

Using χ^2_{\min} to reject models

Fit M parameters to N data points:

$$\chi^2_{\min} = \sum_{i=1}^N \left[\frac{X_i - \mu_i(\alpha_1, \dots, \alpha_M)}{\sigma_i} \right]^2 \sim \chi^2_{N-M}$$
$$\langle \chi^2_{N-M} \rangle = N - M \quad \sigma^2(\chi^2_{N-M}) = 2(N - M)$$

Why $N - M$ degrees of freedom?

Fitting $M = N$ parameters should fit N points exactly.

If model is good, then the best-fit χ^2_{\min} should be:

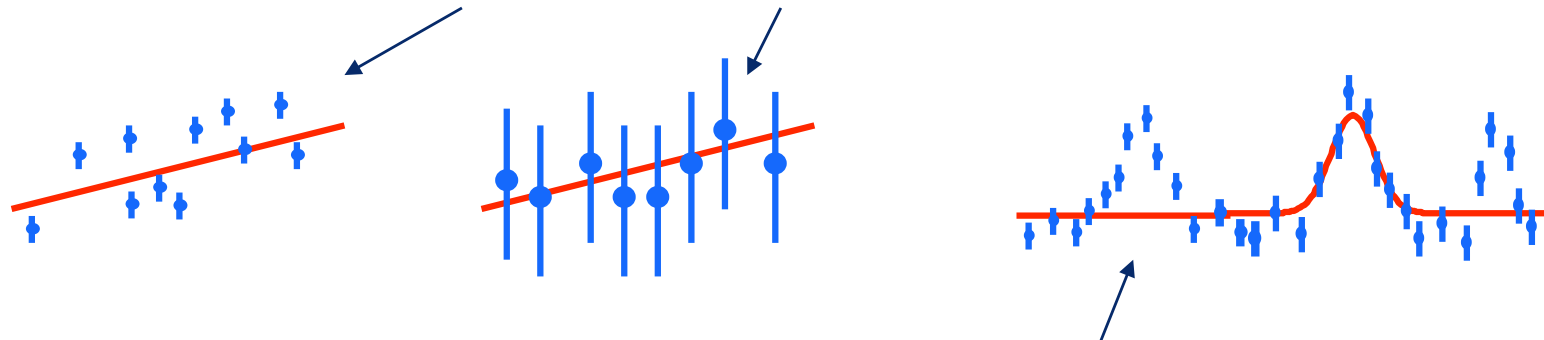
$$\chi^2_{\min} \approx N - M \pm \sqrt{2(N - M)}$$

$$\frac{\chi^2_{\min}}{N - M} \approx 1 \pm \sqrt{\frac{2}{N - M}}$$

What if χ^2_{\min} is too high (or low)?

Several possibilities:

1. Statistical fluke? Use χ^2_{N-M} distribution to estimate probability
2. Wrong model? Use χ^2_{N-M} distribution to reject model
3. Error bars **too small** or **too large**? Re-scale or adjust σ_i ?

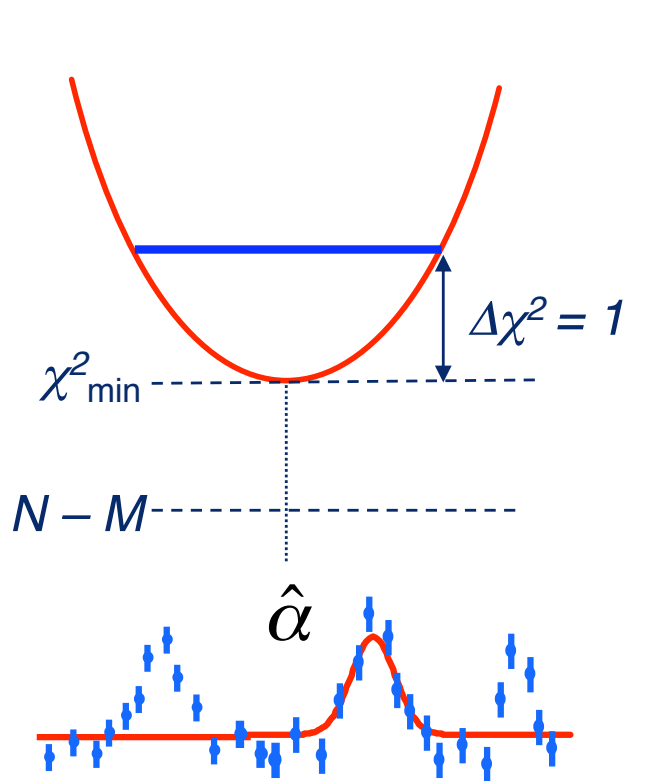


4. Right model, good error bars, but **additional (nuisance) parameters** omitted or not optimised?

Failure to optimise nuisance parameters increases χ^2_{\min} , but may leave the χ^2 curvature the same, **if the nuisance parameters are orthogonal to the parameters of interest.**

