical Inference – Likelihood Surface

Binomial Distribution

- 1-parameter 'surface'

Width reduces with more data reflecting increased certainty about ML parameter

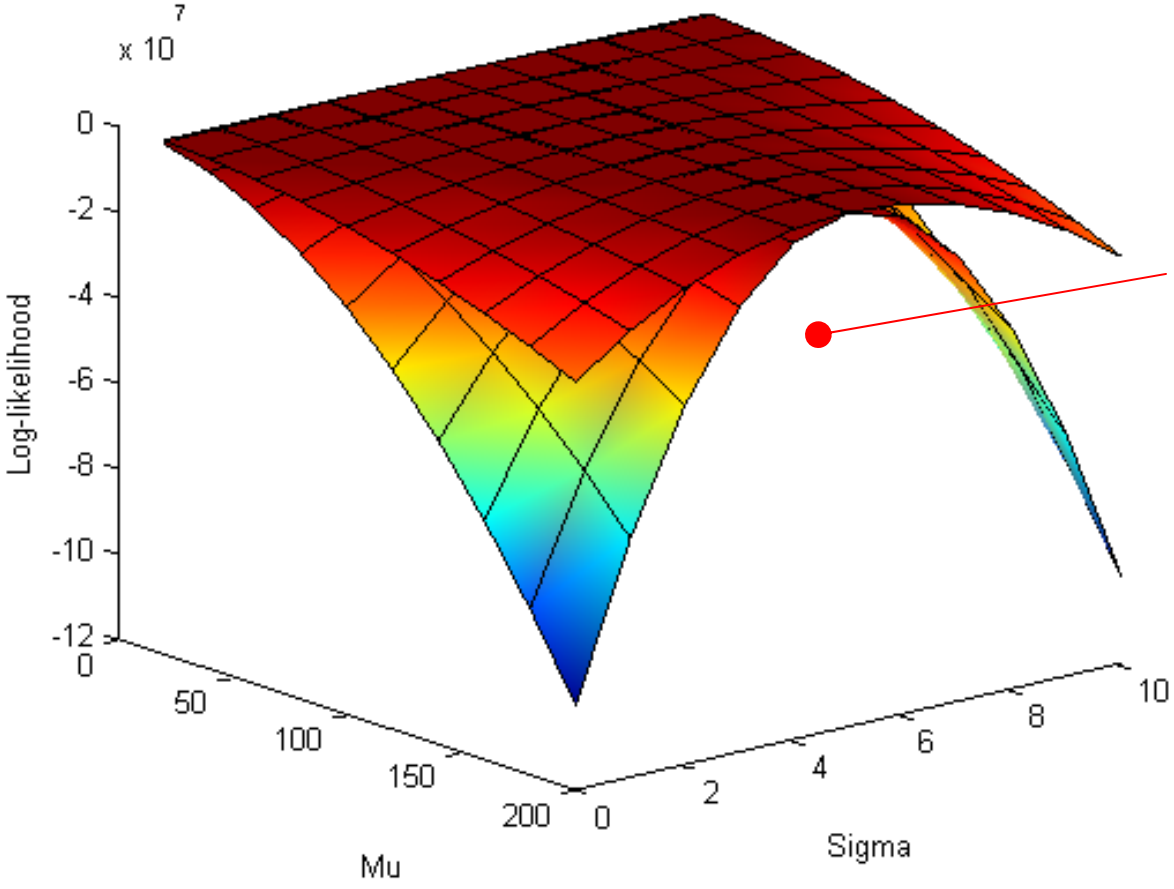
Note: width not symmetric about MLE



TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

Statistical Inference – Likelihood Surface

Normal Distribution



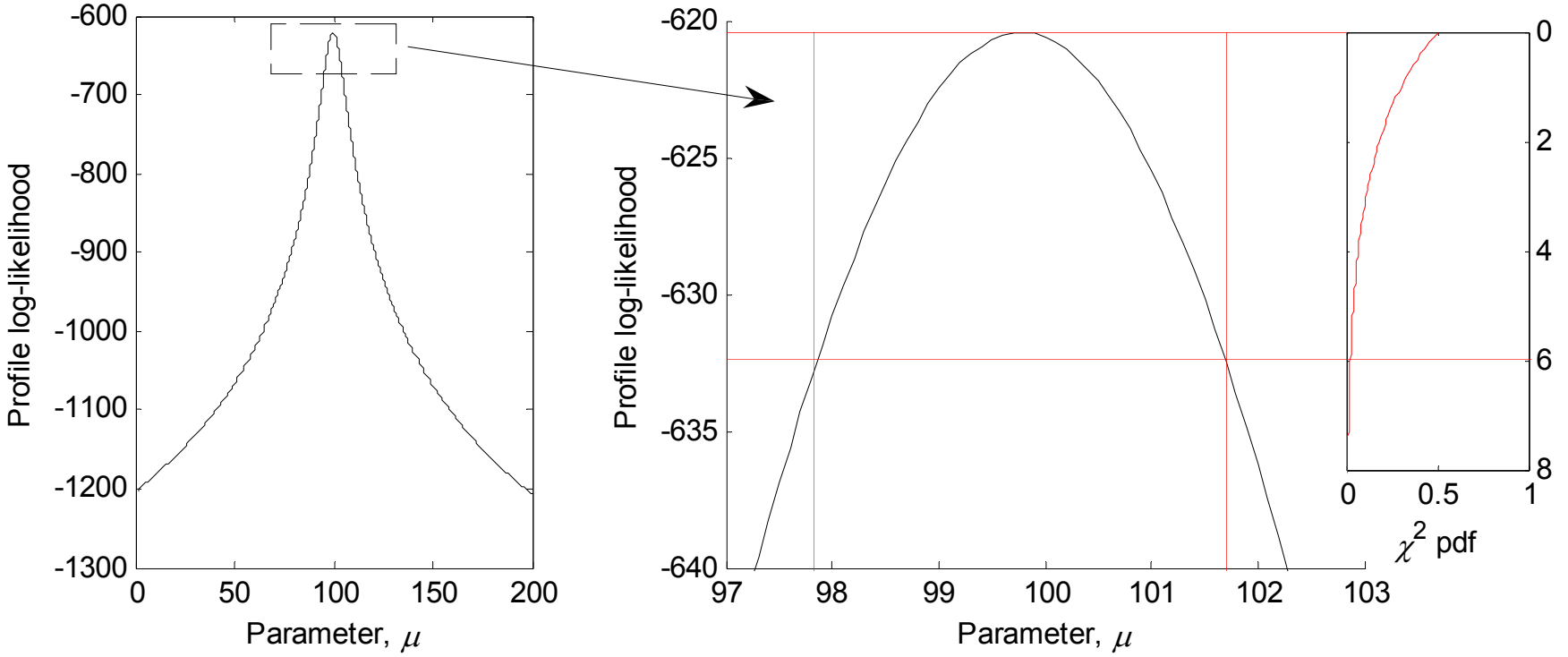
Log-likelihood surfaces of $N(100,5^2)$ for n = 50 and 200.

'Volume' under surface reduces with more data reflecting increased certainty about parameter estimates

Statistical Inference – Profile Likelihood

Can use the surface to get confidence intervals for each parameter

- Look at a 'slice' of surface along each dimension
- No need to assume parameters are normal distributed



Statistical Inference – Other Methods

The likelihood function requires good starting parameter values to obtain a global (and not local) maximum. Some useful methods are:

- Method of Moments (normal method): best for 2-parameter distributions
- Probability Weighted Moments: good for 3-parameters.

For GEV with data *x:*

