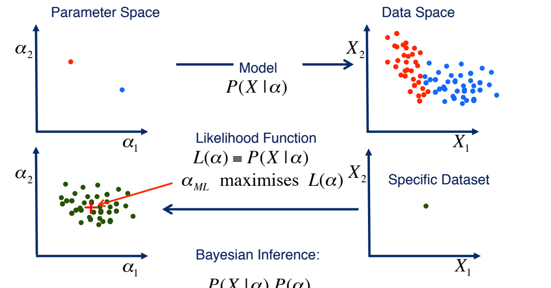
Review: Max Likelihood and Bayesian Inference



$$P(\alpha \mid X) = \frac{P(X \mid \alpha)\, P(\alpha)}{\int P(X \mid \alpha)\, P(\alpha)\, d\alpha} \propto L(\alpha)\, P(\alpha)$$

Posterior Probability

Likelihood modifies the Prior.

Monte-Carlo Error Propagation

1. Create **mock datasets**.

1a. "Jiggle" the data points (using Gaussian random numbers).
* Requires good error bars.

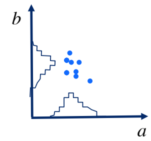
1b. (and/or) "**Bootstrap**" samples:

Pick *N* data points at random, with replacement (some points omitted, some repeated).
* Requires more data than parameters ($N > M$).
* Works with no error bars available.

2. Fit the model to each mock dataset.

$$\langle X_i \rangle = a\, t_i + b$$

3. Observe how the best-fit parameter values "dance".



4. Accumulate histograms approximating the parameter probability distributions.

5. Compute mean, median, variance, etc. of the parameters, or **any function of the parameters**.

Confidence interval on a single parameter

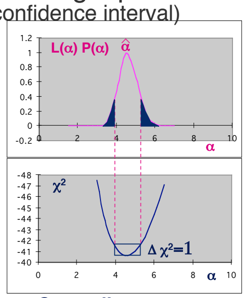
(1-parameter, k-sigma confidence interval)

The **1-σ confidence interval** on α includes 68% of the area under the likelihood function:

$$L(\alpha) \equiv P(X \mid \alpha) \propto \frac{e^{-\chi^2/2}}{\prod_i \sigma_i}$$

or posterior probability distribution, for non-uniform prior $P(\alpha)$:

$$P(\alpha \mid X) \propto L(\alpha)\, P(\alpha)$$



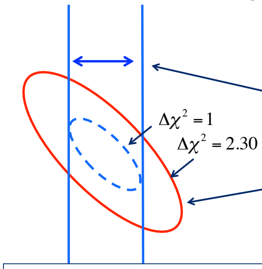
For a k-σ (1-parameter) confidence interval, use $\Delta\chi^2 = k^2$,

$\Delta\chi^2 = 1$ for 1-σ, 68% probability
$\Delta\chi^2 = 4$ for 2-σ, 95.4% probability
$\Delta\chi^2 = 9$ for 3-σ, 99.73% probability …

Generalise:

χ^2 => *-2 ln(L(α) P(α))*

2-parameter 1-sigma Confidence Region



If *Y* is a "**nuisance parameter**", use the **1-parameter 1-sigma confidence interval** in *X*, *tangent to* the $\Delta\chi^2 = 1$ contour in *(X,Y)*.
This interval encloses 68% probability.

If both *X* and *Y* are of interest, use the **2-parameter 1-sigma confidence region**, the $\Delta\chi^2 = 2.30$ contour in *(X,Y)*.
This contour encloses 68% probability.

Use *-2 ln(L(α) P(α))* instead of χ^2, if needed.

Note: Contour enclosing 68% probability must be **wider** than the 1-parameter confidence interval.

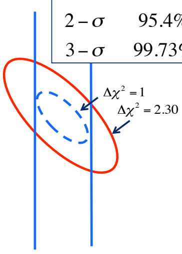
Why?

$$L(\alpha) \equiv P(X \mid \alpha) \propto \frac{e^{-\chi^2/2}}{\prod_i \sigma_i}$$

M - parameter *k* - σ Confidence Regions

$\Delta\chi^2$ thresholds for *M*-parameter *k*-σ Confidence Regions

	Prob	$M=1$	2	3	4
$1-\sigma$	68%	1	2.30	3.53	4.72
$2-\sigma$	95.4%	4	6.17	8.02	9.70
$3-\sigma$	99.73%	9	11.8	14.2	16.3



The ***M*-parameter confidence region** is enclosed by the $\Delta\chi^2$ surface including the desired probability.

