ADA06 - 10am Mon 26 Sep 2022

Recap : χ^2 fitting

Sample Variance S^2
bias correction

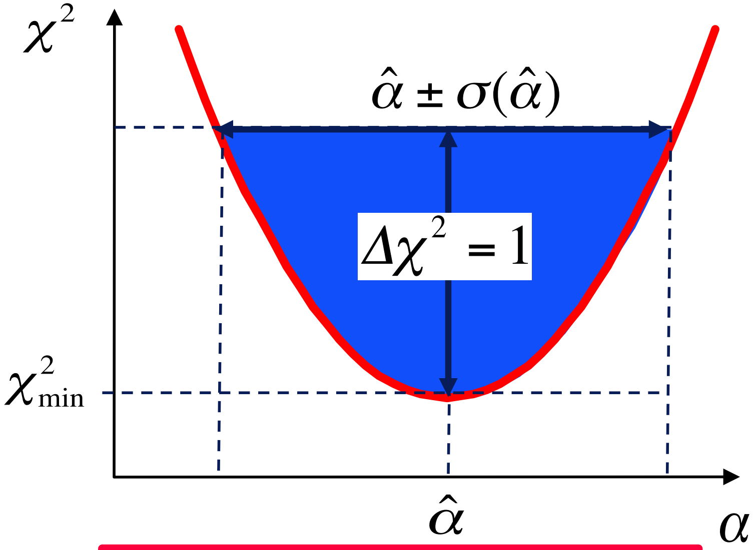
“Robust” Statistics:
Median and MAD
sigma clipping

Recap : Fitting models by minimizing χ^2

Fit M parameters to N data points.

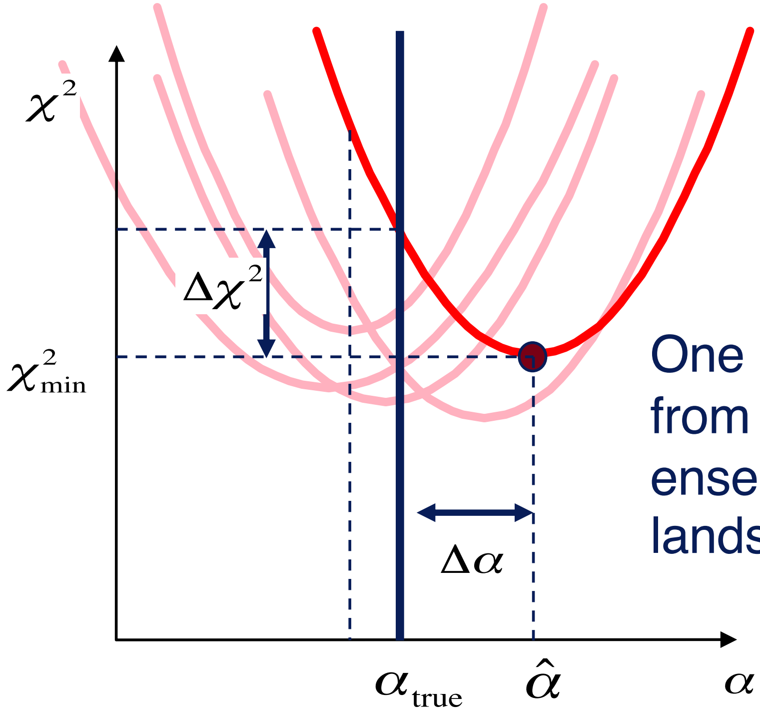
$$\chi^2(X,\sigma,\alpha) \equiv \sum_{i=1}^{N} \left(\frac{X_i - \mu_i(\alpha)}{\sigma_i} \right)^2$$

Best - fit parameters $\hat{\alpha}$ minimise χ^2.



$$\boxed{\sigma^2(\hat{\alpha}) = \frac{2}{\left.\frac{\partial^2 \chi^2}{\partial \alpha^2}\right|_{\alpha=\hat{\alpha}}}}$$

Caveat: Assumes orthogonal parameters.
Generalise to correlated parameters later.



One sample from an ensemble of χ^2 landscapes.

$$\hat{\alpha} \sim G\left(\alpha_{\text{true}}, \sigma^2(\hat{\alpha})\right)$$

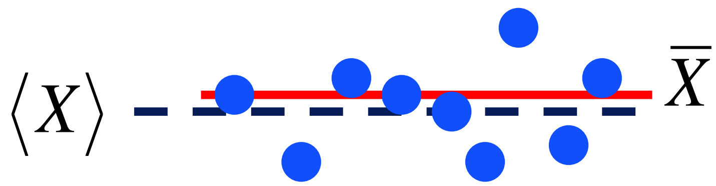
$$\chi^2(\alpha_{true}) \sim \chi^2_N$$

$$\chi^2_{\min} \equiv \chi^2(\hat{\alpha}) \sim \chi^2_{N-M}$$

$$\Delta\chi^2 \equiv \chi^2(\alpha_{\text{true}}) - \chi^2_{\min} \sim \chi^2_M$$

$$\langle \chi^2_N \rangle = N \ , \quad \mathrm{Var}\left[\chi^2_N\right] = 2\,N$$

Data points with no error bars ☹



N data points: $\quad \langle X_i \rangle = \langle X \rangle \quad \mathrm{Cov}\left(X_i \, X_j \right) = \sigma^2 \, \delta_{ij}$

Sample mean: $\bar{X} \equiv \dfrac{1}{N} \sum_i X_i \qquad$ unbiased: $\langle \bar{X} \rangle = \langle X \rangle . \qquad \mathrm{Var}\left(\bar{X} \right) = \dfrac{\sigma^2}{N}$

But σ_{i}^2 are unknown. $\quad$ How can we estimate σ^2 ?