$$b_r = n^{-1}\sum_{j=1}^{n}\frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)}x_j \qquad c = \frac{2b_1 - b_0}{3b_2 - b_0} - \frac{\log 2}{\log 3}$$

$$\hat{\xi} = 7.8590c + 2.9554c^2 \quad \hat{\sigma} = \frac{\hat{\xi}(2b_1 - b_0)}{\Gamma(1+\hat{\xi})(1-2^{-\hat{\xi}})} \quad \hat{\mu} = b_0 + \frac{\hat{\sigma}}{\hat{\xi}}\left[\Gamma(1+\hat{\xi}) - 1\right]$$

There are many others…

Prediction

Because it is the goal of our analysis…

Prediction

Interested in return level, *z*, at probability of occurrence, *p*

Example:

- 1000 events per day
- 250 ('economic') days per year
- Maximum event value in 100 years occurs at

$$p = 1 - \frac{1}{1000 \times 250 \times 100} = 0.99999996$$

Prediction – Return Period

Engineering designs use this concept

- *R* is the inverse of the probability of the event

  e.g. 10 year event has probability of exceedence 1/10 = 0.10 in one year or, 1/(250×10) = 0.0004 in one day

Design codes often say (e.g.) 10% probability of exceedence in 100 years:

- Approximate associated return period is: $R = \frac{100}{0.1} = 1000$ years
- Exact is: $R = \frac{1}{1-\left(1-0.1\right)^{\frac{1}{100}}} = 950$ years

Be careful!