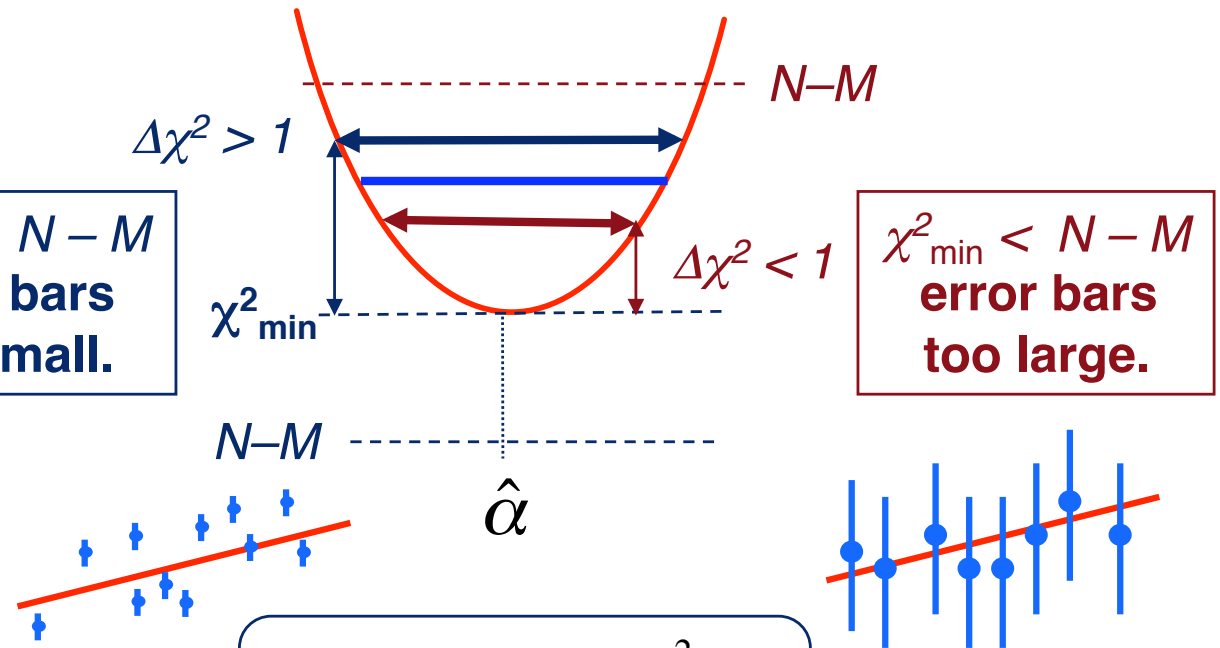
Can then still use $\chi^2_{\min} + 1$ to set $1-\sigma$ confidence intervals on parameters **orthogonal to the nuisance parameters.**

Diagnosis of χ^2_{\min} too large or small



$\chi^2_{\min} > N - M$
error bars too small.

$\chi^2_{\min} > N - M$ due to failure to optimise **orthogonal** nuisance parameters?
 If so, then curvature unchanged.
 Use $\Delta\chi^2 = 1$.



$\chi^2_{\min} < N - M$
error bars too large.

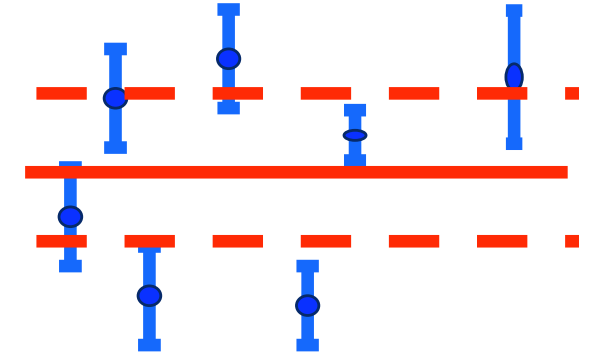
Use:
$$\Delta\chi^2 = \frac{\chi^2_{\min}}{N - M}$$

Equivalent to **re-scaling the error bars**

$$\sigma \Rightarrow \sigma \sqrt{\frac{\chi^2_{\min}}{N - M}}$$

Estimate the “Extra Variance”

Assume two independent noise sources:

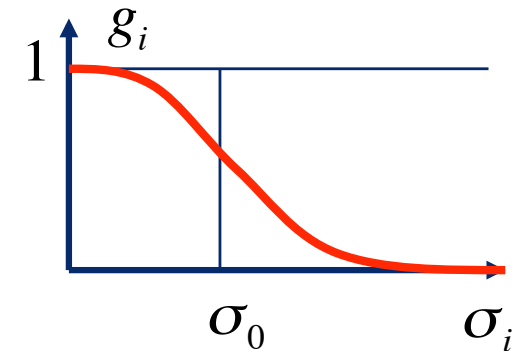


$$\text{Var}[X_i] = \sigma_0^2 + \sigma_i^2 = \frac{\sigma_0^2}{g_i} \quad g_i \equiv \frac{\sigma_0^2}{\sigma_0^2 + \sigma_i^2} = \frac{1}{1 + (\sigma_i/\sigma_0)^2}$$

$$-2 \ln L = \sum_{i=1}^N \frac{(X_i - \mu)^2}{\sigma_0^2 + \sigma_i^2} + \sum_{i=1}^N \ln(\sigma_0^2 + \sigma_i^2) = \sum_{i=1}^N \left(\frac{X_i - \mu}{\sigma_0} \right)^2 g_i + \sum_{i=1}^N \ln(\sigma_0^2 / g_i)$$

$$0 = \frac{\partial(-2 \ln L)}{\partial \mu} = -2 \sum_{i=1}^N \frac{(X_i - \mu)}{\sigma_0^2 + \sigma_i^2} = -2 \sum_{i=1}^N \frac{(X_i - \mu) g_i}{\sigma_0^2}$$

$$0 = \frac{\partial(-2 \ln L)}{\partial \sigma_0^2} = - \sum_{i=1}^N \frac{(X_i - \mu)^2 g_i^2}{\sigma_0^4} + \sum_{i=1}^N \frac{g_i}{\sigma_0^2}$$



$$\hat{\mu} = \frac{\sum \frac{X_i}{\sigma_0^2 + \sigma_i^2}}{\sum \frac{1}{\sigma_0^2 + \sigma_i^2}} = \frac{\sum X_i g_i}{\sum g_i} \quad \text{Var}[\hat{\mu}] = \frac{\sigma_0^2}{\sum g_i} \quad \hat{\sigma}_0^2 = \frac{\sum (X_i - \mu)^2 g_i^2}{\sum g_i}$$