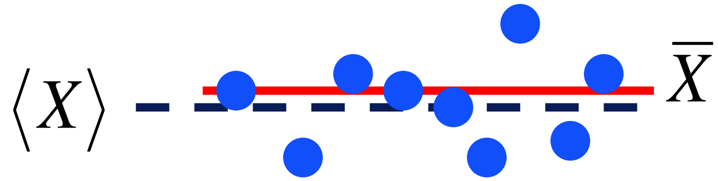
Variance: $\quad \sigma^2 \left(X \right) \equiv \left\langle \left(X - \langle X \rangle \right)^2 \right\rangle$

Try: $\qquad s^2 \equiv \dfrac{1}{N} \sum_i \left(X_i - \bar{X} \right)^2$

Is $\qquad \left\langle s^2 \right\rangle = \sigma^2 \quad ?$

No. $\qquad \left\langle s^2 \right\rangle < \sigma^2 \quad$ We can evaluate and then correct for this bias.

Sample Variance S^2 : Unbiased for σ^2



$$S^2 \equiv A \sum_{i=1}^{N} (X_i - \bar{X})^2 \qquad \text{Pick } A \text{ so that } \left\langle S^2 \right\rangle = A \sum_{i=1}^{N} \left\langle (X_i - \bar{X})^2 \right\rangle = \sigma^2$$

$$\left\langle (X_i - \bar{X})^2 \right\rangle = \left\langle \left[(X_i - \langle X \rangle) - (\bar{X} - \langle X \rangle) \right]^2 \right\rangle$$

$$= \left\langle (X_i - \langle X \rangle)^2 + (\bar{X} - \langle X \rangle)^2 - 2\,(X_i - \langle X \rangle)\,(\bar{X} - \langle X \rangle) \right\rangle$$

$$= \sigma^2(X_i) + \sigma^2(\bar{X}) - 2\,\mathrm{Cov}(X_i, \bar{X})$$

$$= \sigma^2 \qquad + \frac{\sigma^2}{N} \qquad - 2\,\frac{\sigma^2}{N}$$

$$\text{Note : } \quad \mathrm{Cov}(X_i, \bar{X}) = \frac{\sigma^2}{N}$$

$$= \left(1 - \frac{1}{N} \right) \sigma^2 = \left(\frac{N-1}{N} \right) \sigma^2$$

$$\therefore \quad \left\langle S^2 \right\rangle = A \sum_{i=1}^{N} \left(\frac{N-1}{N} \right) \sigma^2 \quad \text{Pick} \quad A = \frac{1}{N-1}$$

$$\boxed{S^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2}$$

Evaluation of Cov(X_i, $\bar{X}$)

$$\mathrm{Cov}(X_i,\bar{X}) \equiv \left\langle (X_i - \langle X_i \rangle)\,(\bar{X} - \langle \bar{X} \rangle) \right\rangle$$

$$\text{Note:}\ \ \langle X_i \rangle = \langle \bar{X} \rangle = \langle X \rangle$$

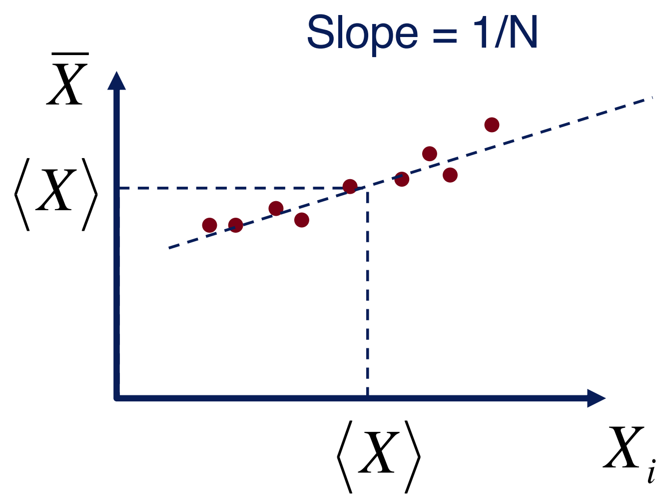
$$\text{Shift coords to put } \langle X \rangle = 0 \ \ :$$

$$\mathrm{Cov}(X_i,\bar{X}) = \left\langle (X_i - 0)\,(\bar{X} - 0) \right\rangle$$

$$= \left\langle X_i \frac{1}{N} \sum_k X_k \right\rangle$$

$$= \frac{1}{N} \sum_k \langle\ X_i\ X_k\ \rangle$$

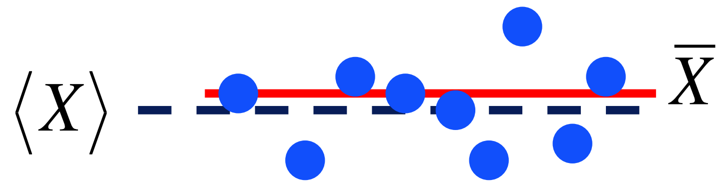
$$= \frac{1}{N} \sum_k \sigma^2 \delta_{ik} = \frac{\sigma^2}{N}$$



$$\mathrm{Cov}(X_i,X_j) \equiv \sigma^2\ \delta_{i\,j}$$

Sample Variance S^2 : Unbiased for σ^2

$$S^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})^2$$



Why $\dfrac{1}{N\text{-}1}$, not $\dfrac{1}{N}$?

Because $\overline{X}$ "chases" the dancing data points,

removing 1 "degree-of-freedom" from the dance.

$$S^2 \sim \frac{\sigma^2}{N-1} \chi^2_{N-1}$$

$$\left\langle S^2 \right\rangle = \frac{\sigma^2}{N-1} \left\langle \chi^2_{N-1} \right\rangle$$

$$= \frac{\sigma^2}{N-1} (N-1) = \sigma^2$$

$$\mathrm{Var}\left[S^2\right] = \left(\frac{\sigma^2}{N-1}\right)^2 \mathrm{Var}\left[\chi^2_{N-1}\right]$$