Prediction Accuracy – Delta Method

- Using ML usually returns the variance-covariance matrix of the parameters
- Also the gradients (first-order derivatives) are got
- The return level is a function of the parameters:

$$z = f(\theta)$$

- Take a Taylor-series expansion:

$$z = f(\theta) \approx g(\hat{\theta}) + (\theta - \hat{\theta}) \frac{d f(\hat{\theta})}{d\theta}$$

- Transformation of variables gives:

$$\mathrm{Var}(z) = \left[\frac{d}{d\theta} f\left(\hat{\theta}\right) \right]^2 \cdot \mathrm{Var}(\theta)$$

Prediction Accuracy – Delta Method

For several parameters this is:

$$\mathrm{Var}(z) = \nabla f(\boldsymbol{\theta}) \cdot \mathbf{V}_\theta \cdot \nabla f(\boldsymbol{\theta})^T$$

where: $\mathbf{V}_\theta$ is the var-covar matrix and $\nabla f(\boldsymbol{\theta})$ is the vector of gradients

For GEV: $\hat{z} = f\left(p;\hat{\theta}\right) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \cdot \left[1 - p^{\hat{\xi}}\right]$

$$\nabla f(\theta) = \nabla \hat{z}^T$$

$$= \left[\frac{\partial z}{\partial \mu}; \quad \frac{\partial z}{\partial \sigma}; \quad \frac{\partial z}{\partial \xi}\right]$$

$$= \left[1; \quad \hat{\xi}^{-1}\left(1 - p^{\hat{\xi}}\right); \quad \sigma\hat{\xi}^{-2}\left(1 - p^{\hat{\xi}}\right) - \sigma\hat{\xi}^{-1} p^{\hat{\xi}} \log p\right]$$

This assumes normal distribution for z

Prediction Accuracy – Profile Likelihood

Can use the likelihood surface to get confidence intervals for prediction:

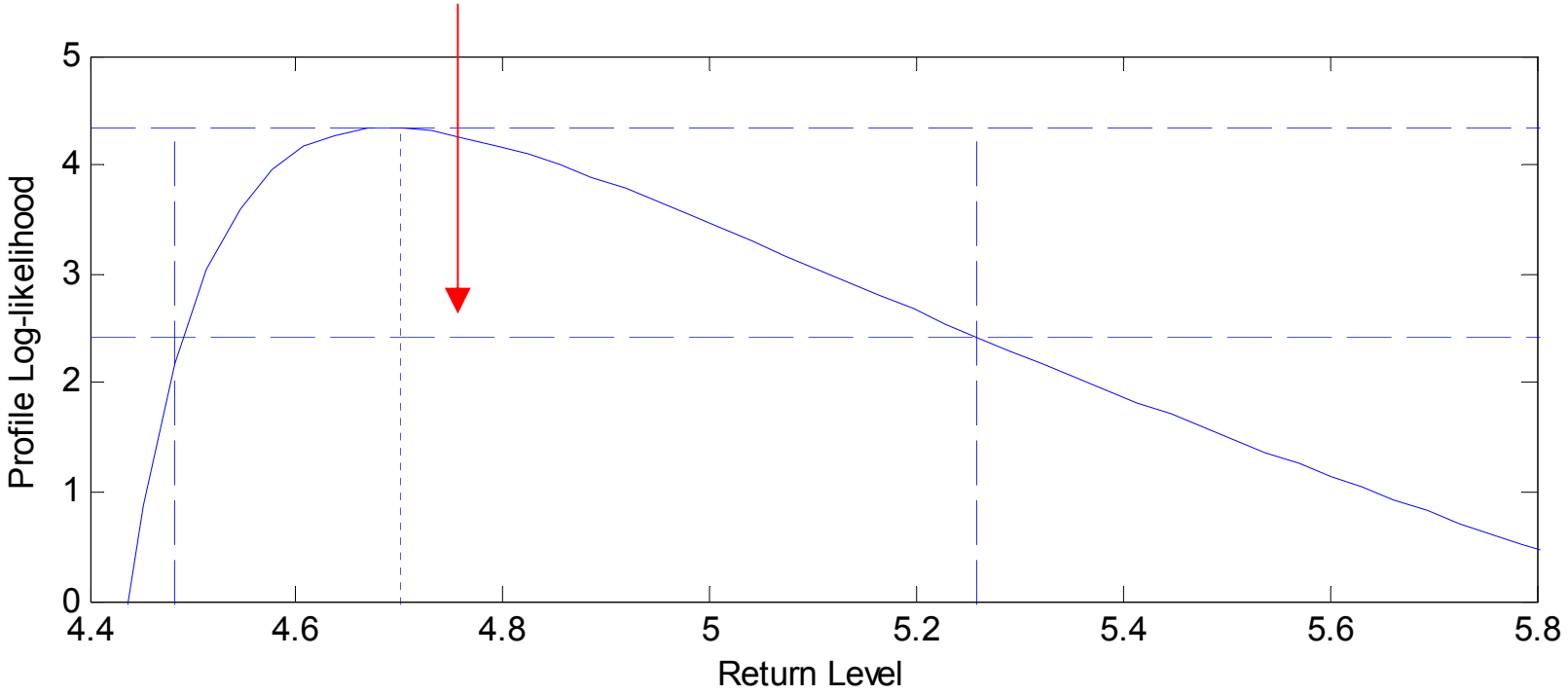
- return level of z at probability p
- rearrange equation so z is now a parameter
- Find MLE
- Draw profile likelihood of z
- Estimate confidence intervals