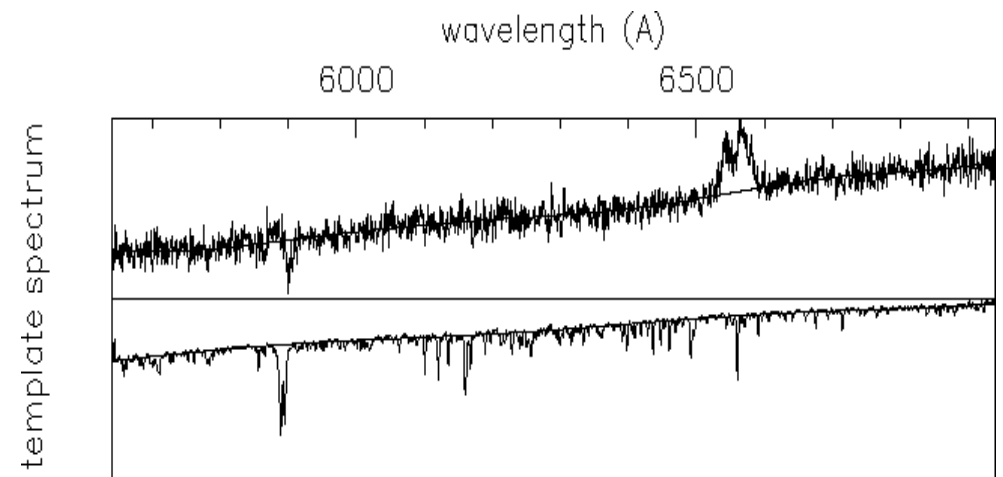
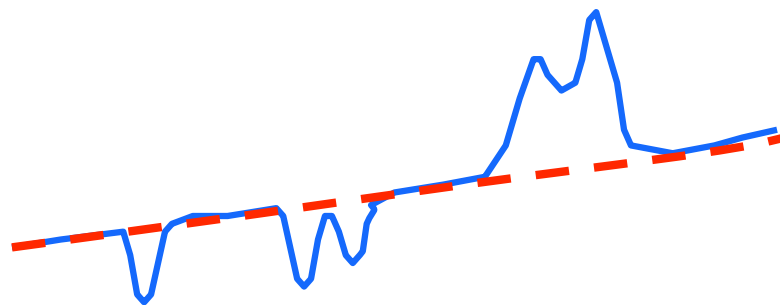
Need to iterate.

Background Functions

Smooth functions with adjustable flexibility.

- Polynomials
- Splines
- Running Optimal Average
 - with sigma-clipping
- Running Median

Example: Continuum fit to a spectrum



Polynomials

Fit $N = 30$ points with
 $M = 1, 2, 3, 4$ polynomial
coefficients.

Higher $M =$ more flexible model.
Use lowest M that gives good fit.

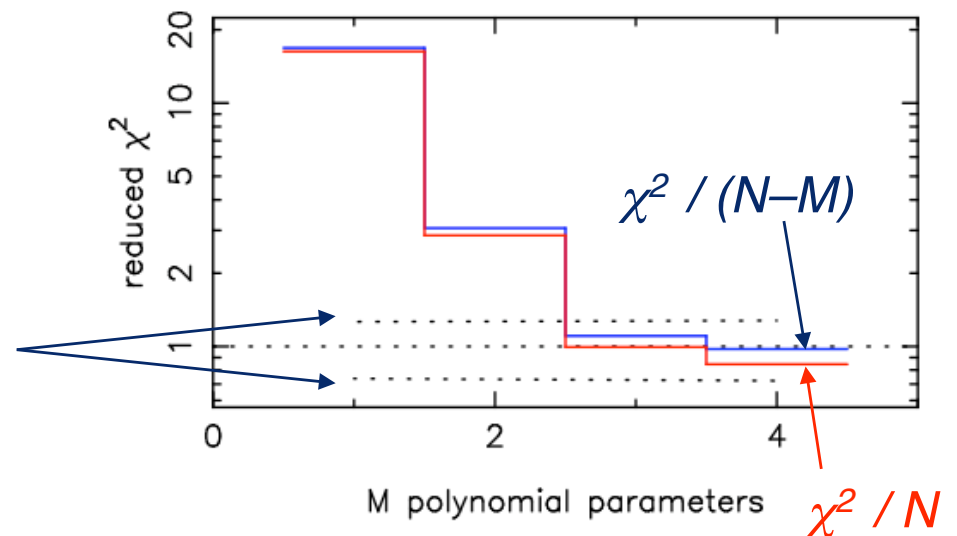
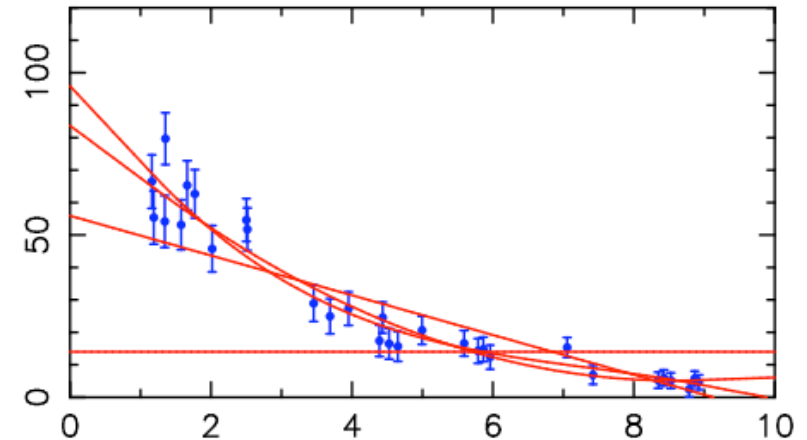
Reject $M = 1, 2.$

Accept $M = 3, 4.$

Based on Reduced χ^2

$$\frac{\chi^2}{N - M} \approx 1 \pm \sqrt{\frac{2}{N - M}}$$

Polynomial Fit $N = 30$ $M = 1 \dots 4$



Splines – e.g. piecewise cubic

N nodes: $x_i, y_i \quad i = 1, \dots, N.$
 x_i fixed, y_i adjustable.