All **nuisance parameters must be re-fitted** (or integrated over) for each set of fixed values for the *M* parameters in the sub-space of interest.

The $\Delta\chi^2$ in the *M*-parameter sub-space has a χ^2_M distribution, with *M* degrees of freedom.

Example: Estimate both μ and σ

$$L(\mu,\sigma) \equiv P(X \mid \mu,\sigma) = \frac{e^{-\chi^2/2}}{(2\pi)^{N/2}\,\sigma^N}$$

$$-2\ln L = \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma}\right)^2 + 2N \ln\sigma + \text{const}$$

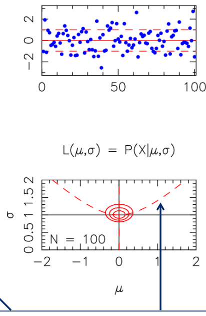
$$0 = \frac{\partial}{\partial\mu}\left[-2\ln L\right] = -2\sum_{i=1}^{N} \frac{X_i - \mu}{\sigma^2}$$

$$0 = \frac{\partial}{\partial\sigma}\left[-2\ln L\right] = -2\sum_{i=1}^{N} \frac{(X_i - \mu)^2}{\sigma^3} + \frac{2N}{\sigma}$$

$$\mu_{ML} = \frac{1}{N}\sum_i X_i \qquad \sigma^2_{ML} = \frac{1}{N}\sum_i (X_i - \mu_{ML})^2$$

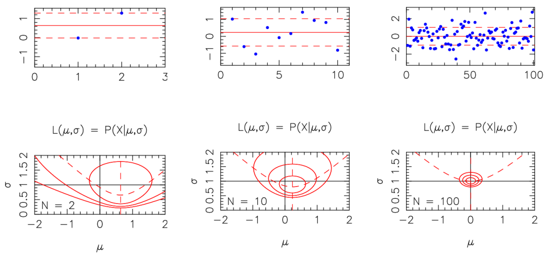
Posterior $\propto$ Likelihood $\times$ Prior

$$P(\mu,\sigma \mid X) \propto L(\mu,\sigma)\, P(\mu,\sigma)$$



Note: ML gives biased estimate for σ.

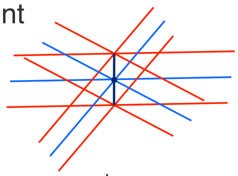
Example: Estimate both μ and σ



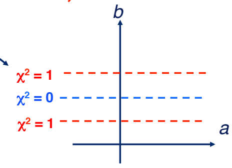
Contours: 1,2,3-sigma 2-parameter confidence regions for μ and σ.
Dashed curves: maximum-likelihood estimates for μ_{ML} and σ_{ML}.
True values: $\mu = 0$ and $\sigma = 1$.

Fit a line to *N=1* data point

Fit $y = a\,x + b$ to *N=1* data point:
- Blue lines : $\chi^2 = 0$
- Red lines : $\chi^2 = 1$



χ^2 contours in the *(a,b)* plane:

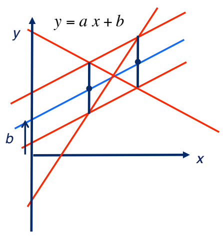


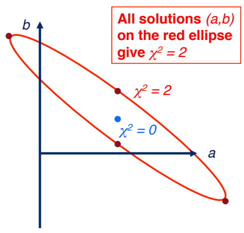
Solution is **degenerate**, since *M=2* parameters are constrained by only *N=1* data point.

Bayes: prior *P(a,b)* needed to determine a unique solution.

Fit a line to *N = 2* data points

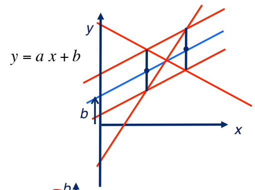
- Fit $y = a\,x + b$ to $N = 2$ data points:
  - red lines give $\chi^2 = 2$
  - blue line gives $\chi^2 = 0$
- Note that *a, b* are not independent.