See Coles (2001) for more information

Prediction Accuracy – Profile Likelihood

Notice confidence intervals are not symmetric about MLE

Beware the assumption of normality!



Prediction Accuracy – Predictive Likelihood

Can jointly fit data and a postulated prediction, *z*, at a probability, *p*, to get the joint likelihood of observing the prediction, given the data, *x:*

$$L_P\left(z \mid x\right) = \max_{\theta} L_x\left(\theta; x\right) L_z\left(\theta; z\right)$$

$$L_y\left(\theta; x\right) = \prod_{i=1}^{n} g\left(x_i; \theta\right) \quad \text{and} \quad L_z\left(\theta; z\right) = g_z\left(z; \theta\right)$$

$$\log\left[L_P\left(z \mid x\right)\right] = \max_{\theta}\left\{\sum_{i=1}^{n} \log\left[g(x_i;\theta)\right] + \log\left[g_z(z;\theta)\right]\right\}$$

Repeat for different *z* and find distribution of prediction from the relative values of L_P

Prediction Accuracy – Predictive Likelihood

For GEV, at a level of *m* repetitions of the data sampling period

$$g\left(x;\theta\right)=G\left(x;\theta\right)\cdot\sigma^{-1}\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi-1}$$

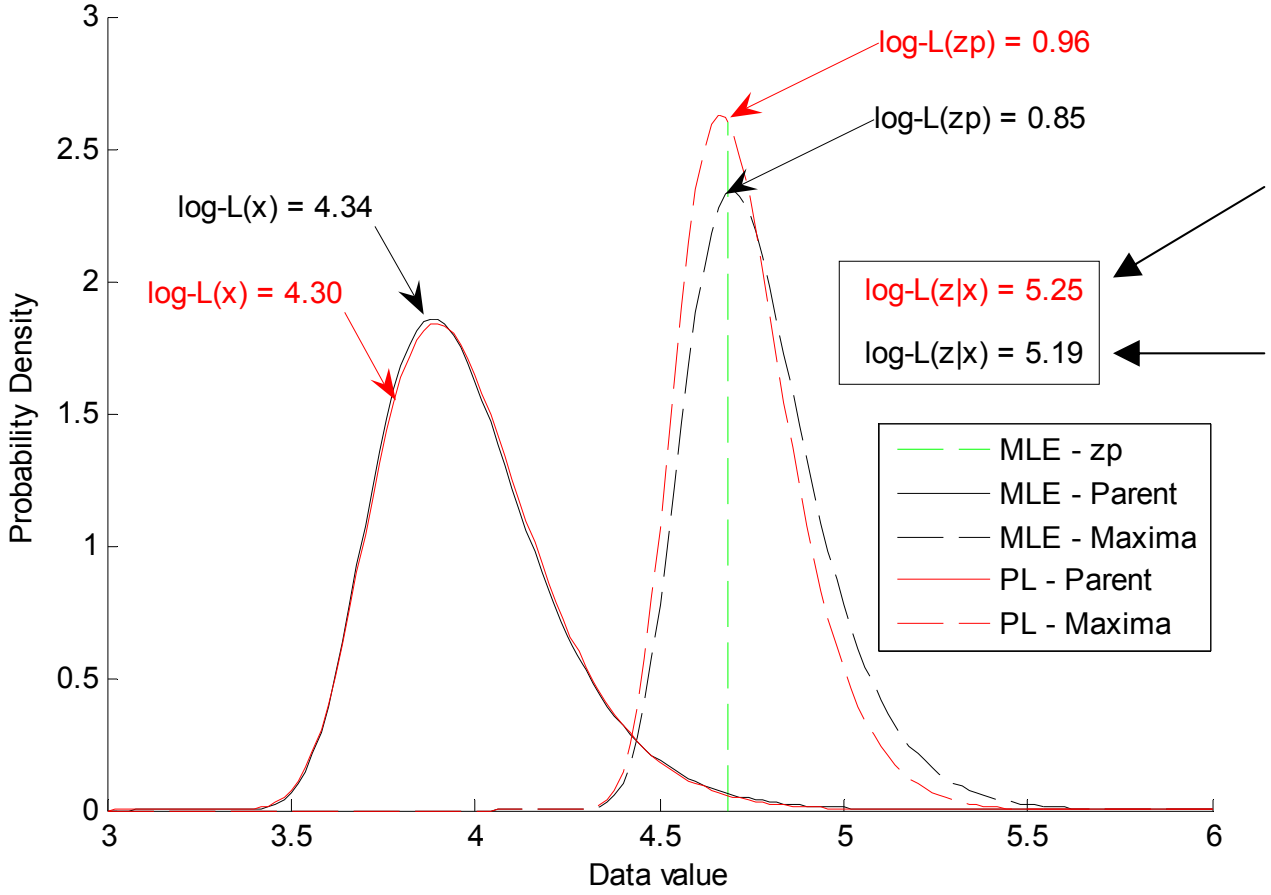
$$g_z\left(z;\theta\right)=m\cdot g\left(z;\theta\right)\cdot\left[G\left(z;\theta\right)\right]^{m-1}$$