$4(N - 1)$ parameters (4 cubic coefficients for each of the $N - 1$ segments)

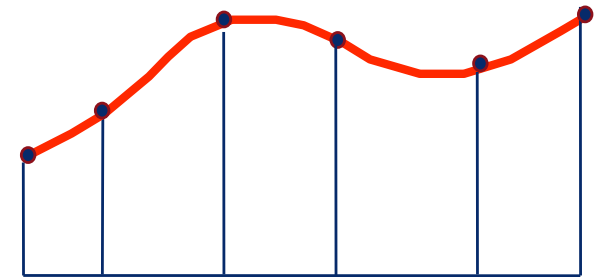
$3(N - 2)$ matching conditions (value, slope, curvature at each of the $N - 2$ internal nodes)

$N + 2$ degrees of freedom (N values y_i values plus either slope or curvature at 2 end points).

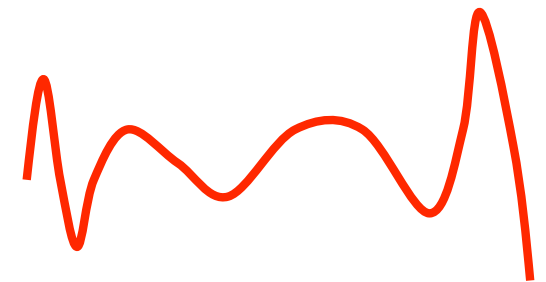
- First, **distribute the nodes x_i** , e.g. equally spaced, or equal weight $\Sigma(1/\sigma^2)$ on each segment.
- Then, **fit the $N + 2$ parameters**, e.g. optimise y_i by χ^2 minimization, set endpoint curvatures (or slopes) to zero.

Low-order polys good for simple background fits.

Splines better than high-order polys. Better control over the x distribution of the degrees of freedom.



8-parameter cubic spline



8-parameter polynomial

Running Optimal Average

$$\hat{X}(t) = \frac{\sum X_i w_i(t)}{\sum w_i(t)} \quad \sigma^2(\hat{X}(t)) = \frac{1}{\sum w_i(t)}$$

$$w_i(t) = \frac{G(t - t_i)}{\sigma_i^2}$$

Memory function $G(t)$

expands the error bars as time-difference increases.

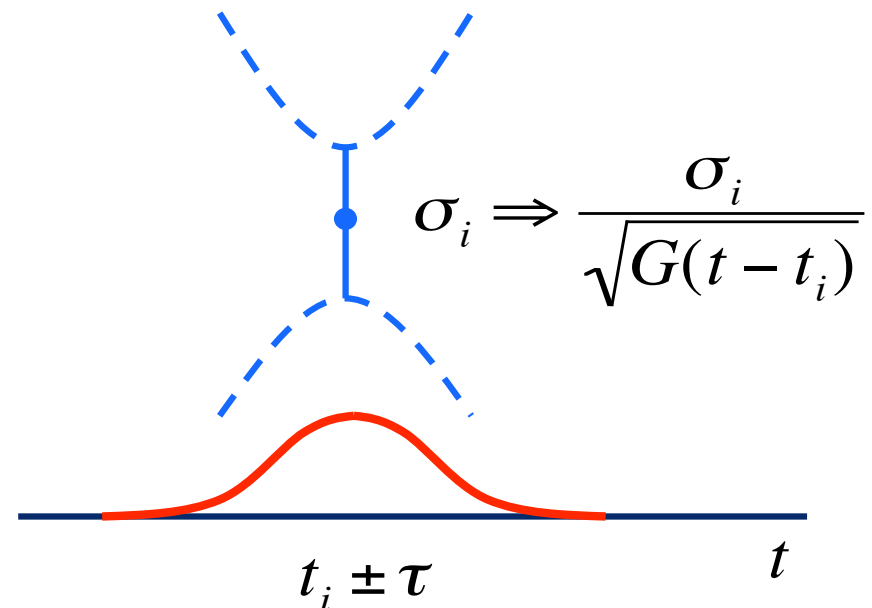
Parameter τ controls time interval over which the data point retains its $1/\sigma^2$ weight.

Memory functions:

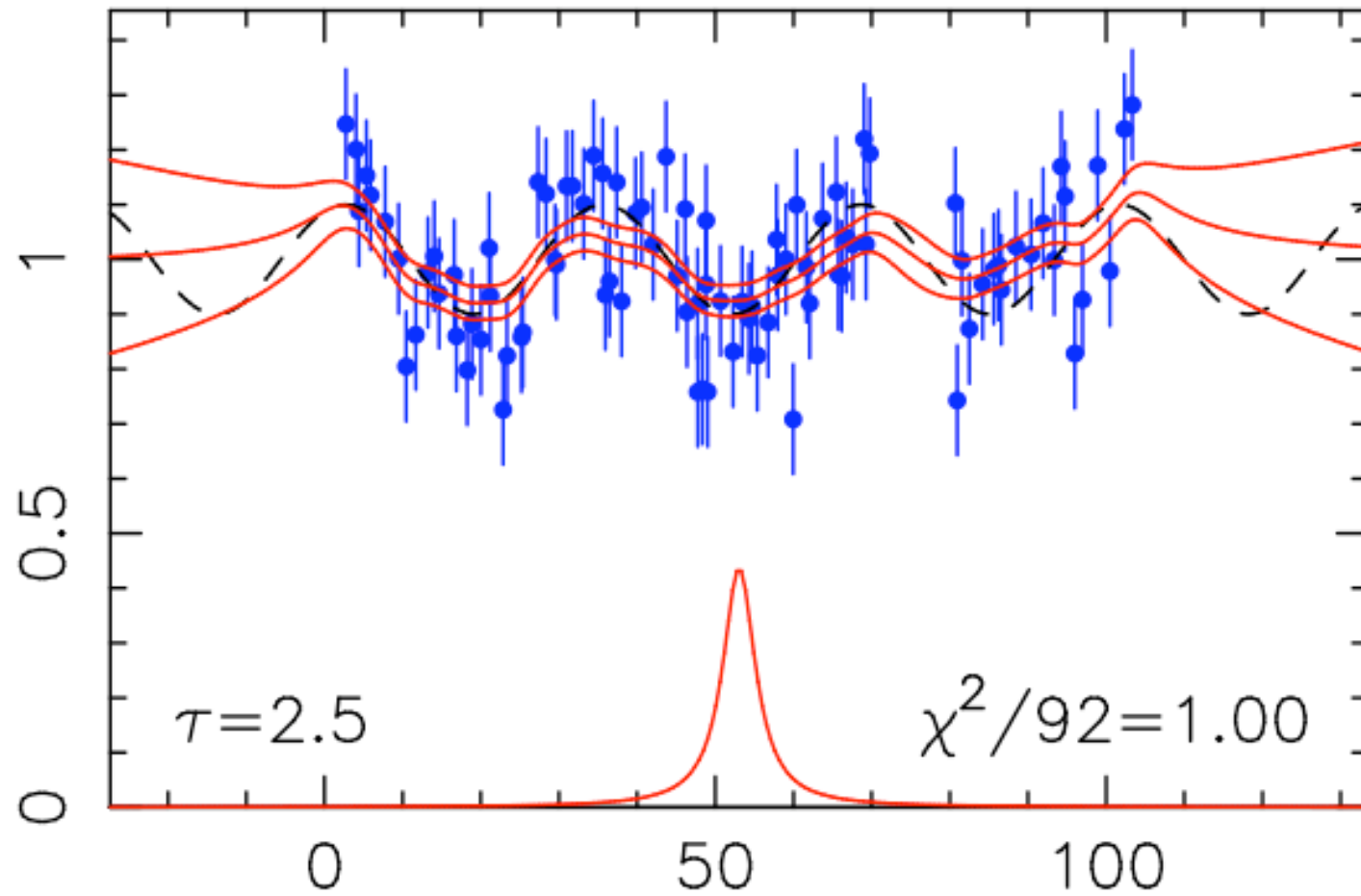
boxcar: $G(t) = \begin{cases} 1 & |t| < \tau \\ 0 & |t| > \tau \end{cases}$

Gaussian: $= \exp\left\{-\frac{1}{2}\left(\frac{t}{\tau}\right)^2\right\}$

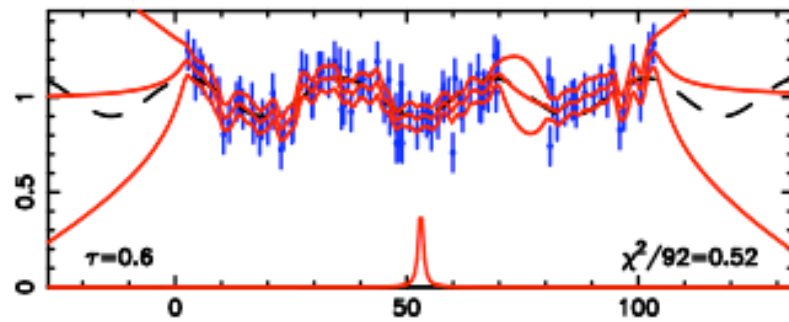
Lorentzian: $= \frac{1}{1 + (t/\tau)^2}$



