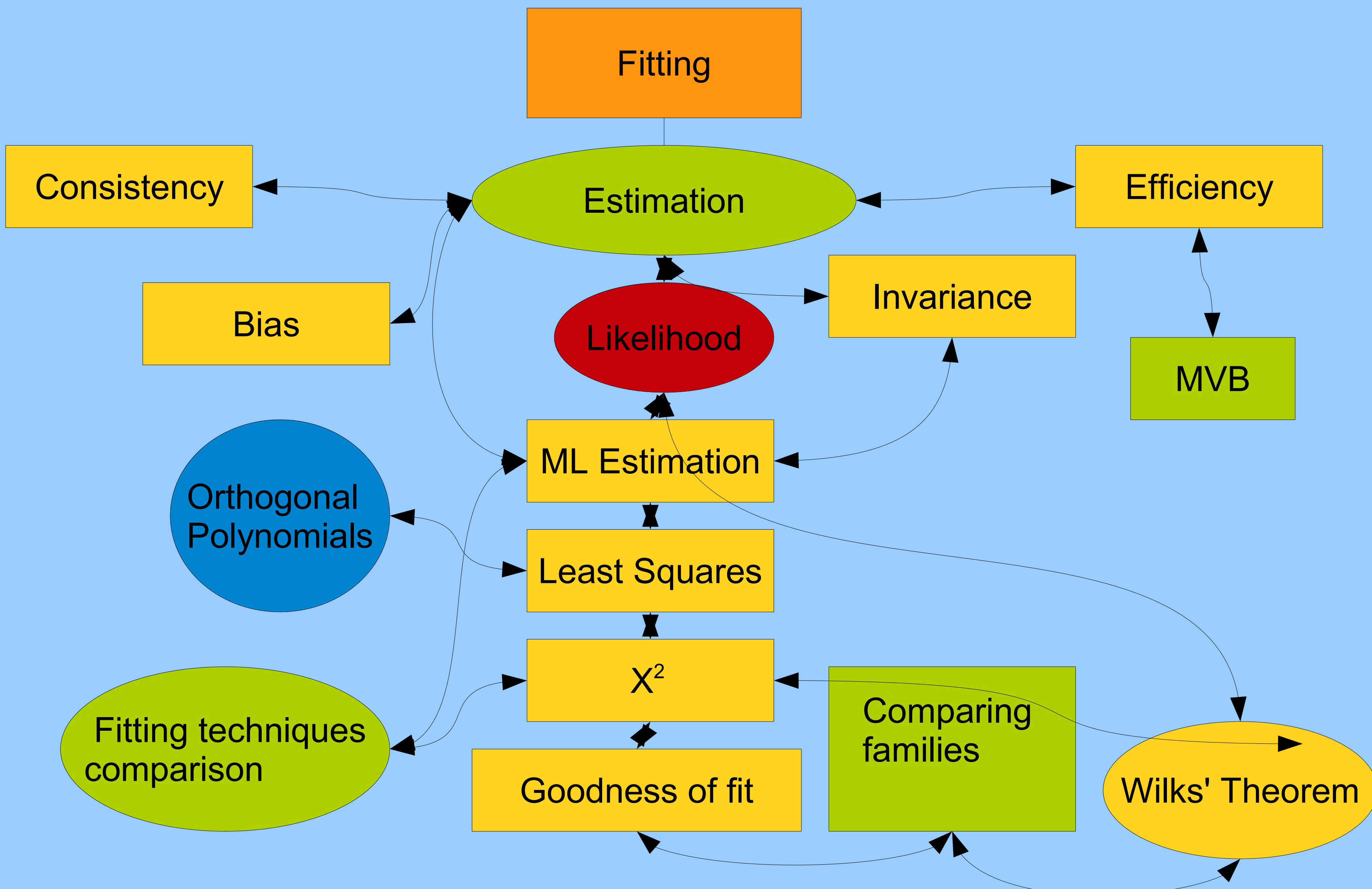
Statistics (2) Fitting

Roger Barlow
Manchester University

IDPASC school
Sesimbra
14th December 2010

Summary


Fitting
Consistency
Estimation
Efficiency
Invariance
Bias
Likelihood
MVB
ML Estimation
Orthogonal Polynomials
Least Squares
X2
Fitting techniques comparison
Comparing families
Goodness of fit
Wilks' Theorem

Fitting and Estimation

Data sample $\{x_1,x_2,x_3,...\}$ confronts theory – pdf $P(x;a)$

(*a* may be multidimensional)

Estimator $\hat{a}(x_1,x_2,x_3,..)$ is a process returning a value for *a*.