Since *m* and *z* are known, we maximize parameters only

TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

Prediction Accuracy – Predictive Likelihood

Single optimization and resulting L_P value



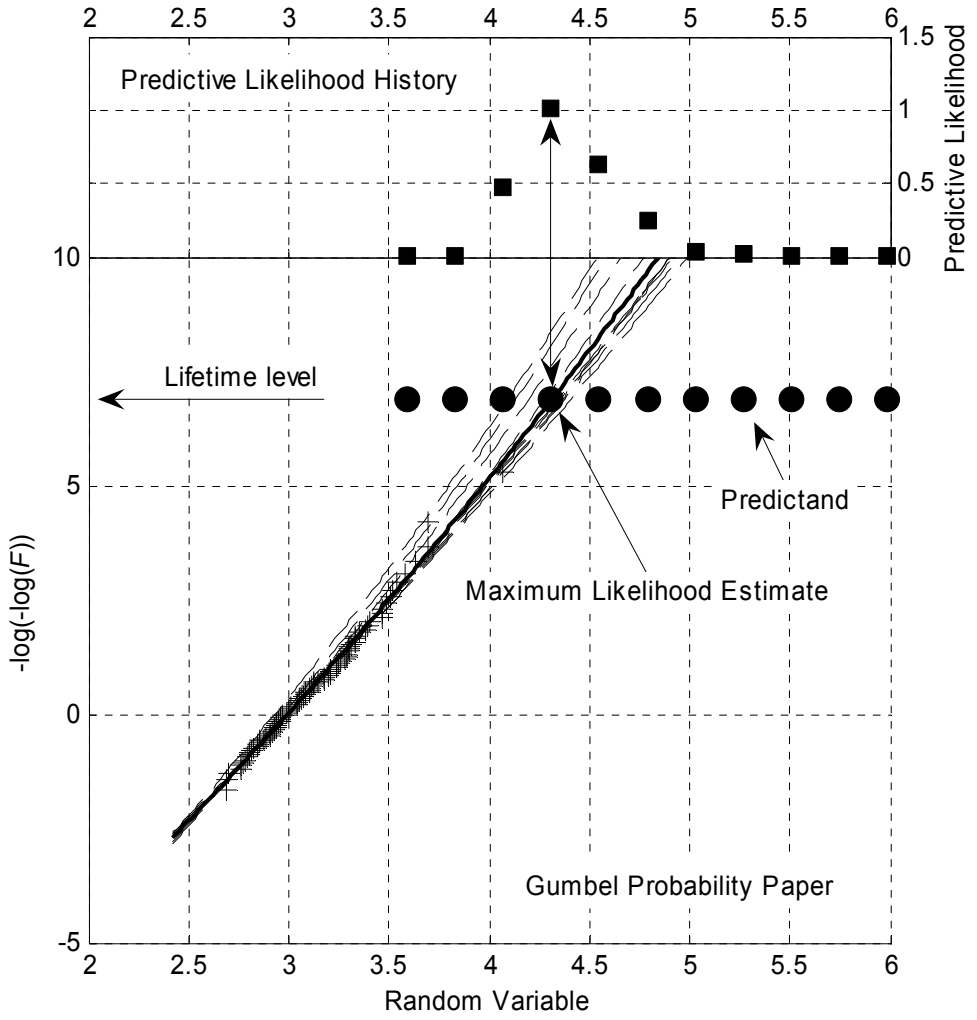
Fit to data and prediction together

Fit to data only

Prediction Accuracy – Predictive Likelihood

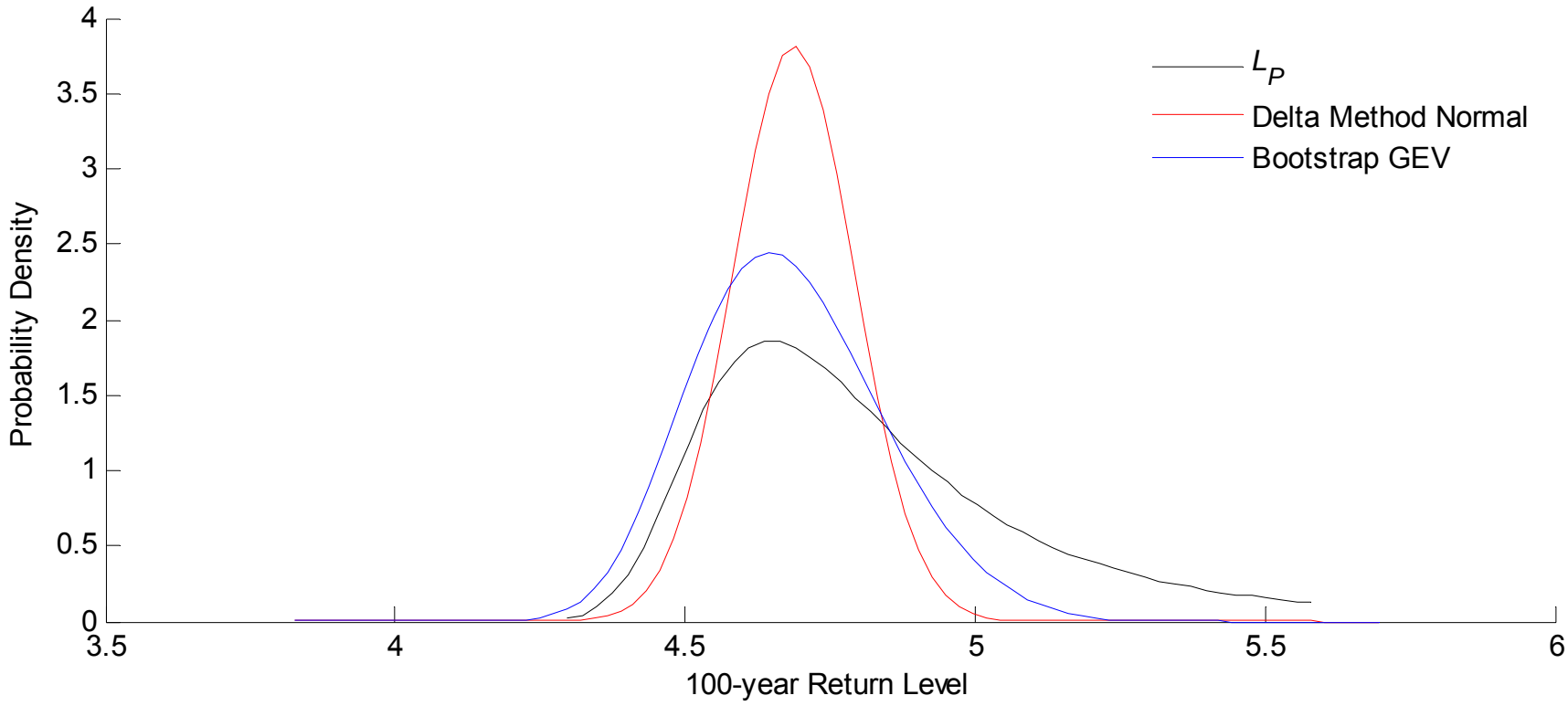
All optimizations

Ratios of L_P values obtained used to estimate distribution of z



Prediction Accuracy – Predictive Likelihood

Result comparison



Computational Tools

Because the right tool makes the job easier…

Computational Tools

What follows are just my own opinions…

- Your problem may be different
- You may have different preferences
- Other considerations may apply (e.g. research team preference, using old code, supervisor's wishes, etc…)

With the proviso done…

Computational Tools – Random Number Generator

Your RNG is the workhorse – get it right by DieHard testing it…

RNGs repeat after a period – some can be quite short ($2^{32} - 1$)

RNGs repeat the same sequence if the seed is the same:

good for debugging – bad if you forget and use for results!

Beware system RNGs:

- MS Excel – very bad!
