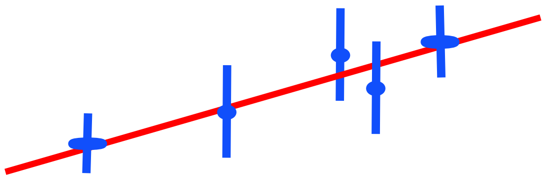
ADA 11 - 9am Thu 06 Oct 2022

Orthogonal Patterns (= orthogonal vectors using the data space metric)
e.g. Gram-Schmidt Orthogonalisation

Occam's Razor (model selection)
Information Criteria (AIC,BIC)

Review: Data Space Metric

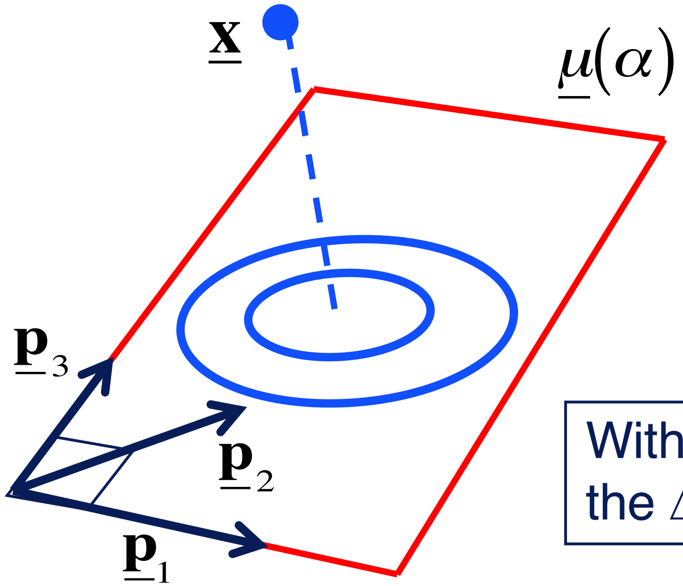


Data $X_i \pm \sigma_i \qquad i = 1 \ldots N$

Data space metric: $g_{ij} = \dfrac{\delta_{ij}}{\sigma_i^2}$

Distance from data to model:

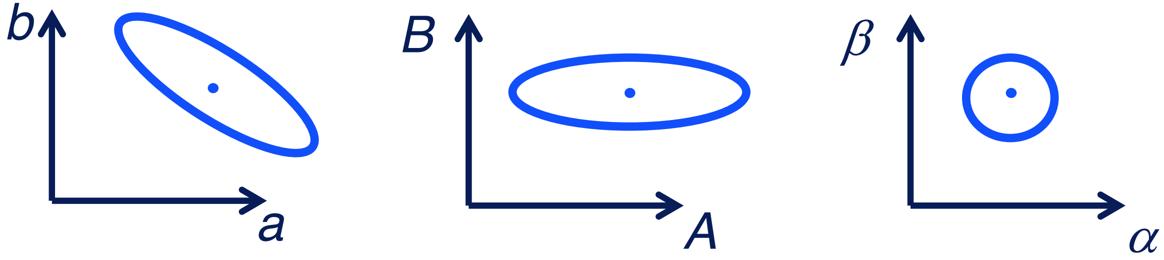
$$\left\| \underline{\mathbf{x}} - \underline{\mu}(\alpha) \right\|^2 = \sum_{i=1}^{N} \left(\frac{X_i - \mu_i(\alpha)}{\sigma_i} \right)^2 = \chi^2 .$$



With **the data-space metric**, the $\Delta\chi^2$ contours are **circular** !

For **linear models** (scaling patterns), the model surface $\underline{\mu}(\alpha)$ is a flat *M*-dimensional hyper-plane spanned by *M* vectors $\underline{\mathbf{p}}_k$

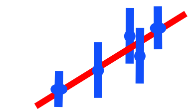
Non-orthogonal vs **Orthogonal** vs **Ortho-normal** patterns:



$$\underline{\mu} = \left(b\,\underline{\mathbf{p}}_1 + a\,\underline{\mathbf{p}}_2 \right) = \left(B\,\underline{\mathbf{p}}_1 + A\,\underline{\mathbf{p}}_3 \right) = \left(\beta\,\frac{\underline{\mathbf{p}}_1}{||\underline{\mathbf{p}}_1||} + \alpha\,\frac{\underline{\mathbf{p}}_3}{||\underline{\mathbf{p}}_3||} \right)$$

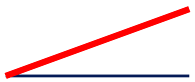
Scaling Orthogonal Patterns

Scaling **non-orthogonal patterns**:



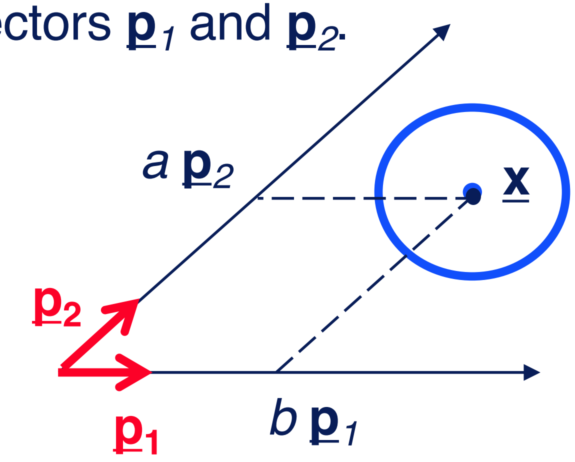
$$y = b + a\,x$$
$$= b\,\underline{\mathbf{p}}_1 + a\,\underline{\mathbf{p}}_2$$

$$\underline{\mathbf{p}}_1 = \{\,1, 1, \ldots, 1\,\}$$



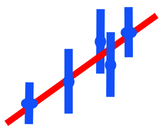
$$\underline{\mathbf{p}}_2 = \{x_1, x_2, \ldots, x_N\}$$

Model surface in the 2-dimensional data space is the plane spanned by vectors $\underline{\mathbf{p}}_1$ and $\underline{\mathbf{p}}_2$.



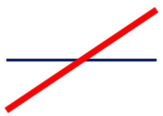
Data vector $\underline{\mathbf{x}}$ (above or below the model plane).

Scaling **orthogonal patterns**:



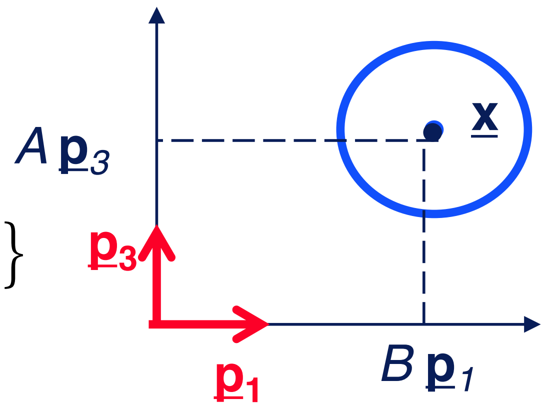
$$y = B + A\,(x - \hat{x})$$
$$= B\,\underline{\mathbf{p}}_1 + A\,\underline{\mathbf{p}}_3$$

$$\underline{\mathbf{p}}_1 = \{1, 1, \ldots, 1\}$$



$$\underline{\mathbf{p}}_3 = \{\,(x_1 - \hat{x}), (x_2 - \hat{x}), \ldots, (x_N - \hat{x})\,\}$$
$$= \quad \underline{\mathbf{p}}_2 - \hat{x}\,\underline{\mathbf{p}}_1$$

Model plane also spanned by orthogonal vectors $\underline{\mathbf{p}}_1$ and $\underline{\mathbf{p}}_3$.

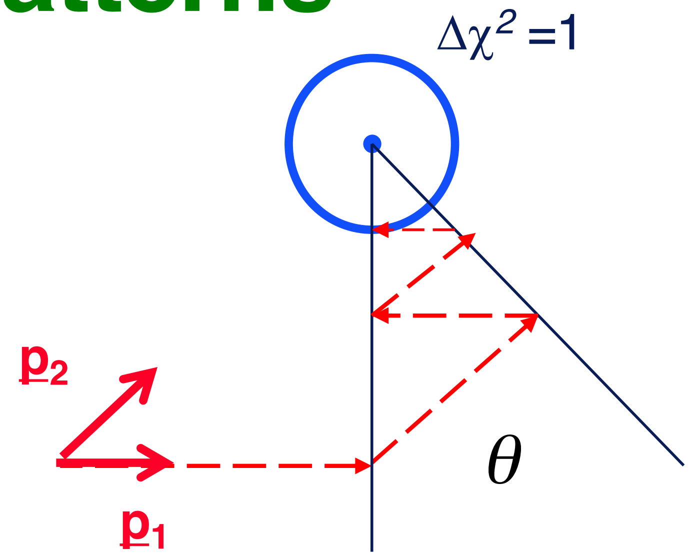


Since we can rotate these vectors, **orthogonal patterns are not unique.**

Scaling Orthogonal Patterns

Scaling **non-orthogonal patterns** can be <u>SLOW</u>.