A 'good' estimator is

- Consistent
- Unbiassed
- Invariant
- Efficient

These qualities depend on the estimator and on the particular pdf

Explanations follow. Introduce (again) the Likelihood

$$L(x_1,x_2,x_3,...;a) = P(x_1;a)\,P(x_2;a)\,P(x_3;a)\,...$$

Consistency

Introduce the Expectation value

$$\langle f \rangle = \iiint \ldots \; f(x_1,x_2,x_3,\ldots)\, L(x_1,x_2,x_3,..a)\; dx_1\, dx_2\, dx_3,..$$

Integrating over the space of results but not over *a*.

It is the average you would get from a large number of samples. Analogous to Quantum Mechanics.

Consistency requires: $\mathrm{Lt}_{N \to \infty} \langle \hat{a} \rangle = a$

i.e. given more and more data, the estimator will tend to the right answer

This is normally quite easy to establish

Bias

Require <â>=a (even for finite sample sizes)

If a bias is known, it can be corrected for

Standard example: estimate mean and variance of pdf from data sample

$$\hat{\mu} = \frac{1}{N} \sum x_i \qquad \hat{V} = \frac{1}{N} \sum (x_i - \hat{\mu})^2$$

This tends to underestimate V. Correct by factor N/(N-1)

Invariance

Desirable to have a procedure which is transparent to the form of a, i.e. need not worry about the difference between $\hat{a}^2$ and $\widehat{a^2}$

This is incompatible with unbiassedness. The well known formula (previous slide) is unbiassed for V but biassed for σ

Efficiency

Minimise <(â-a)^2>

The spread of results of your estimator about the true value

Remarkable fact: there is a limit on this

(Minimum Variance Bound, or Cramer-Rao bound)

$$V(\hat{a}) \geq \frac{-1}{\left\langle \frac{d^2 \ln L}{da^2} \right\rangle}$$

Some examples

Repeated Gaussian measurements

$$\hat{\mu} = \frac{1}{N} \sum x_i$$

Bias
$$\iiint dx_1\, dx_2\, dx_3 \left(\frac{(x_1-\mu)}{N} + \ldots \right) \frac{e^{-(x_1-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \ldots = 0$$

Variance
$$\iiint dx_1\, dx_2\, dx_3 \left(\frac{(x_1-\mu)^2}{N^2} + \ldots \right) \frac{e^{-(x_1-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \ldots = \frac{\sigma^2}{N}$$

MVB
$$\ln L = \sum \frac{-(x_i-\mu)^2}{2\sigma^2} - N \ln(\sigma\sqrt{2\pi}); \qquad \frac{d^2 \ln L}{d\mu^2} = \frac{-N}{\sigma^2}$$

More examples

Centre of a top hat function: ½(max + min)

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

More efficient than the mean.

Several Gaussian measurements with different σ: weight each measurement by $(1/\sigma)^2$. - normalised

But don't weight Poisson measurements by their value....