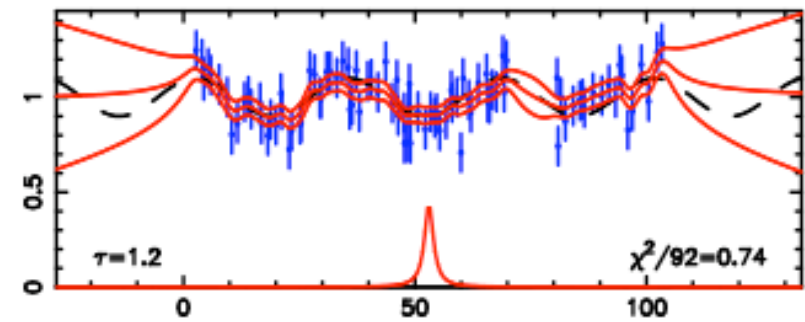
Running Optimal Average



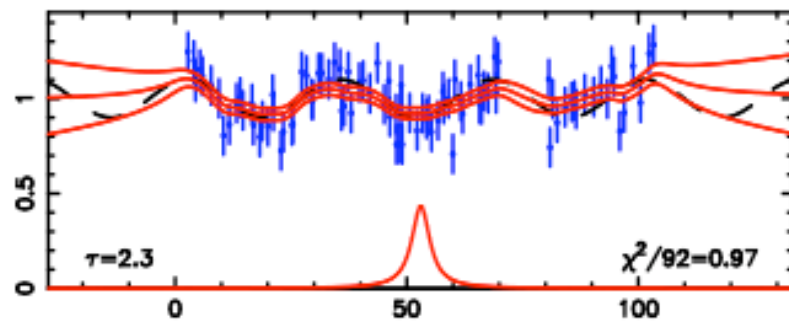
running optimal average



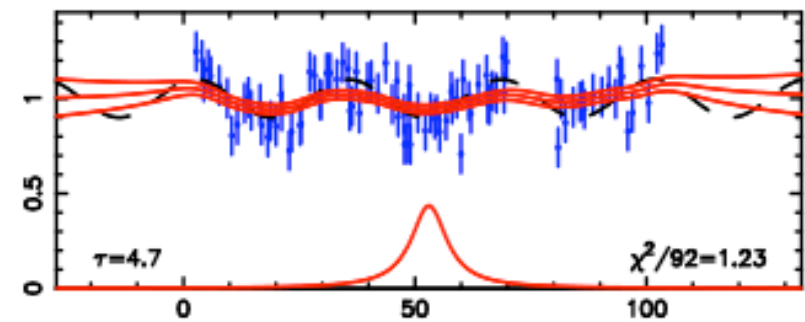
running optimal average



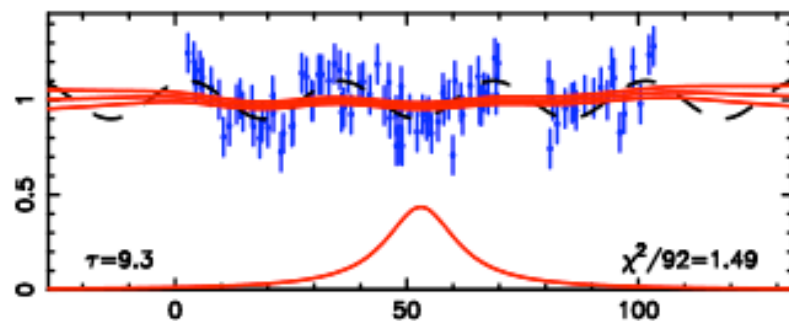
running optimal average



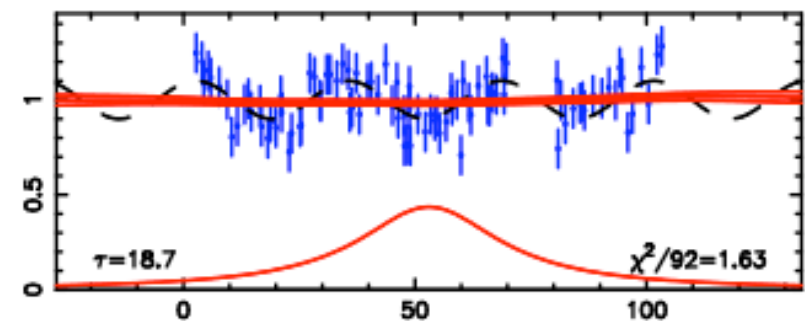
running optimal average



running optimal average

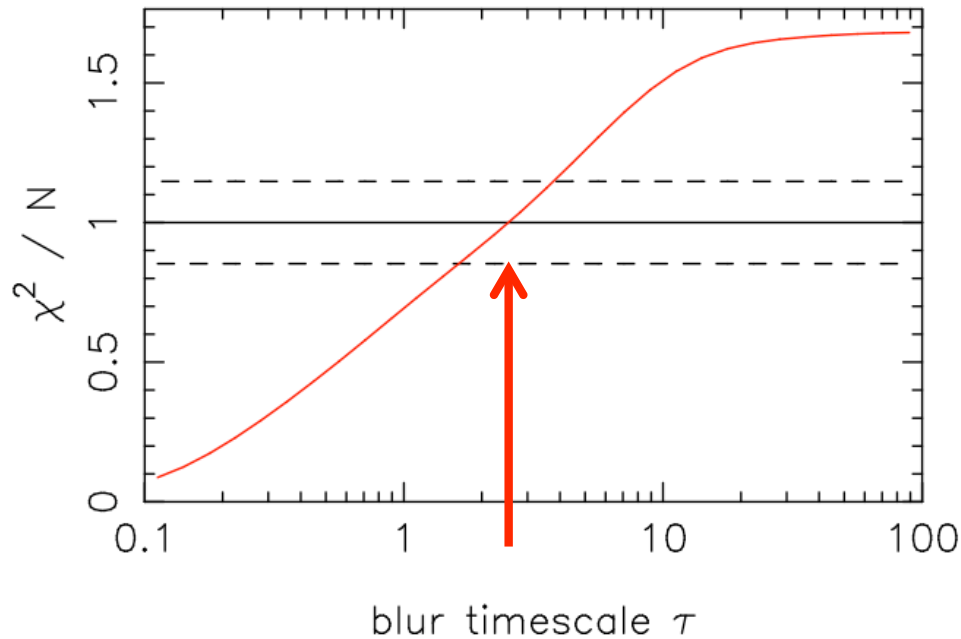


running optimal average



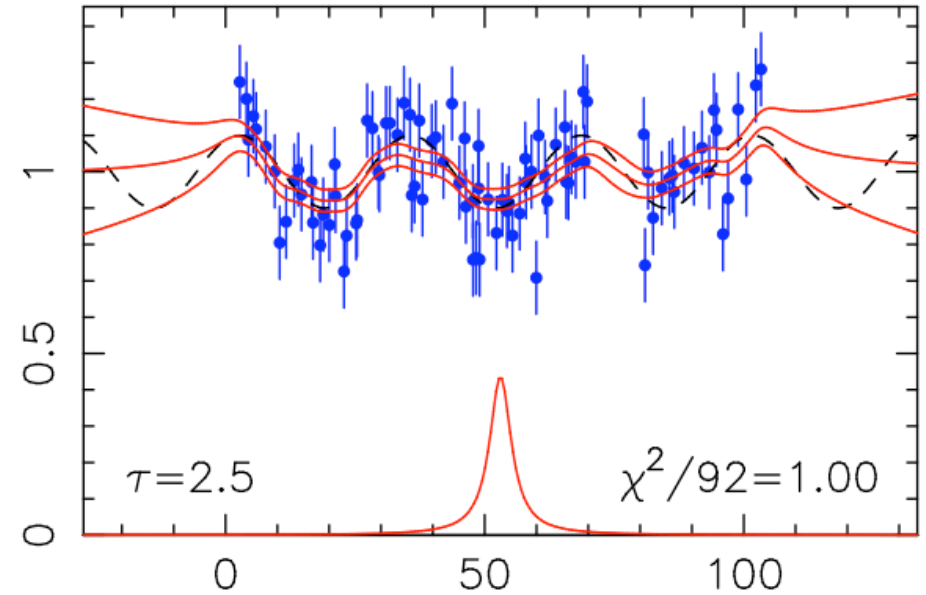
Running Optimal Average

running optimal average $N = 92$



Blur timescale τ
chosen to make $\chi^2 / N = 1$.