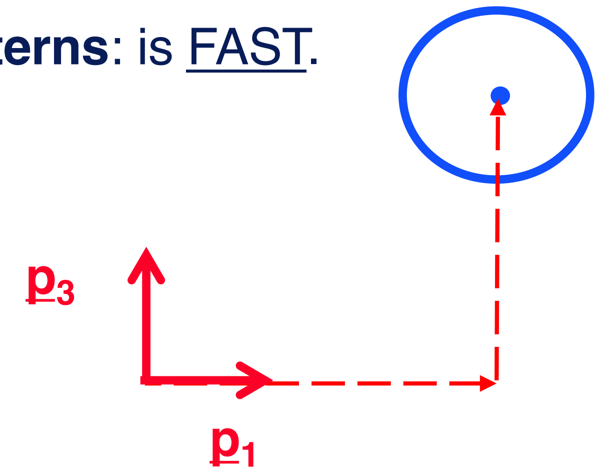
$$\text{Reach } k\sigma \text{ after 1 step,}$$
$$k\sigma\cos\theta \text{ after 2 steps,}$$
$$k\sigma\left(\cos\theta\right)^{n-1} \text{ after } n \text{ steps.}$$



Scaling **orthogonal patterns**: is <u>FAST</u>.

$$\theta = 90^\circ \qquad \cos\theta = 0$$

2 steps only !



Iterated Optimal Scaling

$$\hat{\underline{\mu}} = \sum_k^M \hat{\alpha}_k \underline{p}_k = \sum_k^M \hat{\beta}_k \frac{\underline{p}_k}{\left\|\underline{p}_k\right\|}$$

$$\hat{\alpha}_k = \frac{\underline{X} \bullet \underline{p}_k}{\underline{p}_k \bullet \underline{p}_k} \qquad \mathrm{Var}\left[\hat{\alpha}_k\right] = \frac{1}{\underline{p}_k \bullet \underline{p}_k}$$

$$\hat{\beta}_k = \underline{X} \bullet \frac{\underline{p}_k}{\left\|\underline{p}_k\right\|} \qquad \mathrm{Var}\left[\hat{\beta}_k\right] = 1$$

Scaling **ortho-normal patterns:**

$$\underline{p} \Rightarrow \frac{\underline{p}}{\left\|\underline{p}\right\|} \qquad \left\|\underline{p}\right\|^2 = \underline{p} \bullet \underline{p} = \sum_i \frac{p_i^2}{\sigma_i^2}$$

How to construct Orthogonal Patterns

- 1. Diagonalise Hessian Matrix
- 2. Gram-Schmidt Process
- 3. Differences between successive χ^2 fits

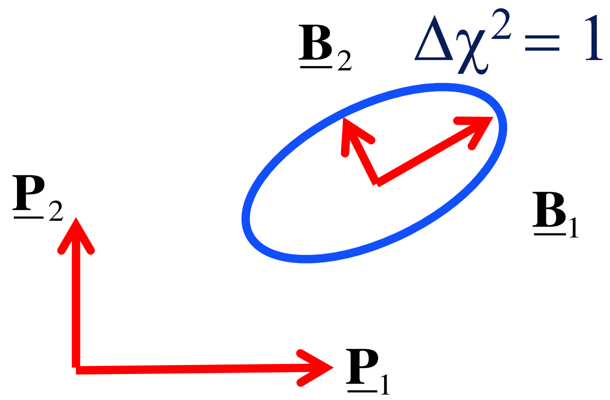
1. Diagonalise Hessian Matrix

- Quadratic approximation to χ^2 surface:

$$\underline{\mu}(\alpha) = \underline{\mu}(\hat{\alpha}) + \sum_i \underline{P}_i \, \Delta\alpha_i + \ldots \qquad \Delta\alpha_i \equiv \alpha_i - \hat{\alpha}_i$$

$$\chi^2(\alpha) = \chi^2(\hat{\alpha}) + \sum_{i,j} \Delta\alpha_i \, H_{ij} \, \Delta\alpha_j + \ldots$$

$$H_{ij} \equiv \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial\alpha_i \, \partial\alpha_j} \right|_{\alpha=\hat{\alpha}} \qquad \left(H^{-1}\right)_{ij} = \mathrm{Cov}\left[\hat{\alpha}_i, \hat{\alpha}_j\right]$$



- **Orthogonal basis** vectors $\underline{\mathbf{B}}_i$ are the **eigenvectors** of H_{ij} along the principal axes of the χ^2 contours.

$$\underline{\mu}(\beta) = \underline{\mu}(\hat{\beta}) + \sum_j \underline{B}_j \, \Delta\beta_j + \ldots$$

$$\chi^2(\beta) = \chi^2(\hat{\beta}) + \sum_j \left(\frac{\Delta\beta_j}{\sigma\left(\hat{\beta}_j\right)} \right)^2 + \ldots$$

$$H_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial\beta_i \, \partial\beta_j} \right|_{\beta=\hat{\beta}} = \frac{\delta_{ij}}{\sigma^2\left(\hat{\beta}_i\right)}$$

$$\left(H^{-1}\right)_{ij} = \delta_{ij} \, \sigma^2\left(\hat{\beta}_i\right)$$

Hessian Matrix for Non-Linear Models

Model and derivatives: $\underline{\mu}(\alpha)\,,\quad \frac{\partial\,\underline{\mu}}{\partial\,\alpha_k} \equiv \underline{\mathbf{P}}_k\,,\quad \frac{\partial^2\,\underline{\mu}}{\partial\,\alpha_k\,\partial\,\alpha_j} \equiv \underline{\mathbf{C}}_{kj}$

M Gradient vectors: $\underline{\mathbf{P}}_k$ *M(M+1) / 2* Curvature vectors: $\underline{\mathbf{C}}_{jk}$

Badness-of-fit:

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 = \left\|\underline{\mathbf{x}}-\underline{\mu}\right\|^2$$

$$\frac{\partial\,\chi^2}{\partial\,\alpha_k} = -2\sum_{i=1}^{N}\frac{x_i-\mu_i}{\sigma_i^2}\frac{\partial\,\mu_i}{\partial\,\alpha_k} = -2\left(\underline{\mathbf{x}}-\underline{\mu}\right)\bullet\underline{\mathbf{P}}_k$$

Hessian Matrix: $$H_{jk} = \frac{1}{2}\frac{\partial^2\,\chi^2}{\partial\,\alpha_j\,\partial\,\alpha_k} = \underline{\mathbf{P}}_j\bullet\underline{\mathbf{P}}_k - \left(\underline{\mathbf{x}}-\underline{\mu}\right)\bullet\underline{\mathbf{C}}_{jk}$$

Best fit Parameters: $$0 = \frac{\partial\,\chi^2}{\partial\,\alpha_k} = -2\left(\underline{\mathbf{x}}-\underline{\mu}\right)\bullet\underline{\mathbf{P}}_k \quad\Rightarrow\quad \underline{\mu}\left(\hat{\alpha}\right)\bullet\underline{\mathbf{P}}_k = \underline{\mathbf{x}}\bullet\underline{\mathbf{P}}_k$$

Parameter Error Bars: $$\mathrm{Cov}\left[\hat{\alpha}_j,\hat{\alpha}_k\right] = \left(H^{-1}\right)_{jk}$$