Maximum Likelihood

Estimate *a* by choosing the value which maximises $L(x_1,x_2,x_3,..a)$. Or, for convenience, $\ln L = \Sigma \ln P(x_i,a)$

Consistency	Yes
Bias-free	No
Invariance	Yes
Efficiency	Yes, in large N limit

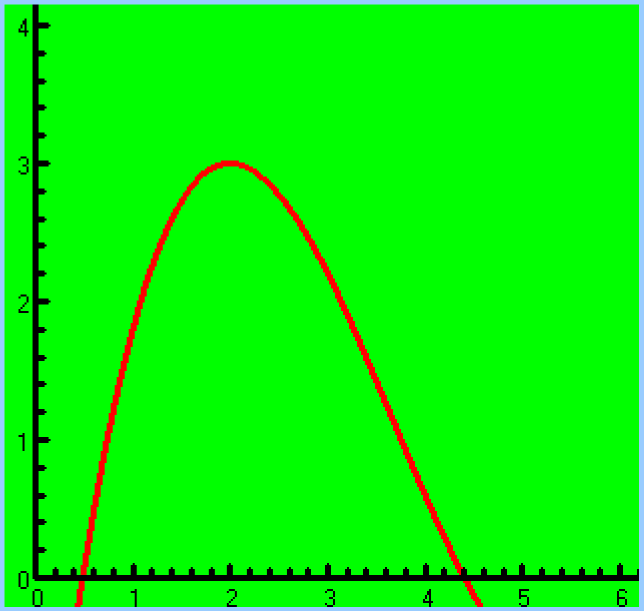
This is a technique, but not the only one.

Use by algebra in simple cases or numerically in tougher ones

Numerical ML

Adjust *a* to maximise Ln L

If you have a form for (dln L/da) that helps a lot.



Use MINUIT or ROOT or...., especially if *a* is multidimensional

Algebraic ML

Maximising: requires $\Sigma\, d \ln P(x_i,a)/da = 0$

This leads to fractions with no nice solution
– unless P is exponential.

Given set of x_i, measured y_i, predictions $f(x_i)$ subject to Gaussian smearing – Max likelihood mean minimising

$$\chi^2 = \frac{\sum (y_i - f(x_i;a))^2}{\sigma_i^2}$$

Classic example: straight line fit *f(x)=mx+c*

$$m = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}; \qquad c = \overline{y} - m\overline{x}$$

The Normal Equations

If f is linear function of $a_1, a_2, a_3 \ldots a_M$

– $f_i = f(x_i) = \Sigma a_j g_j(x_i)$

Maximum Likelihood = Minimum χ^2

$$\sum 2\left(y_i - \sum a_j g_j(x_i)\right) g_k(x_i) = 0$$

$$\sum y_i g_k(x_i) = \sum a_j \sum g_j(x_i) g_k(x_i)$$

Solve for the coefficients a_j by inverting matrix

Orthogonal Polynomials

Good trick: construct the g(x) functions so that the matrix is diagonal

If fitting polynomial up to 5th power (say), can use $1,x,x^2,x^3,x^4,x^5$ or $1,x,2x^2-1,4x^3-3x,8x^4-8x^2+1,16x^5-20x^3+5x$, or whatever

Choose $g_0=1$

Choose $g_1 = x-(\Sigma x)/N$ so that makes $\Sigma g_0 g_1=0$

And so on iteratively $g_r(x)=x^r + \Sigma c_{rs} g_s(x)$

$$c_{rs}=-\Sigma x_i^r g_s(x_i)/\Sigma g^2{}_s(x_i)$$

These polynomials are orthogonal over a specific dataset

Fitting histograms

Raw data $\{x_1,x_2,x_3,...x_N\}$

Often sorted into bins $\{n_1,n_2,n_3,...n_m\}$

Number of entries in bin is Poisson

$$\chi^2=\sum\frac{(n_i-f(x_i;a))^2}{\sigma_i^2}\rightarrow\sum\frac{(n_i-f(x_i;a))^2}{f(x_i;a)}\rightarrow\sum\frac{(n_i-f(x_i;a))^2}{n_i}$$

Last form sometimes used as a definition for χ^2, though really only an approximation

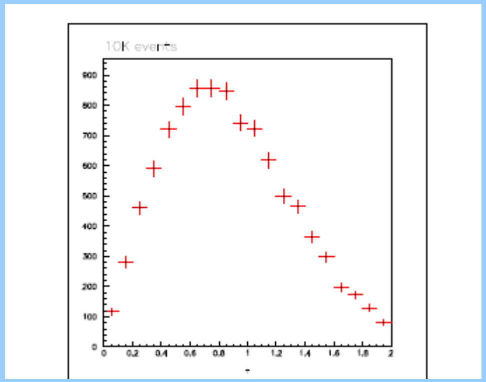
Fit function to histogram by minimising χ^2.