running optimal average

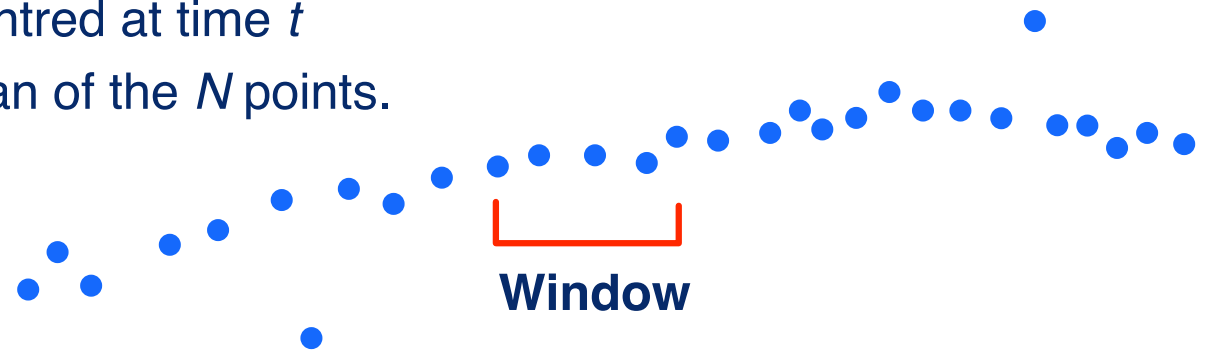


Interpolates across gaps.
Extrapolates past ends.
Averages appropriately.
Error bars provided.
(Almost) model-free.

Median Filter and Sigma-Clip

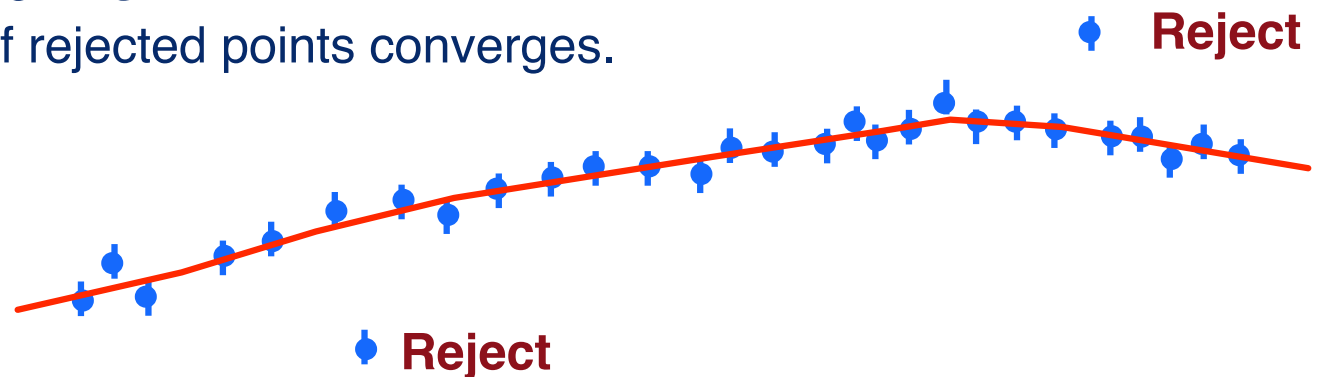
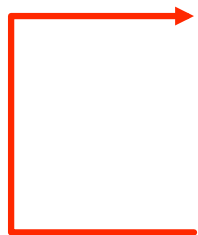
Median filter:

- window of N points centred at time t
- $\text{medfilt}(t)$ is the median of the N points.



Sigma-clip:

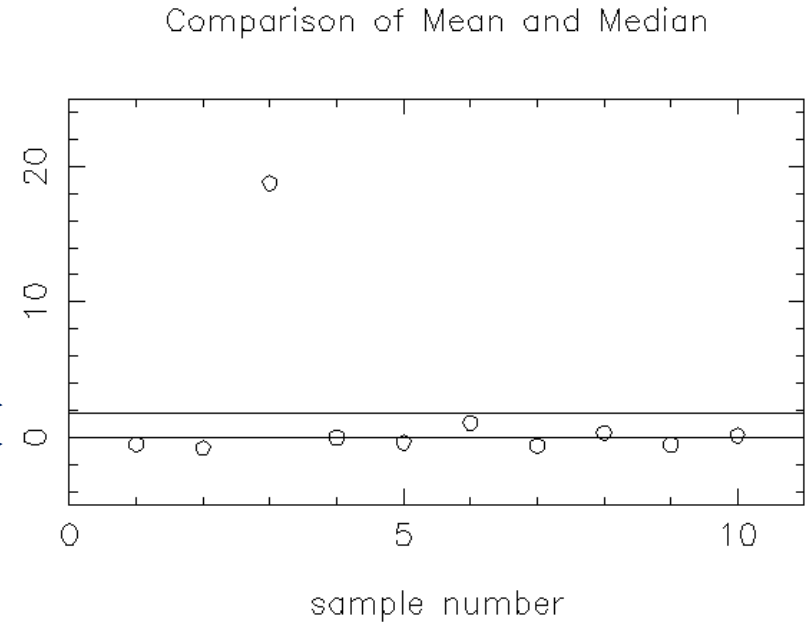
- Fit all points by minimising χ^2
- Set threshold K and check for outliers at $\pm K \sigma$ or more
- Repeat fit omitting **largest** outlier
- Iterate until set of rejected points converges.