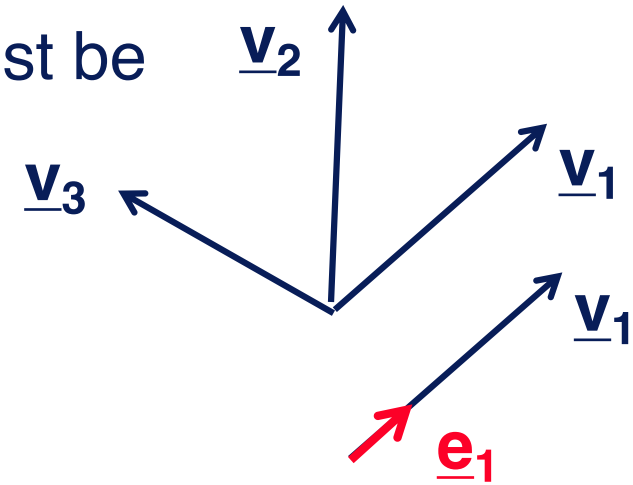
Linear Model: $$\underline{\mathbf{C}}_{jk} = 0 \qquad H_{jk} = \underline{\mathbf{P}}_j\bullet\underline{\mathbf{P}}_k$$

2. Gram-Schmidt Orthogonalization

The **Gram-Schmidt process**:

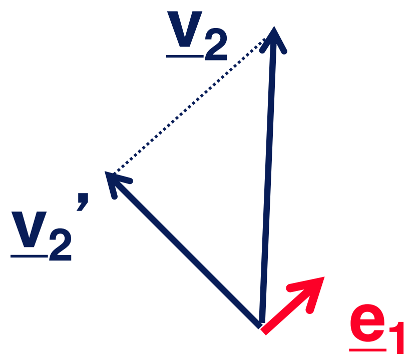
- 1. Start with *M* vectors $\underline{\mathbf{V}}_i$, $i = 1 \ldots M$. They must be independent, i.e. no two of them parallel.
- 2. Normalize vector $\underline{\mathbf{V}}_1$:

$$\underline{\mathbf{e}}_1 \equiv \frac{\underline{\mathbf{V}}_1}{\| \underline{\mathbf{V}}_1 \|}$$



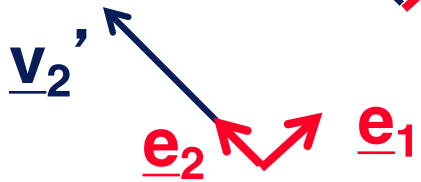
- 3. Make $\underline{\mathbf{V}}_2'$ perpendicular to $\underline{\mathbf{e}}_1$:
  - i.e. subtract component of $\underline{\mathbf{V}}_2$ in direction of $\underline{\mathbf{e}}_1$

$$\underline{\mathbf{V}}_2' = \underline{\mathbf{V}}_2 - (\underline{\mathbf{V}}_2 \bullet \underline{\mathbf{e}}_1)\, \underline{\mathbf{e}}_1$$



- 4. Normalize $\underline{\mathbf{V}}_2'$:

$$\underline{\mathbf{e}}_2 \equiv \frac{\underline{\mathbf{V}}_2'}{\left\| \underline{\mathbf{V}}_2' \right\|}$$



2. Gram-Schmidt Orthogonalization

- 5. Make $\underline{\mathbf{V}}_3'$ perpendicular to $\underline{\mathbf{e}}_1$:

$$\underline{\mathbf{V}}_3' = \underline{\mathbf{V}}_3 - (\underline{\mathbf{V}}_3 \bullet \underline{\mathbf{e}}_1)\, \underline{\mathbf{e}}_1$$

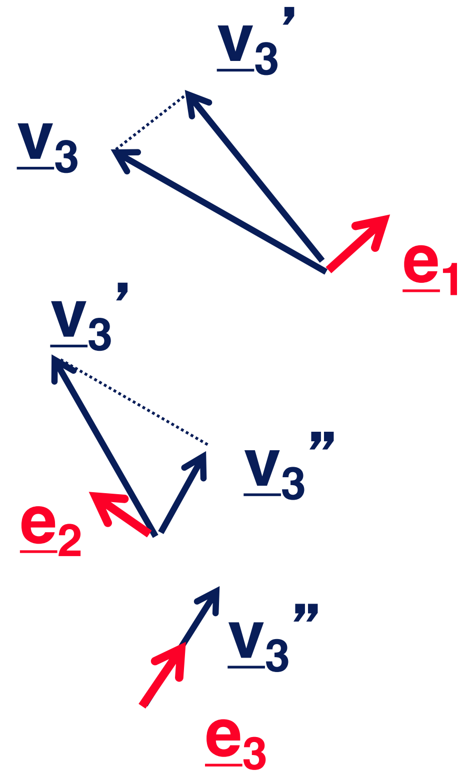
- 6. Make $\underline{\mathbf{V}}_3''$ perpendicular to $\underline{\mathbf{e}}_2$:

$$\underline{\mathbf{V}}_3'' = \underline{\mathbf{V}}_3' - (\underline{\mathbf{V}}_3' \bullet \underline{\mathbf{e}}_2)\, \underline{\mathbf{e}}_2$$

  – Note: $\underline{\mathbf{V}}_3''$ is perpendicular to $\underline{\mathbf{e}}_1$ AND $\underline{\mathbf{e}}_2$.

- 7. Normalize $\underline{\mathbf{V}}_3''$:

$$\underline{\mathbf{e}}_3 \equiv \frac{\underline{\mathbf{V}}_3''}{\left\| \underline{\mathbf{V}}_3'' \right\|}$$



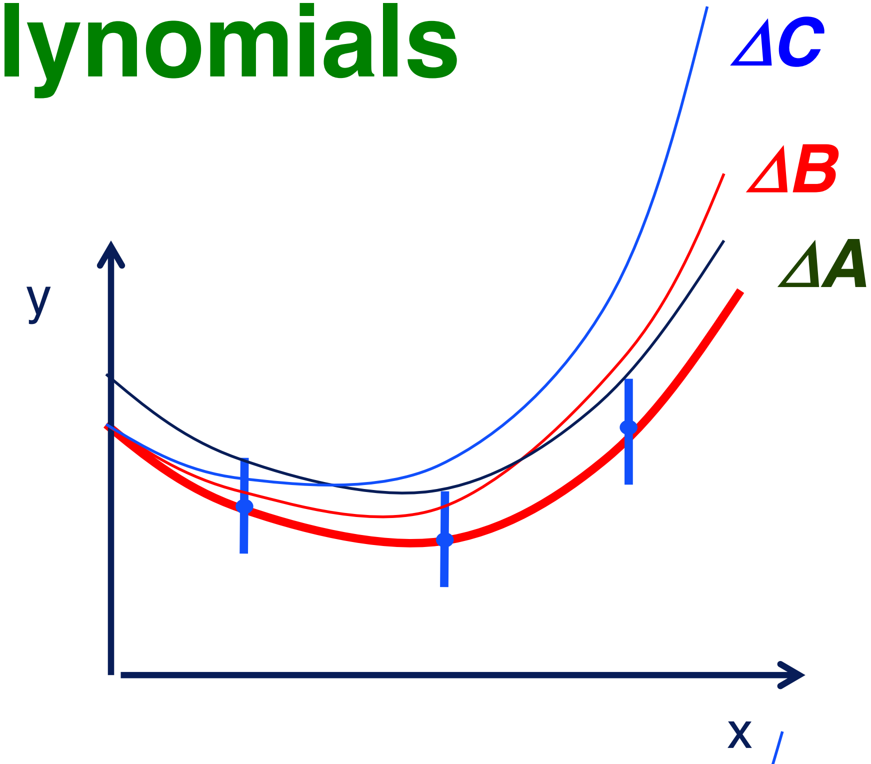
… Make $\underline{\mathbf{V}}_4$ perpendicular to $\underline{\mathbf{e}}_1$, $\underline{\mathbf{e}}_2$, $\underline{\mathbf{e}}_3$ and normalise to get $\underline{\mathbf{e}}_4$.

Repeat up to $\underline{\mathbf{V}}_M$ to get **complete ortho-normal basis** $\underline{\mathbf{e}}_1$, $\underline{\mathbf{e}}_2$, ..., $\underline{\mathbf{e}}_M$.

Easy to code ! (Try it !)

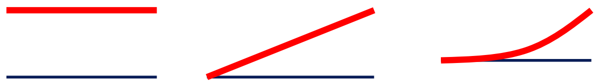
.