4 Techniques

1) Minimise naïve χ^2. Computationally easy as problem linear
2) Minimise full χ^2. Slower as problem nonlinear due to terms in the denominator
3) Binned Maximum Likelihood. Write the Poisson probability for each bin $e^{-f_i} f_i^{n_i}/n_i!$ and maximise the sum of logs
4) Full maximum likelihood without binning

Consumer test

Fit $$f(x) = \frac{1}{2a} x e^{-ax^2}$$



Try (many times) with10,000 events

All methods give same results

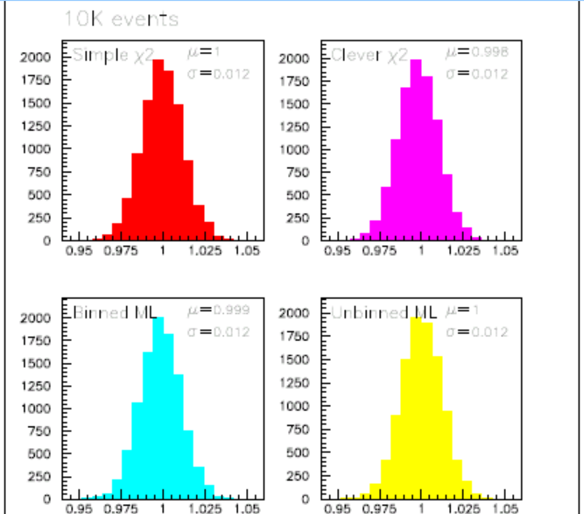


Figure 7: Results with 10,000 events.

Histogram fitting (contd)

- With small sample (100 events)
- Simple χ^2 goes bad due to bins with zeros
- Full χ^2 not good as Poisson is not Gaussian
- Two ML methods OK