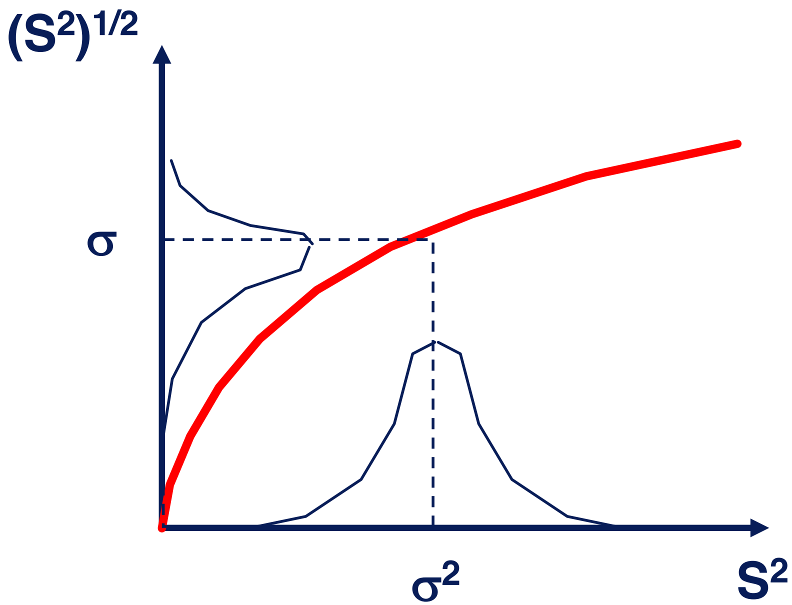
$$= \left(\frac{\sigma^2}{N-1}\right)^2 2(N-1) \quad = \frac{2\,\sigma^4}{N-1}$$

$$\frac{\sigma\left(S^2\right)}{\left\langle S^2 \right\rangle} = \left(\frac{2}{N-1}\right)^{1/2} = \text{fractional accuracy}$$

Is $(S^2)^{1/2}$ unbiased for σ ?

- The sample variance S^2 is unbiased for σ^2.
- i.e. $\langle S^2 \rangle = \sigma^2$
- Is $(S^2)^{1/2}$ unbiased for σ ?
- No. The square root introduces a bias:

$$S^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$



Homework:
Work out the bias correction, as a function of N, to construct an unbiased estimate for σ^p.

$$\left\langle \sqrt{S^2} \right\rangle < \sigma, \text{ even though } \sqrt{\sigma^2} = \sigma.$$

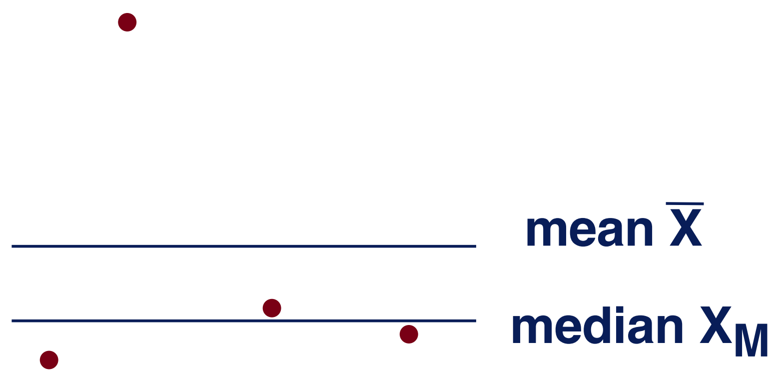
“Robust” estimation methods

Robust Statistics :

less sensitive to “bad” data

Example: use the **median**

rather than the **mean**.



Sample Mean $\bar{X}$ minimizes the **Sample Variance**:

$$S^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \mu \right)^2$$

$$\frac{\partial}{\partial \mu} \left[\sum_{i=1}^{N} (X_i - \mu)^2 \right] = 0$$

$$\text{for} \quad \mu = \bar{X}$$

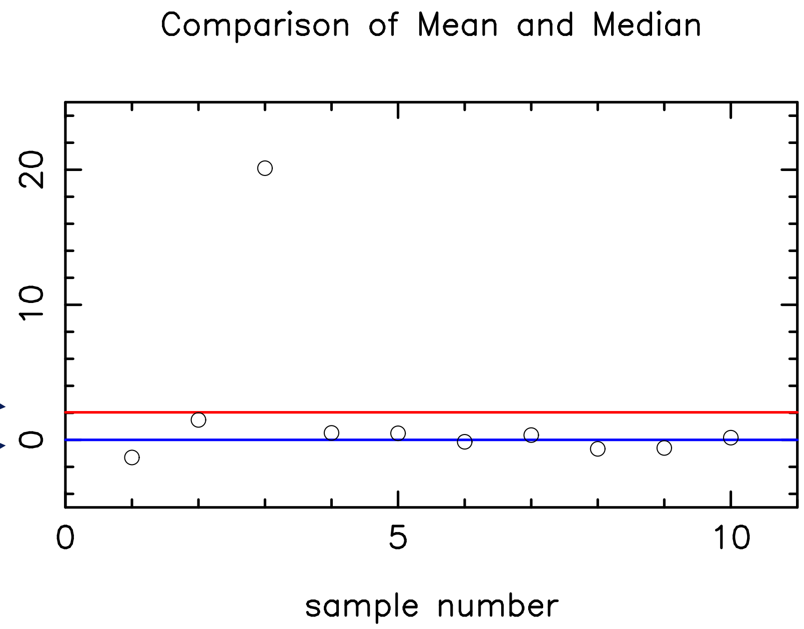
Median X_M minimizes the “**Mean Absolute Deviation**” :

$$\mathrm{MAD} \equiv \frac{1}{N} \sum_{i=1}^{N} \left| X_i - \mu \right|$$

$$\frac{\partial}{\partial \mu} \left[\sum_{i=1}^{N} \left| X_i - \mu \right| \right] = 0$$