Orthogonal Polynomials



non-orthogonal polynomial:

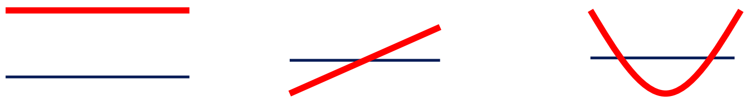
$$y = A + B\,x + C\,x^2$$

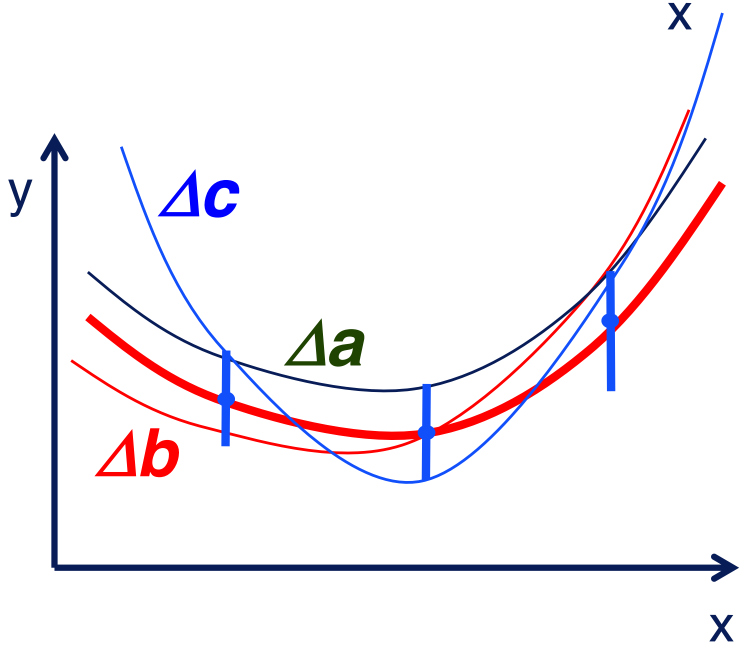


orthogonal polynomials:

$$P_i \bullet P_j \equiv \sum_{k=1}^{N} \frac{P_i(x_k)\,P_j(x_k)}{\sigma_k^{\,2}} = \frac{\delta_{ij}}{\mathrm{Var}\left[\alpha_i\right]}$$

$$y = a\,P_0(x) + b\,P_1(x) + c\,P_2(x)$$

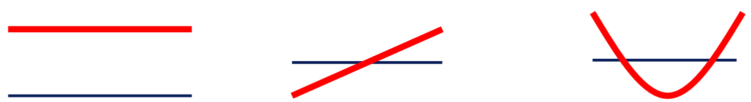




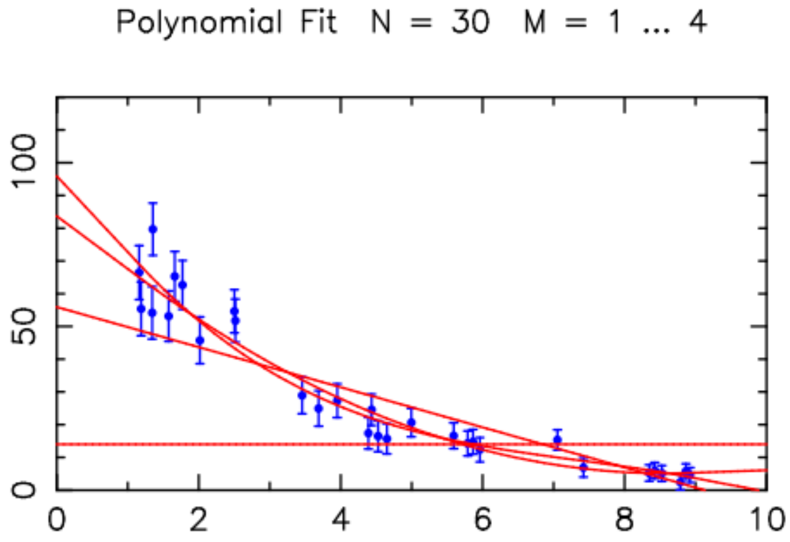
Note: every dataset has its own $1/\sigma^2$ weights, defines its own orthogonal polynomials.

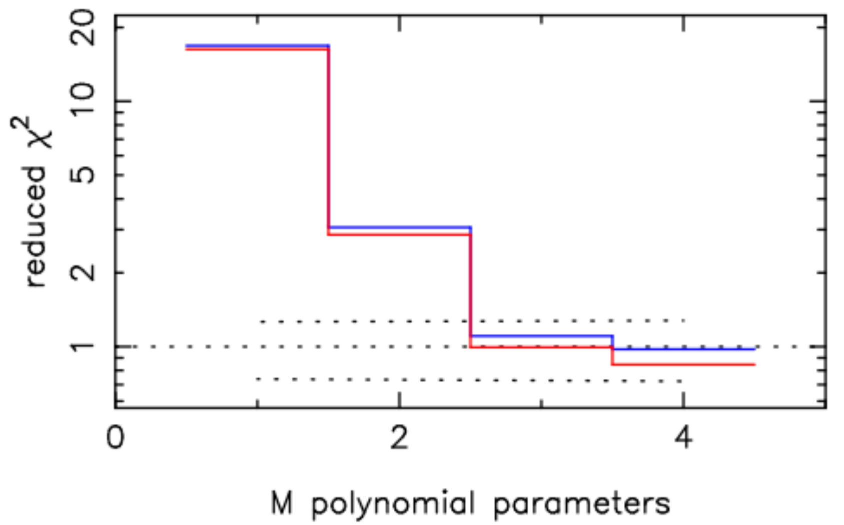
3. Differences between successive χ^2 fits

- Fit: $A + B\,x + C\,x^2$
  - A, B, C are not independent
  - $1, x, x^2$ are not orthogonal
- If $P_k(x)$ is a polynomial of degree k fitted to the data, then $P_k(x) - P_{k-1}(x)$ **are orthogonal**:
- $a\,P_0(x) + b\,[\,P_1(x) - P_0(x)\,] + c\,[\,P_2(x) - P_1(x)\,]$
  - a, b, c are independent



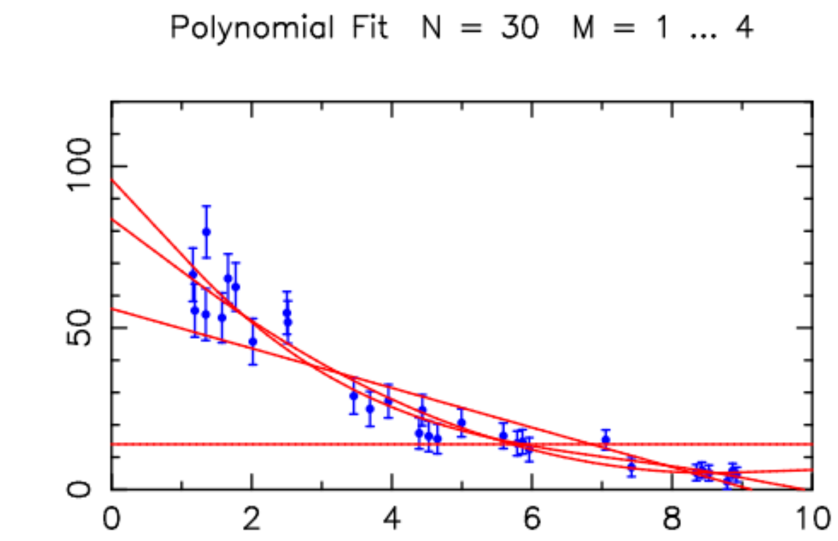
Note: every dataset has its own $1/\sigma^2$ weights, defines its own orthogonal polynomials.

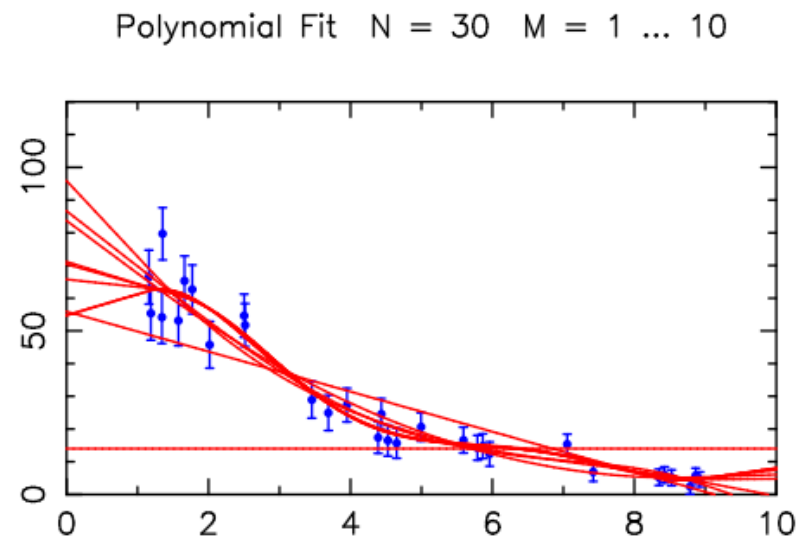




Mock data, true model is an exponential.