- Matlab – good in later versions (7.7+)
  – beware: the default seed is always the same

Computational Tools – Random Number Generator

Generation of RNs of different distributions often requires multiple uniform RNs (e.g. Box-Muller for Normal distribution)

State-of-the-art is the Mersenne Twister (period ~ $2^{199636} - 1$) or for quicker computation use L'Ecuyer's multiple streams: MRG32k3a (2^{191})

For small probabilities be careful of your *machine epsilon*: the smallest number your computer can handle (mine is 2.2×10^{-16} – what's yours?)

Read:

- Park & Miller 1984 – *RNGs: good ones are hard to find*
- *Numerical Recipes in C++*
- the work of L'Ecuyer (and his free source codes!)

TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

Computational Tools - Matlab

Good points:

- quick easy to use and get results
- huge library of optimized functions for specialized tasks
- Good plots and figures

Bad points:

- Is a high level language so uncompiled complex code can run slow
- Not great support for advanced statistics work

Get:

- WAFO toolbox (free) – www.maths.lth.se/matstat/wafo/
- Matlab's own Statistics Toolbox (not free!)

Computational Tools - *R*

R is freeware statistics software (similar to commercial *S-Plus*)

- Get it at www.r-project.org

Good points:

- Huge library of packages (functions) written by statistics researchers
- Even the most obscure problems can be found
  (e.g. generating random multivariate extreme value numbers!)
- Very quick to execute

Bad points:

- language very different to others and so hard to learn
- Less user-friendly – more obviously a research tool

Computational Tools – Programming

For your own algorithms that must run fast use C++ or Fortran

For graphical user interface use .NET framework or similar

> But avoid if at all possible as GUI code can take longer to write than the computation (or productive) code (e.g. use console)

Try to think about the long-term of your code:

- lots of comments
- compile to dll when perfected – for use in other programs?

Don't spend time making it useable for other people (unless necessary)

Computational Tools – Programming

Use Object-Oriented Programming (OOP)

Please…you will thank me later!

Benefits:

- Encapsulation: easier to write, maintain, extend, and reuse
- Inheritance: generalize and then specialize as necessary
- Polymorphism: much more flexibility of your code