$$\text{for} \quad \mu = X_M \equiv \mathrm{Median}\left(X_i \right)$$

Mean vs Median

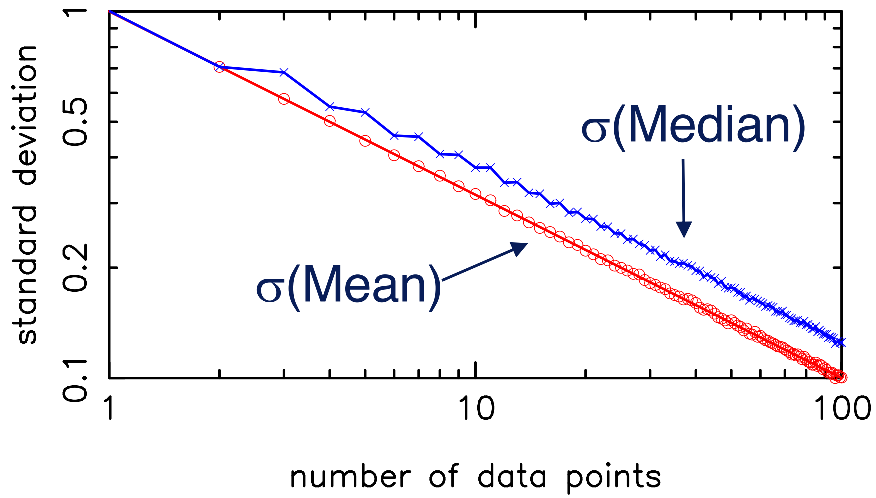


- The median is less sensitive to outliers than the mean.

- The median is **unbiased**, but **not** a minimum-variance estimator.

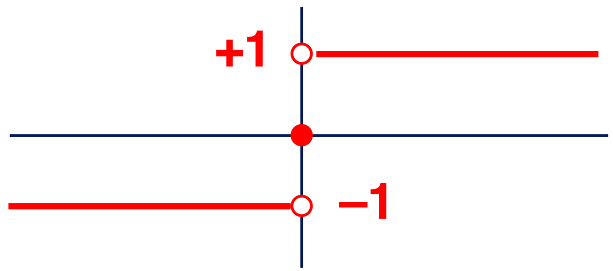
- Note how the standard deviations of the median and of the mean vary with sample size.

$$\sigma\left(\bar{X}\right) = \frac{\sigma}{\sqrt{N}} \quad \le \quad \sigma\left(X_{\text{Med}}\right)$$



"Proof" that the Median minimises the MAD

$$H(x) \equiv \begin{cases} +1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \qquad \frac{\mathrm{d}H}{\mathrm{d}x} = 2\delta(x)$$



$$MAD \equiv \frac{1}{N}\sum_{i=1}^{N} \left| X_i - \mu \right| \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \left(X_i - \mu \right) H\left(X_i - \mu \right)$$

$$\frac{d\,MAD}{d\,\mu} = \frac{1}{N}\sum_{i=1}^{N} \left[(-1)\,H\left(X_i - \mu \right) + \cancel{\left(X_i - \mu \right)(-1)\,H'\left(X_i - \mu \right)} \right]$$

$= 0$

since $H'(x) = 0$

whenever $x \neq 0$

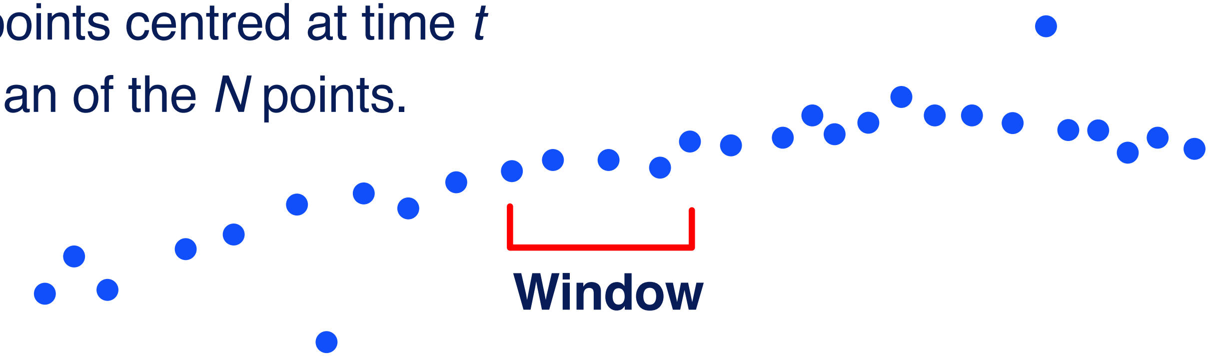
$$= \frac{-1}{N}\sum_{i=1}^{N} H\left(X_i - \mu \right) = \frac{-1}{N}\left(\sum_{X_i > \mu} (+1) \; + \sum_{X_i < \mu} (-1) \right)$$

$$= 0 \quad \text{if} \quad \mu = \text{median}(X_i)$$

Median Filter and Sigma-Clip

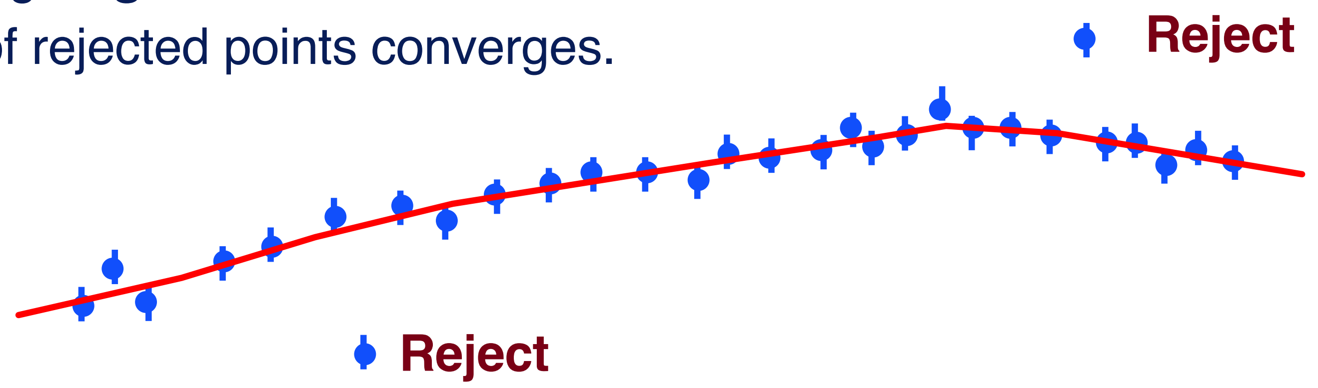
Median filter:

- Window encloses N points centred at time t
- 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.