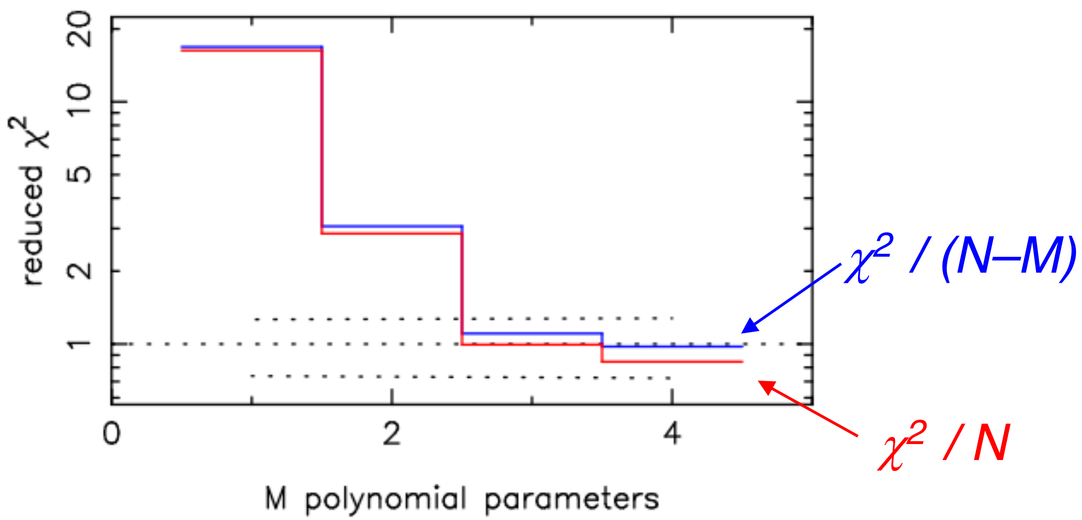
Polynomial Fits

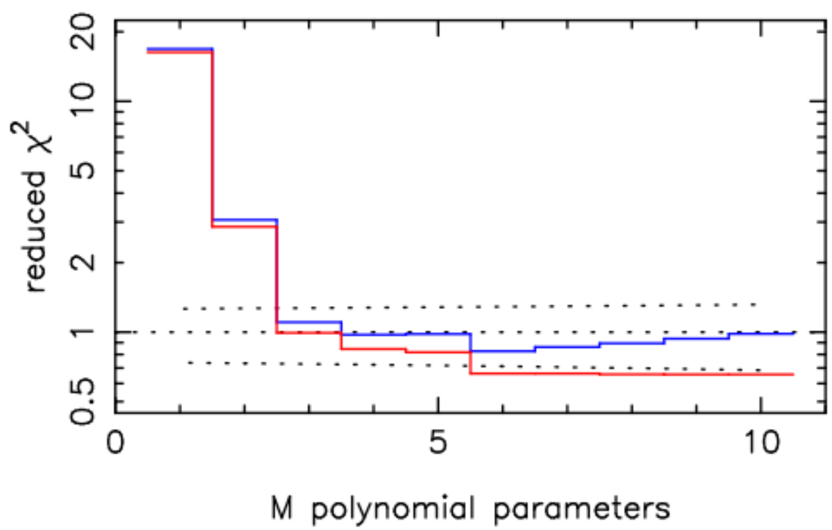




Badness-of-Fit : $\chi^2 / N \rightarrow 0 \text{ as } M \rightarrow N$

$$\frac{\chi^2}{N-M} \approx 1 \pm \left(\frac{2}{N-M} \right)^{1/2}$$

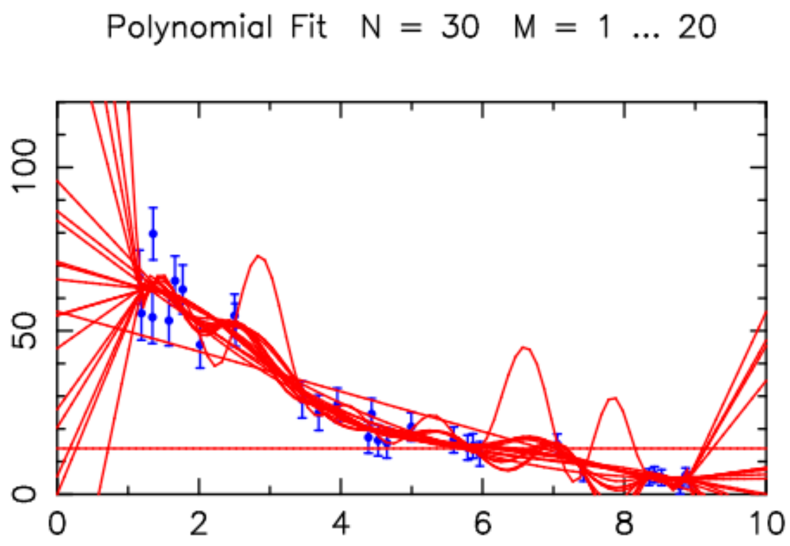




How many parameters to use ?

Fit *N = 30* data points with
M = 1, 2, ... 20 poly coefficients.

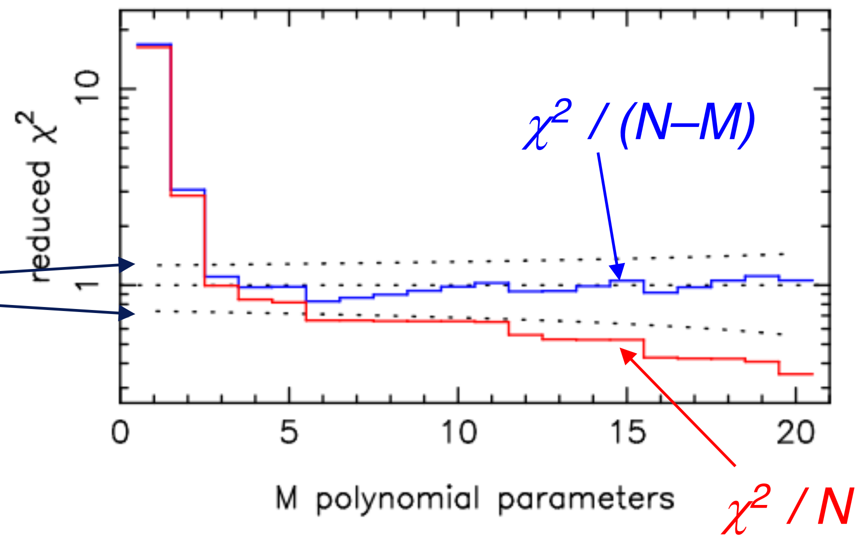
> Higher *M* = more flexible model.
>
> Lower χ^2, but less satisfactory fit.



Note “flailing” in data gaps
and beyond ends for high *M*

χ^2_{min} rejects *M = 1, 2 .*
accepts *M = 3, 4, ...*

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

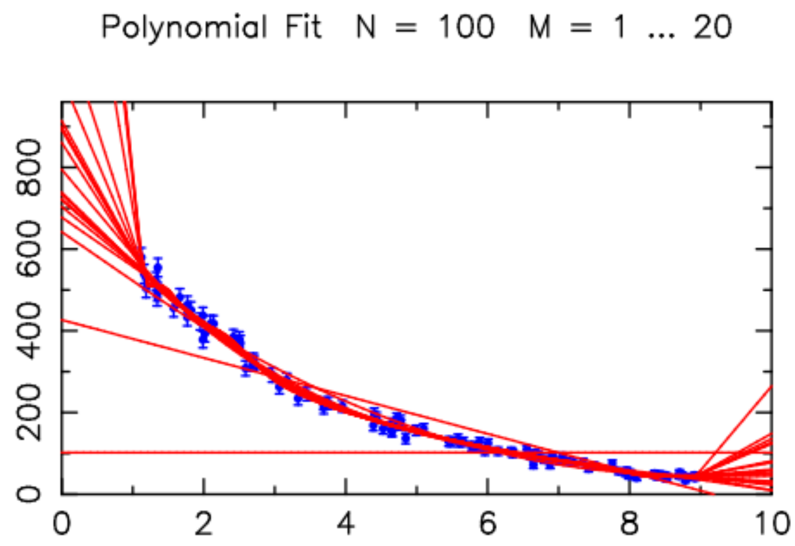


How many parameters to use ?