Use the C++ *Standard Template Language*

highly-optimized library for efficiency, memory, and time

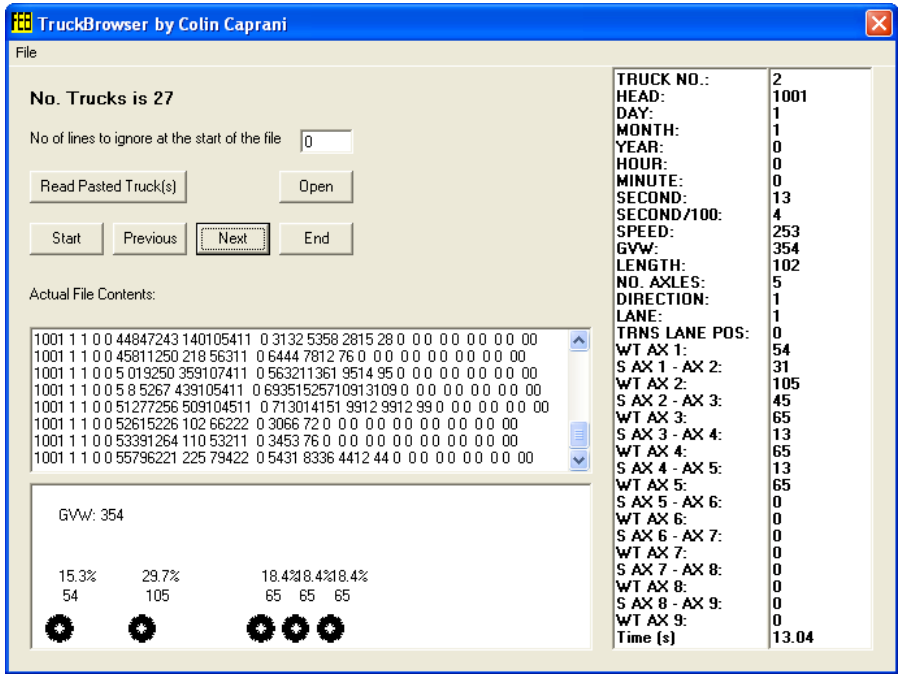
Read Bjarne Stroustrop: *Programming: Principles and Practice using C++*

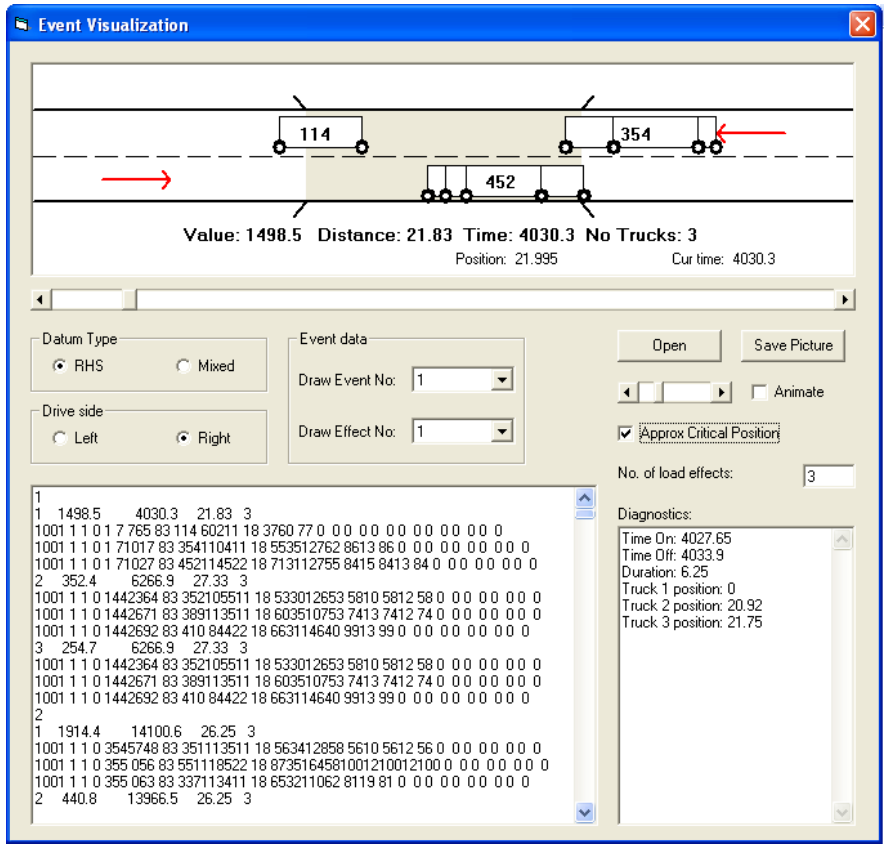
TEAM TRAINING in EUROPEAN ASSET MANAGEMENT

Computational Tools – Diagnostics

Develop lots of small diagnostic programs that are just for you:

- to help visualize phenomenon
- to assist with debugging logic errors





TRAINING in EUROPEAN ASSET MANAGEMENT

Epilogue

Because all good things come to an end…

Epilogue

This presentation and some related files will be available at:

www.colincaprani.com/research/projects/team-project-talk/

Password: *Paris051010*

Thank you for listening and I hope this helps you…

Any questions?