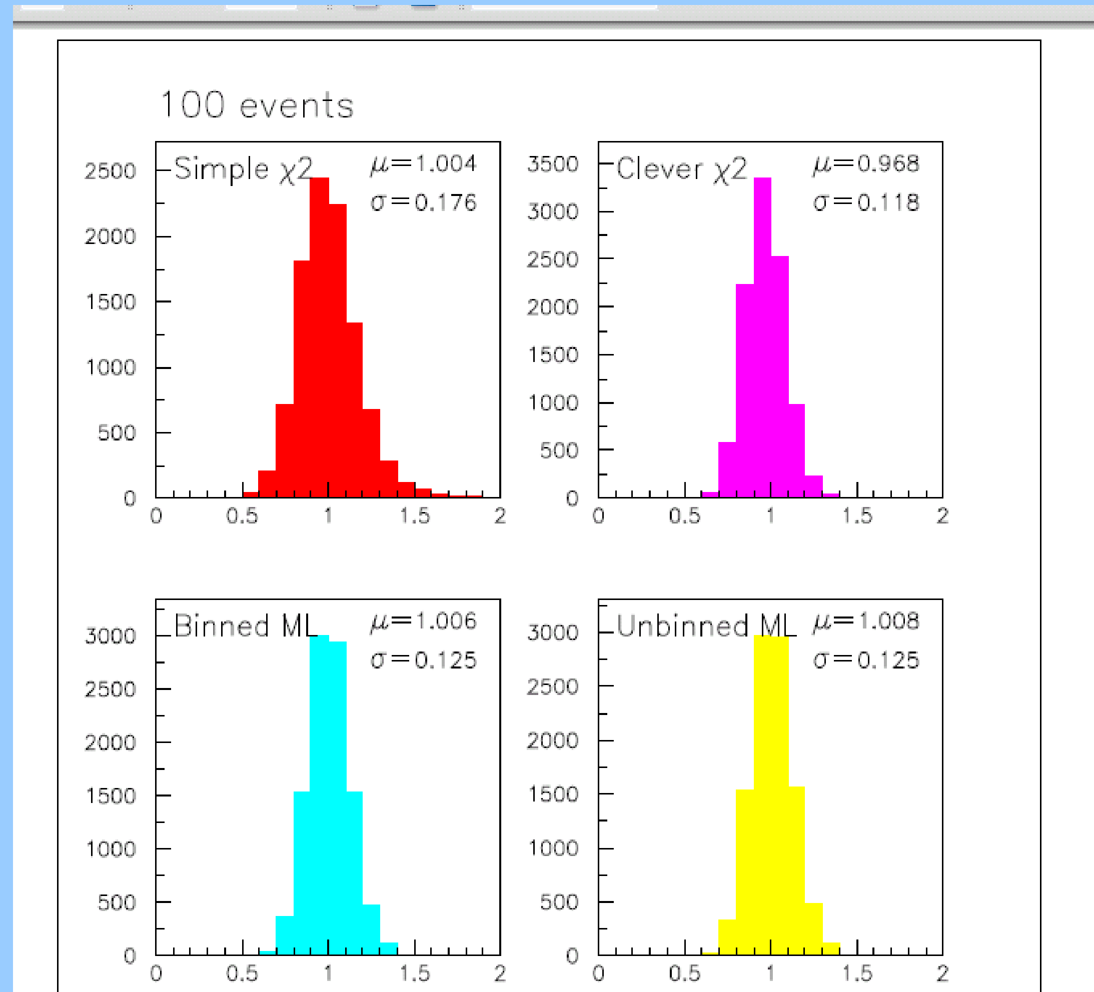


Figure 9: Results with 100 events

Goodness of fit

- Each term is clearly of order 1.
- Full treatment by integrating multi-d gaussian gives χ^2 distribution $P(\chi^2,N)$
- Mean indeed N. Shapes vary
- If the fit is bad, χ^2 is large

$$\chi^2 = \sum \left(\frac{y_i - f(x_i;a)}{\sigma_i} \right)^2$$

$$\int_{\chi^2}^{\infty} P(\chi'^2;N)\, d\chi'^2$$

Is a p value.
Often called " χ^2 probability"

Goodness of fit

Large χ^2 >> N, low p value means:

- The theory is wrong
- The data are wrong
- The errors are wrong
- You are unlucky

If you histogram the p values from many cases (e.g. kinematic fits) the distribution should be flat.
This is obvious if you think about it in the right way

Small χ^2 << N, p value~1 means:

- The errors are wrong
- You are lucky

Exact χ^2 = N means the errors have been calculated from this test, and it says nothing about goodness of fit

Nice extra feature

If one (or more) of the parameters in the function have been fitted to the data, this improves the χ^2 by an amount which corresponds to 1 less data point

Hence 'degrees of freedom' $N_D=N-N_P$

Likelihoood and Goodness of fit

No test available, sorry

Likelihoood and Goodness of fit!!!

Take a 'Toy Monte Carlo' which simulates your data many times, fit and find the likelihood.

Use this distribution to obtain a p value for your likelihood

This is not in most books as it is computationally inefficient. But who cares these days?

Wilks' Theorem

Often stated that $\Delta \ln L = -2\chi^2$

This strictly relates to changes in likelihood caused by an extra term in model. Valid for relative comparisons within families