Fit $N = 100$ data points with

$M = 1, 2, \ldots 20$ poly coefficients.

More data points and smaller error bars than before.

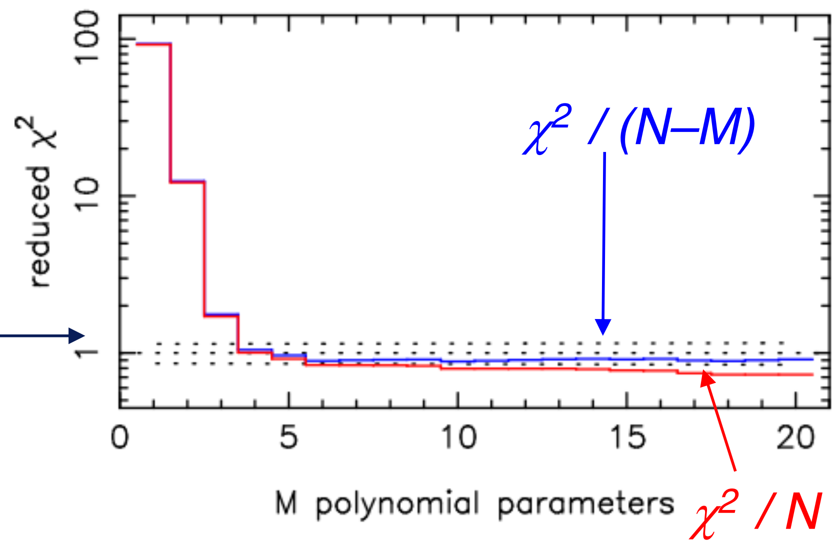


Note "flailing" beyond the range of the data for high *M*

χ^2_{min} rejects $M = 1, 2, 3.$

accepts $M = 4, 5, \ldots.$

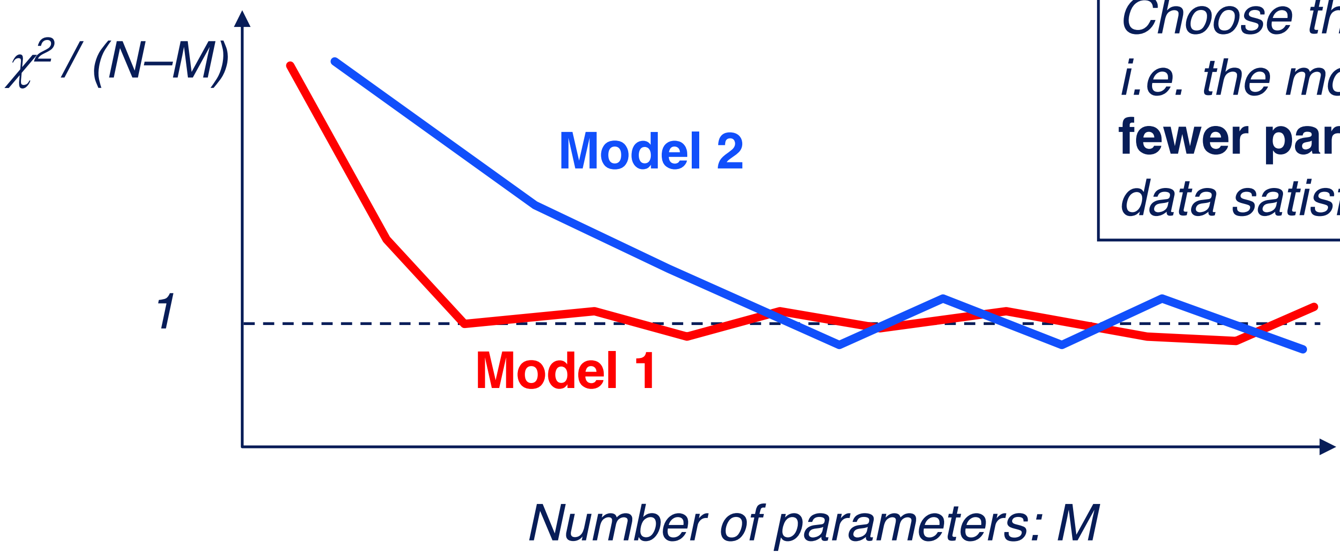
$$\frac{\chi^2}{N-M} \approx 1 \pm \sqrt{\frac{2}{N-M}}$$



Occam's Razor - "Keep it Simple"

- William of Occam (ca. 1286–1347):
  "It is futile to do with more, what can be done with fewer"
  or: "Keep it simple!"
- Fit 2 different models, 1 and 2, to the same data.
- Each model has M =1,2,… parameters,
  e.g. increasing numbers of polynomial coefficients.
- **Prefer the simpler model**.

Choose the **simpler model**, *i.e. the model needing* **fewer parameters** *to fit the data satisfactorily.*



Information Criteria: AIC, AICc, BIC

- Each parameter improves the fit: -2 ln(L) decreases.
- Include a penalty for each new parameter.
- Does the reduction in -2 ln(L) offset the penalty?

Data points: N Parameters: M Likelihood: L

Akaike Information Criterion: $AIC \equiv -2\ln(L) + 2\,M$

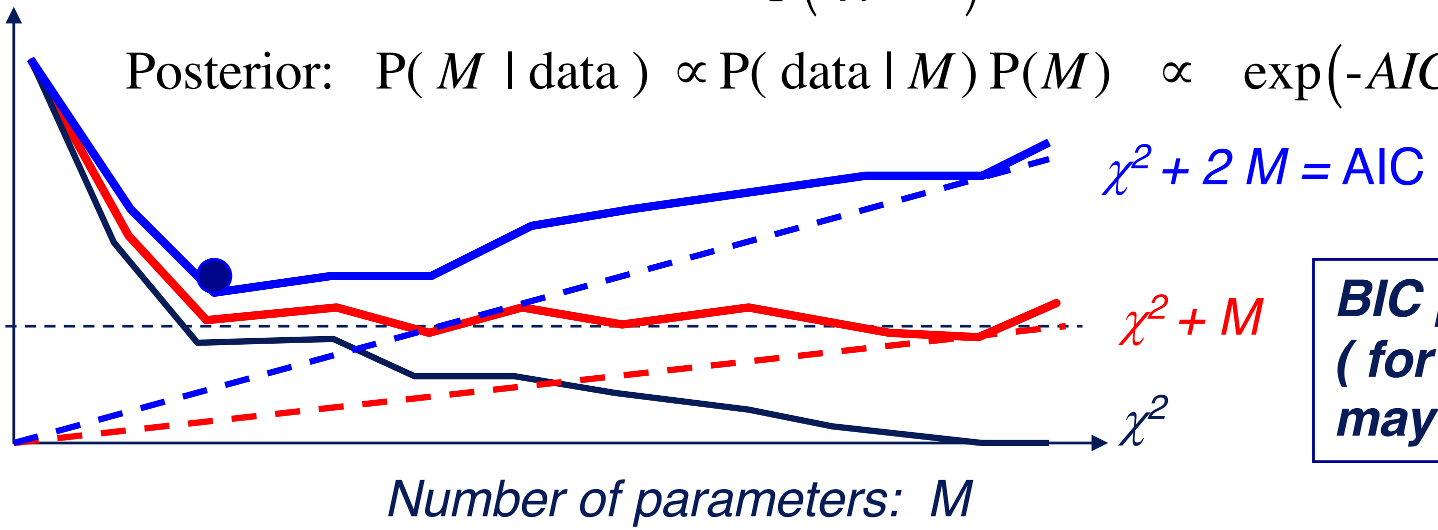
Corrected AIC: $AICc \equiv -2\ln(L) + 2M \Big/ \left(1 - \dfrac{M-1}{N}\right)$