Various "Badness-of-Fit" Statistics

Sample Variance — mean

$$S^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\mu_i\right)^2 \quad\rightarrow\quad \bar{X}$$

Chi-squared — optimal average

$$\chi^2 \equiv \sum_{i=1}^{N}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2 \quad\rightarrow\quad \hat{X}$$

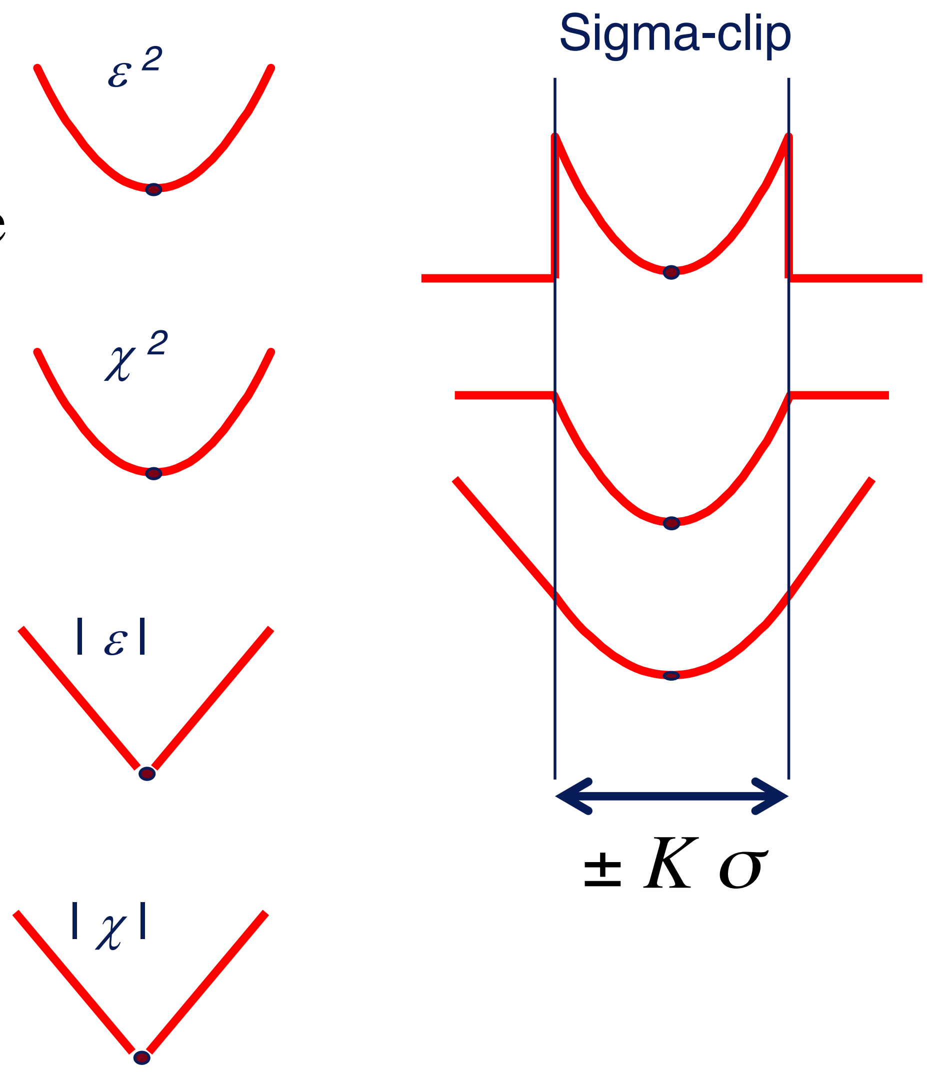
Mean Absolute Deviation — median

$$MAD \equiv \frac{1}{N}\sum_{i=1}^{N}\left|X_i-\mu_i\right| \quad\rightarrow\quad X_M$$

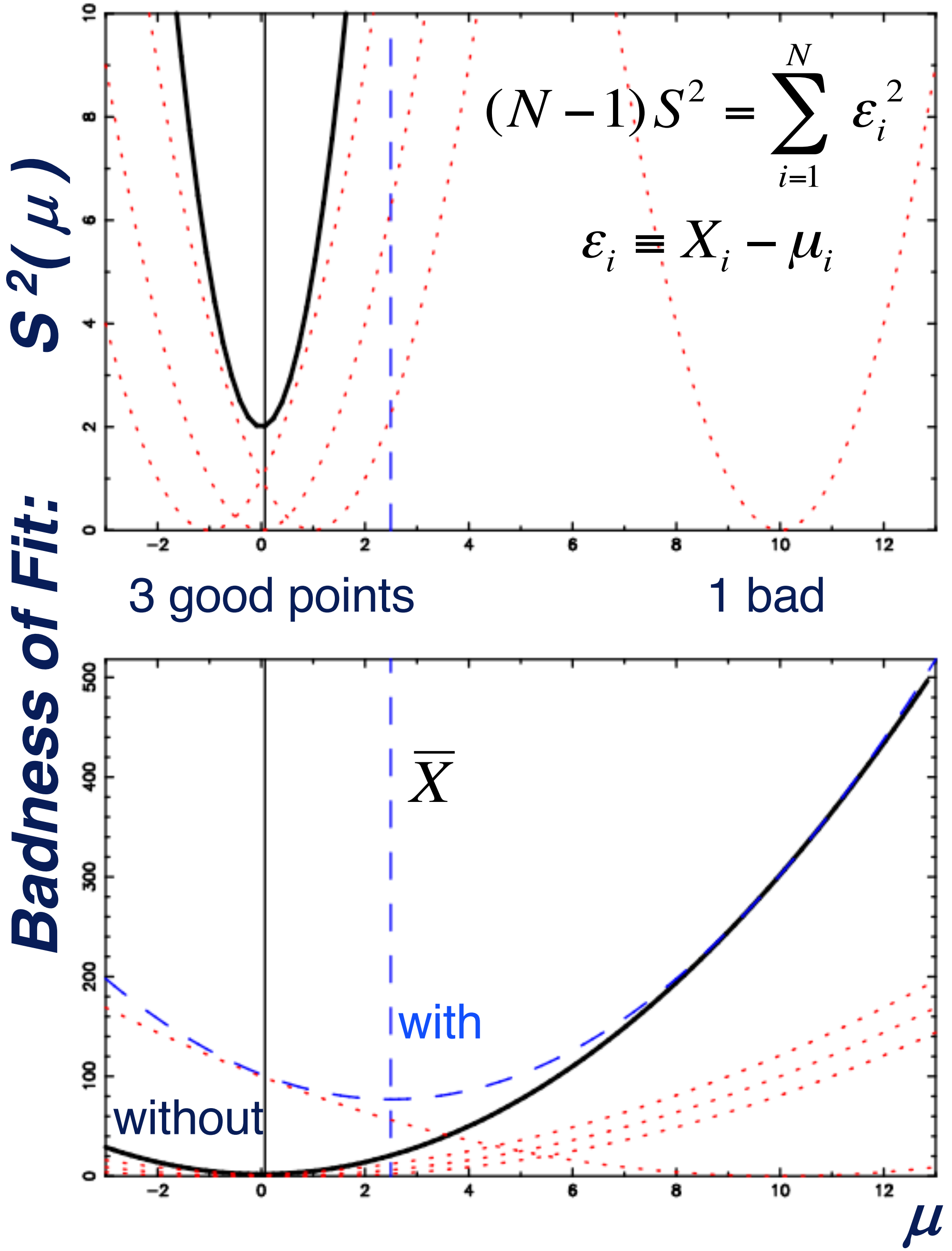
Sum Absolute Normalised Errors:

$$SANE \equiv \sum_{i=1}^{N}\left|\frac{X_i-\mu_i}{\sigma_i}\right|$$

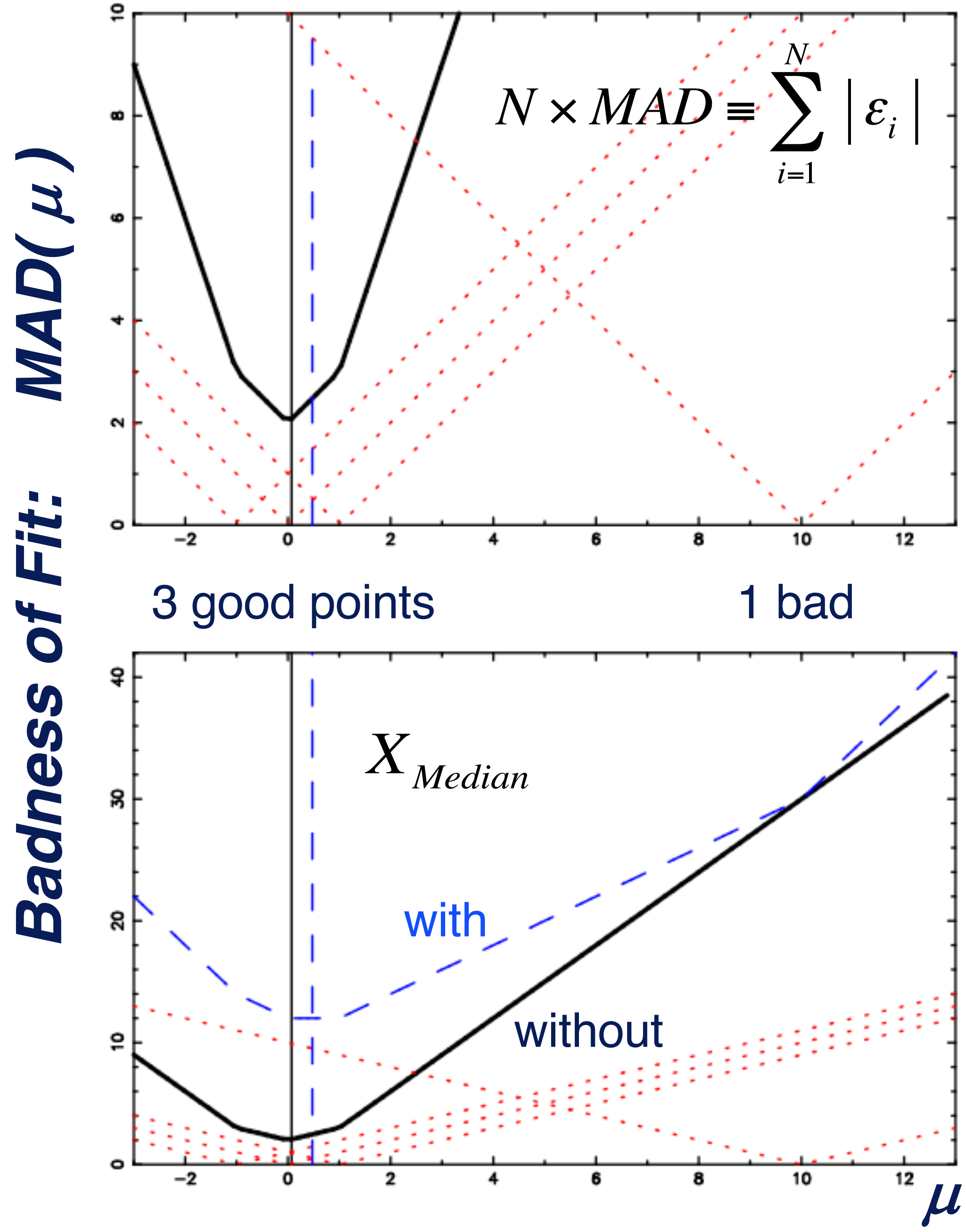
Badness functions:

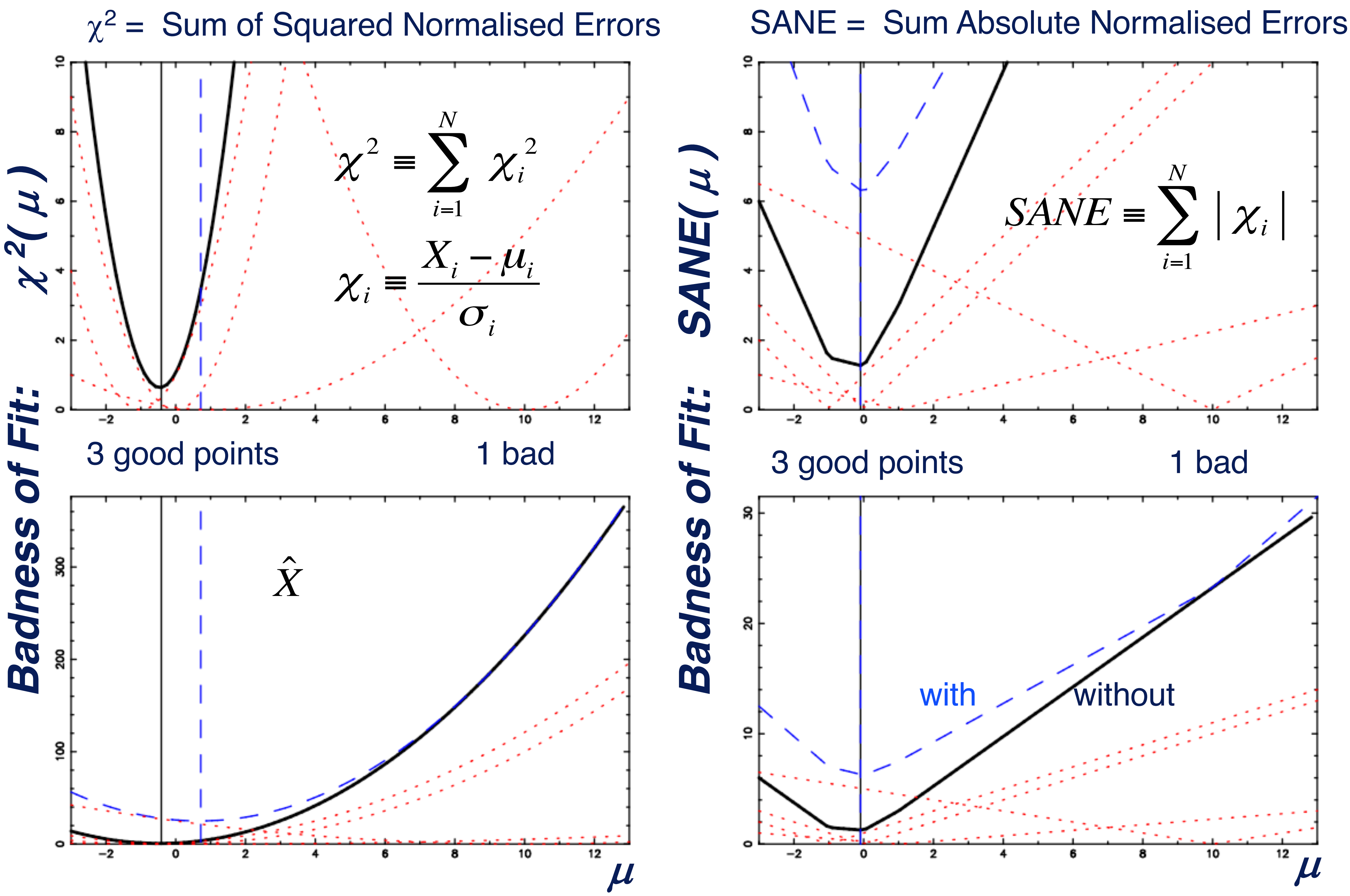


S^2 = Sample Variance



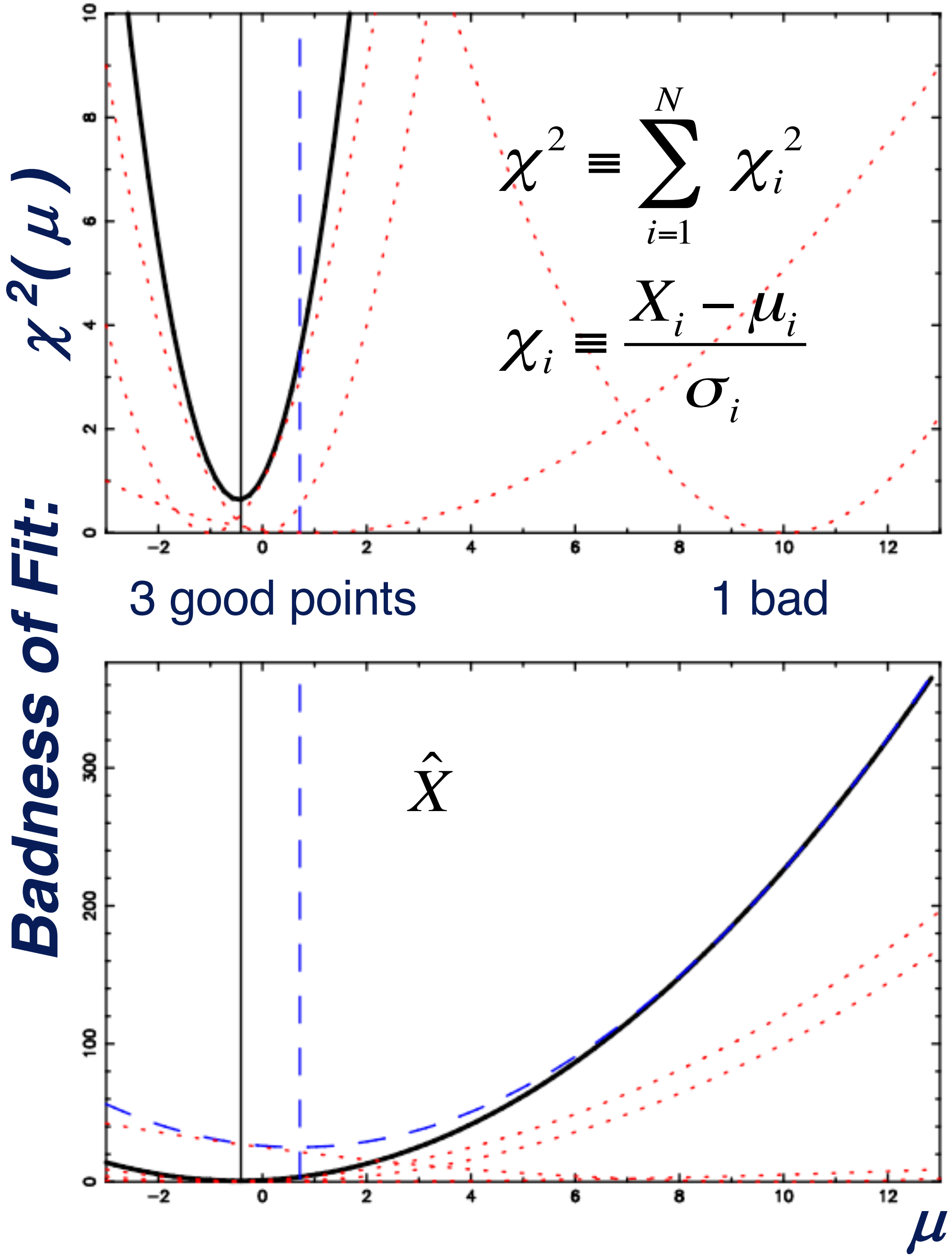
MAD = Mean Absolute Deviation



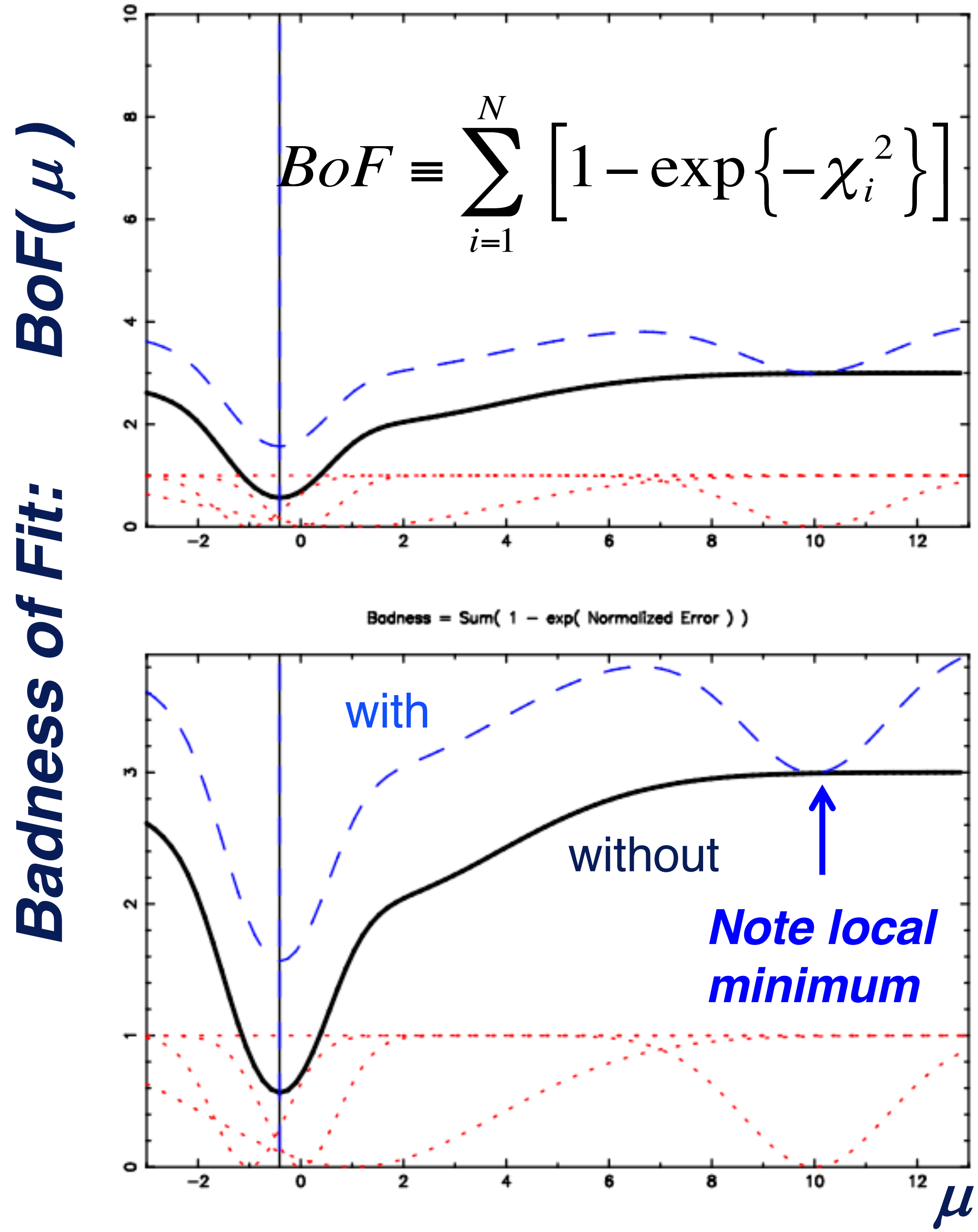


χ^2 = Sum of Squared Normalised Errors
SANE = Sum Absolute Normalised Errors
$\chi^2 \equiv \sum_{i=1}^{N} \chi_i^2$
$\chi_i \equiv \frac{X_i - \mu_i}{\sigma_i}$
$SANE \equiv \sum_{i=1}^{N} |\chi_i|$
Badness of Fit: $\chi^2(\mu)$
Badness of Fit: SANE(μ)
3 good points
1 bad
3 good points
1 bad
$\hat{X}$
with
without
μ
μ

χ^2 = Sum of Squared Normalised Errors



A "clipped" Badness of Fit Statistic



Fini -- ADA 06