E.g. Fit data to straight line. χ^2 sort of OK

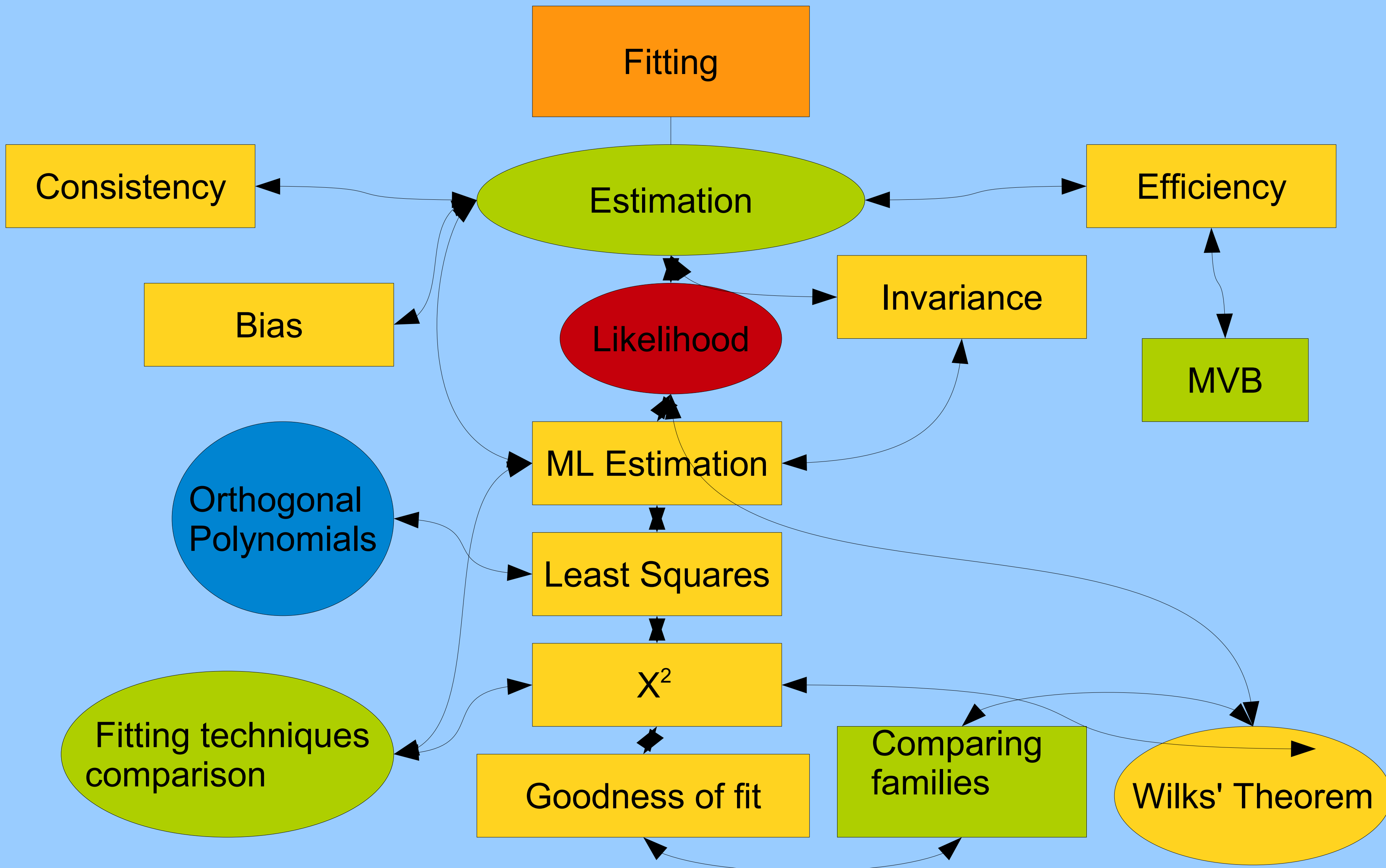
Fit using parabola. χ^2 improves. If this improvement is >>1 the parabola is doing a better job. If only ~1 there is no reason to use it

GoF comparisons and likelihood

Wilks' theorem lets you compare the merit of adding a further term to your parametrisation: yardstick for whether improved likelihood is significant. Does not report absolute merit as χ^2 does

Caution! Not to be used for adding bumps at arbitrary positions in the data.

Summary


Fitting
Consistency
Estimation
Efficiency
Bias
Likelihood
Invariance
MVB
ML Estimation
Orthogonal Polynomials
Least Squares
Fitting techniques comparison
X²
Goodness of fit
Comparing families
Wilks' Theorem