Bayesian Information Criterion: $BIC \equiv -2\ln(L) + \ln(N)\,M$

***Minimise the AIC** (or **AICc** or **BIC**) to choose the **simplest model** (needing the **fewest parameters**) that fits the data.*

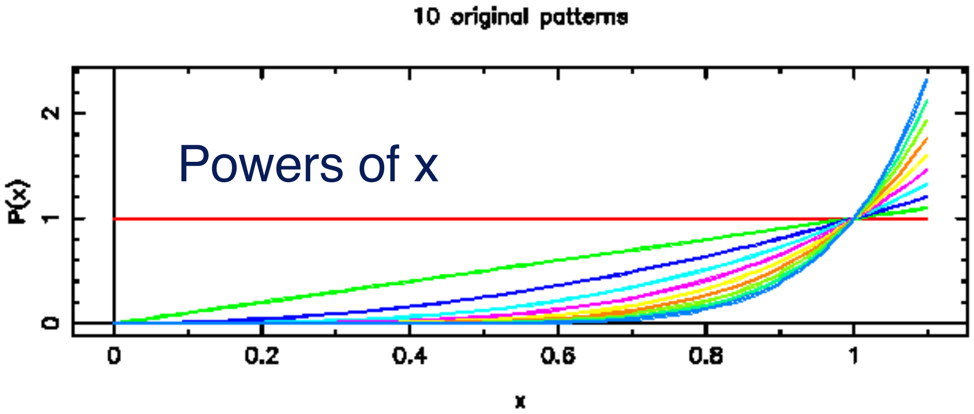
Likelihood: $\mathrm{P}(\text{ data} \mid M\,) \propto \exp\left(-\chi^2/2\right)$ Prior: $\mathrm{P}(M\,) \propto \exp(-M\,)$

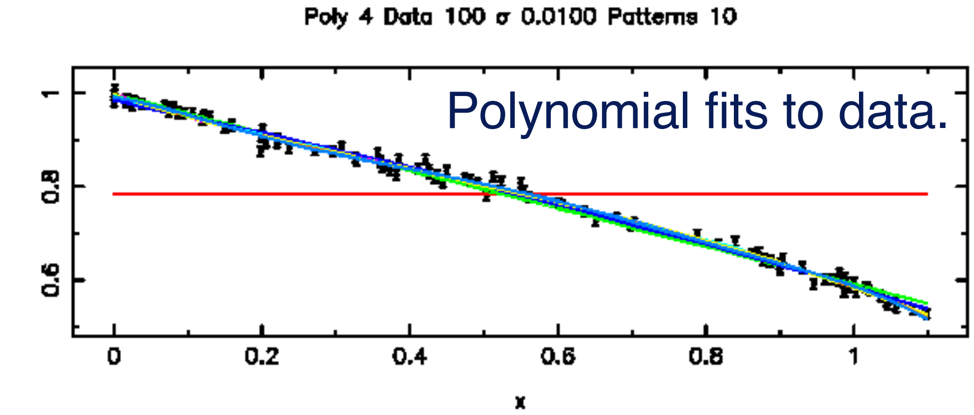
Posterior: $\mathrm{P}(\,M \mid \text{data}\,) \propto \mathrm{P}(\text{ data} \mid M\,)\,\mathrm{P}(M\,) \propto \exp\left(-AIC/2\right)$

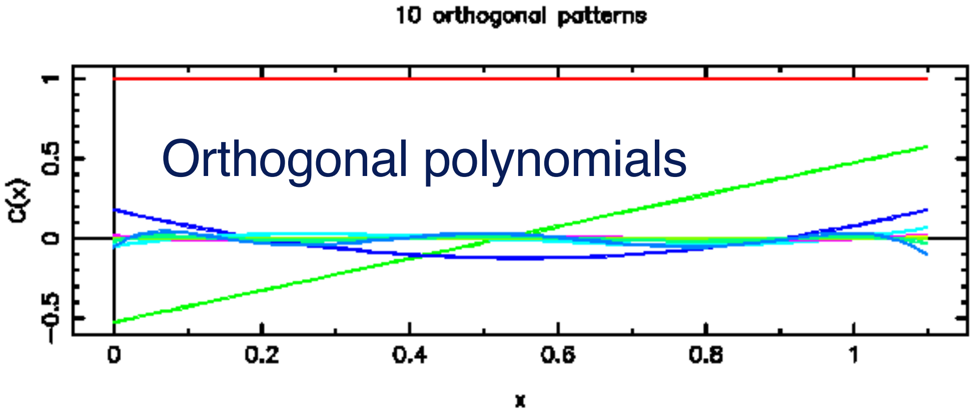


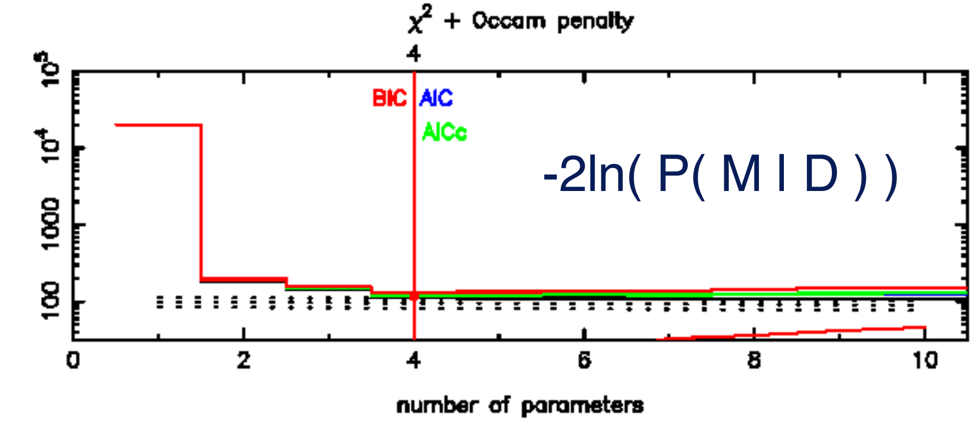
BIC prefers simpler models (for ln(N) > 2, N > 7) and may be better than AIC.