Mean vs Median

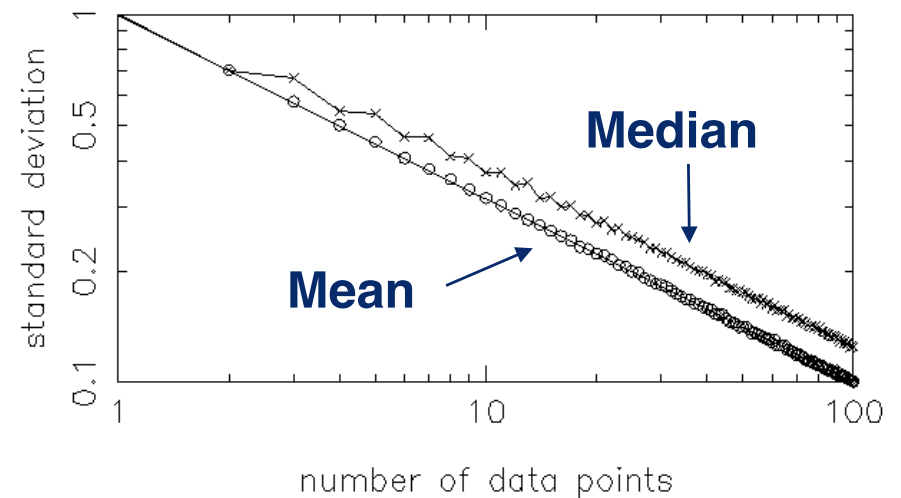
- The median is **less sensitive to outliers** than the mean.

Mean →
Median →



- The median is **unbiased** but **not a minimum-variance estimator**.
- Note how **standard deviation of the median** varies with sample size N in comparison to **standard deviation of the mean**.

Variance of the Median is larger by a factor $\pi / 2 = 1.57$ (for large N) than the Variance of the Mean.



Variance of Median vs Mean

N gaussian random numbers:

$$\langle X_i \rangle = 0 \quad \text{Var}[X_i] = \sigma^2 \quad i = 1 \dots N$$

$$f(x) = dF/dx = \exp\{-x^2/2\} / (2\pi\sigma^2)^{1/2}$$

P = fraction of positive values:

$$p_i = \begin{cases} 1 & X_i > 0 \\ 0 & X_i < 0 \end{cases} \quad \langle p_i \rangle = \frac{1}{2} \quad \sigma^2(p_i) = \frac{1}{4}$$

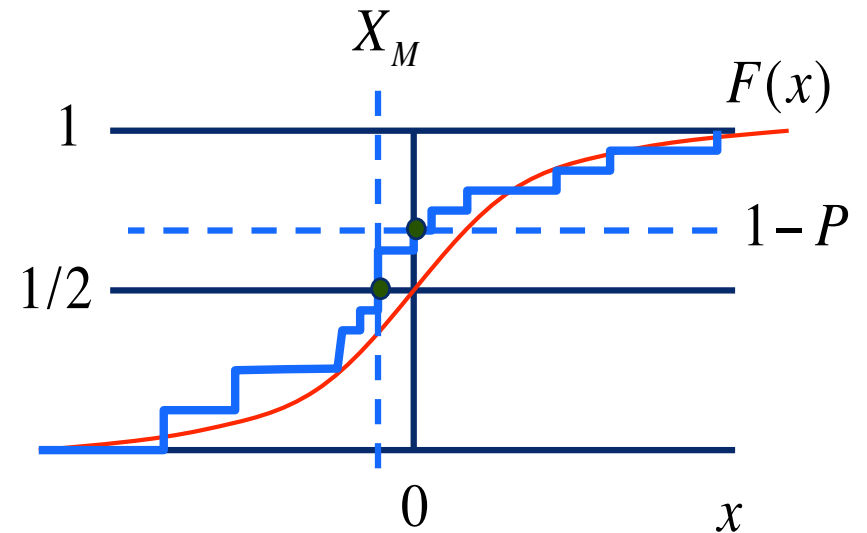
$$P \equiv \frac{1}{N} \sum_{i=1}^N p_i \quad \langle P \rangle = \frac{1}{2} \quad \sigma^2(P) = \frac{1}{4N}$$

$$\text{Median: } X_M \approx \frac{P - \langle P \rangle}{dF/dx|_{x=0}} = \left(P - \frac{1}{2}\right) (2\pi\sigma^2)^{1/2}$$

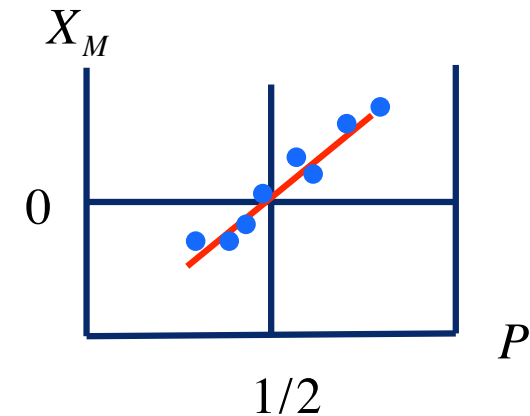
$$\frac{\partial X_M}{\partial P} = \frac{1}{dF/dx|_{x=0}} = (2\pi\sigma^2)^{1/2}$$

$$\sigma^2(X_M) = \sigma^2(P) \left| \frac{\partial X_M}{\partial P} \right|^2 = \frac{1}{4N} (2\pi\sigma^2) = \frac{\pi\sigma^2}{2N}$$

$$\sigma^2(\bar{X}) = \frac{\sigma^2}{N}$$



P co-varies with X_M :



Variance of the Median is larger by a factor $\pi/2 = 1.57$ (for large N) than the Variance of the Mean.