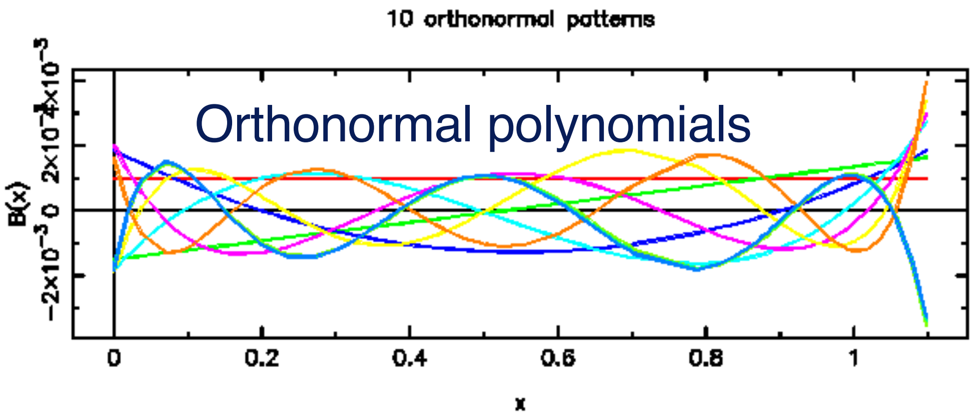
Comparison of AIC, AICc, BIC

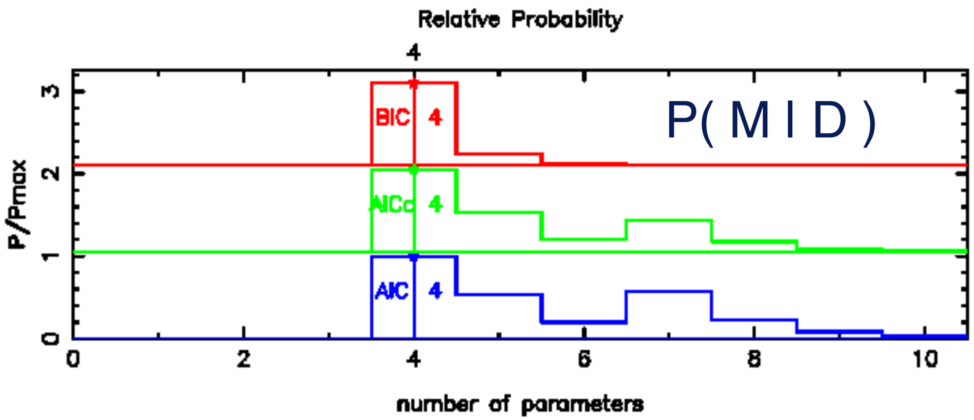




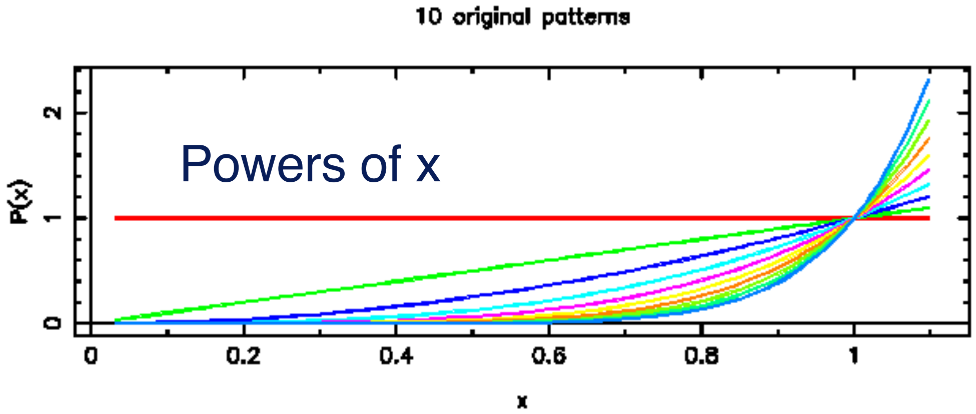


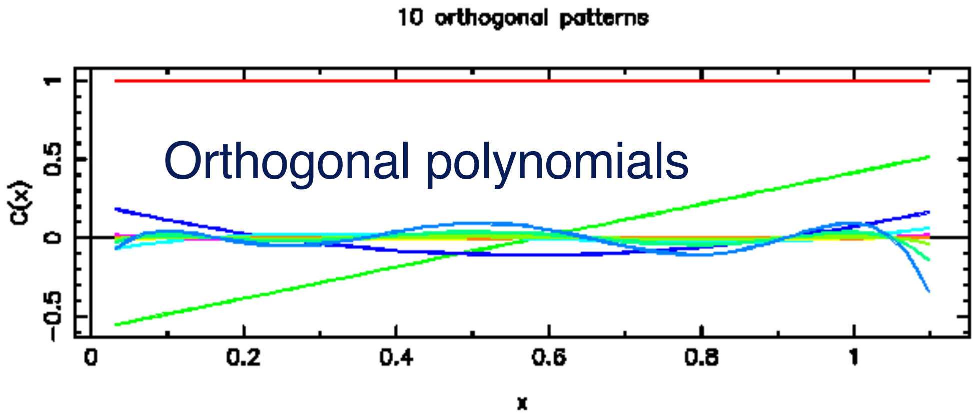


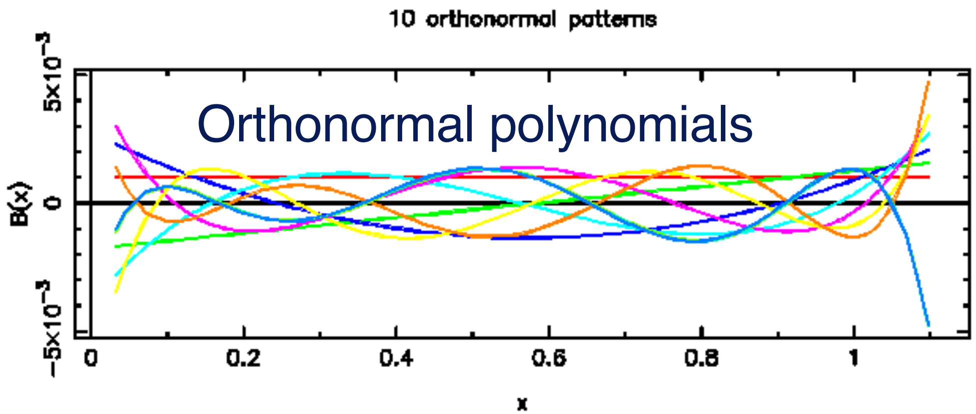


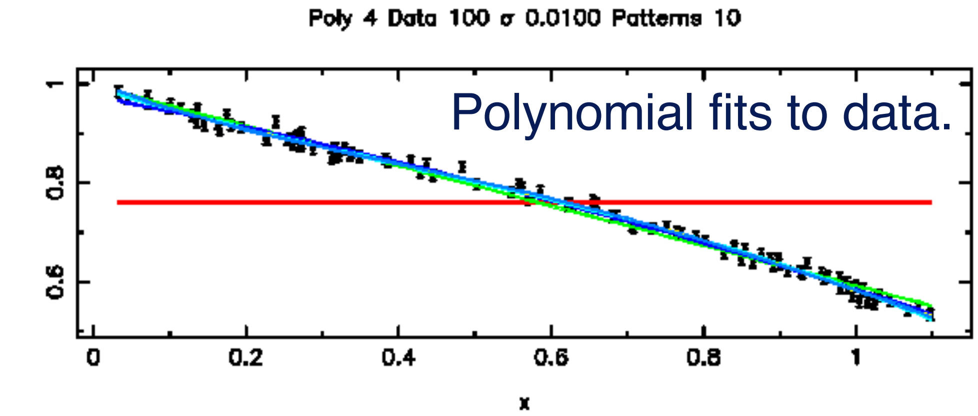


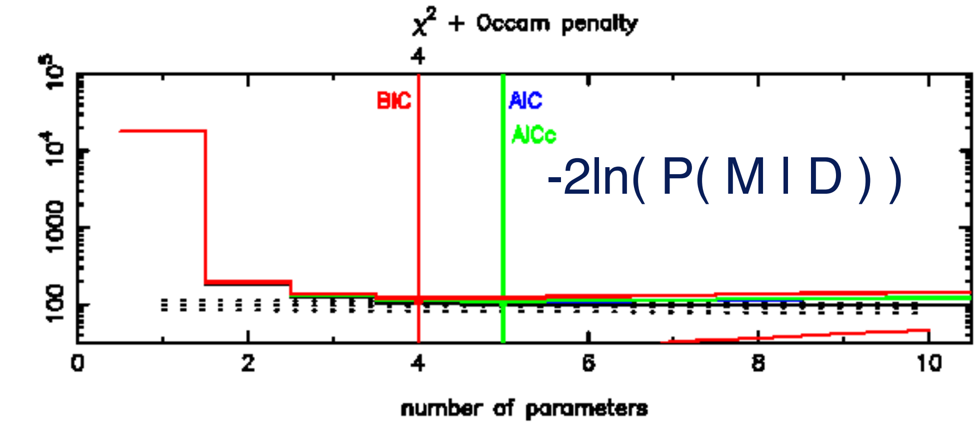
Comparison of AIC, AICc, BIC

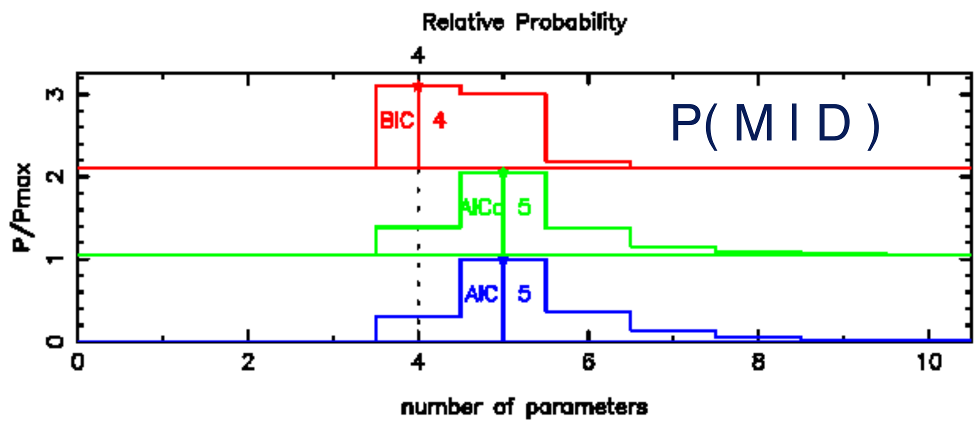












Fini -- ADA 11