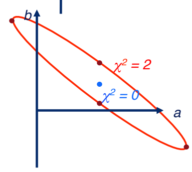




Correlated Parameters ☹

- **Parameters *a* and *b* are correlated :(**
- To find the optimal (*a, b*) we must:
  - minimize χ^2 with respect to *a* at a sequence of fixed *b* values
  - then minimise the resulting χ^2 values with respect to *b*.
- If *a* and *b* were independent, then all slices through the χ^2 surface at each fixed *b* would have same shape and minimum.
- Similarly for *a*.
- We could then optimize *a* and *b* independently, saving a lot of calculation
- **How to make *a* and *b* independent of each other**?



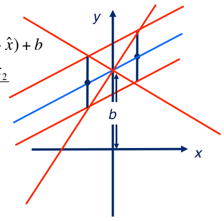


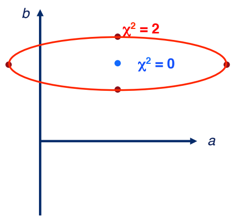
Orthogonal Parameters for fitting a line to *N = 2* data points

- Fit $y = a(\,x - \hat{x}\,) + b$ Different parameters for same model.
- Note: *a, b* are now independent ! 🙂

$y = a\,(x - \hat{x}) + b$

$\hat{x} = \frac{x_1 + x_2}{2}$



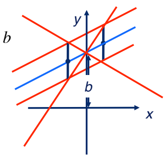


Orthogonal slope and intercept

Analysis using the algebra of random variables:

$y = a\,(x - \hat{x}) + b$

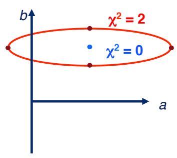
$\hat{x} = \frac{x_1 + x_2}{2}$



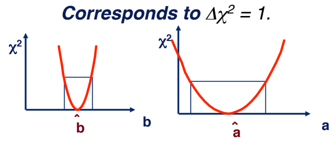
$\hat{b} = \hat{y} = \frac{y_1 + y_2}{2}$ $\qquad \hat{a} = \frac{y_2 - y_1}{(x_2 - x_1)}$

$\sigma^2(\hat{b}) = \frac{2\,\sigma^2}{4}$ $\qquad \sigma^2(\hat{a}) = \frac{2\,\sigma^2}{(x_2 - x_1)^2}$

$\sigma(\hat{b}) = \frac{\sigma}{\sqrt{2}}$ $\qquad \sigma(\hat{a}) = \sqrt{2}\,\frac{\sigma}{(x_2 - x_1)}$

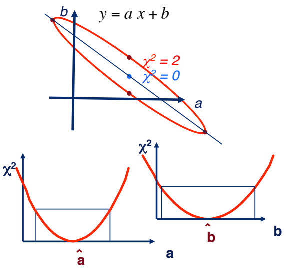


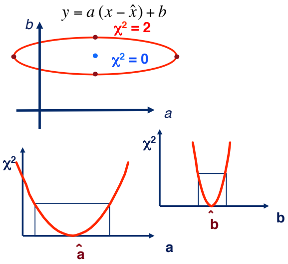
Corresponds to $\Delta\chi^2 = 1$.



Orthogonal vs Correlated Parameters

Correlated Parameters ☹ **Orthogonal Parameters** ☺

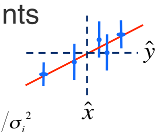




For each *a*, a different *b* minimises χ^2.
For each *b*, a different *a* minimises χ^2.

For any *a*, the same *b* minimises χ^2.
For any *b*, the same *a* minimises χ^2.

Fit a line to *N* data points



- If we use $y = a\,x + b$
  then *a, b* are correlated.
- Make *a, b* orthogonal:
  $y = a\,(x - \hat{x}) + b$ $\quad \hat{x} = \dfrac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$
- Intercept: Set *a = 0* and optimise *b*:
  optimal average: $\hat{b} = \hat{y} = \dfrac{\sum y_i/\sigma_i^2}{\sum 1/\sigma_i^2}, \quad \text{Var}\left[\hat{b}\right] = \dfrac{1}{\sum 1/\sigma_i^2}$
- Slope: Set *b = 0* and optimise *a*:
  optimal scaling of pattern: $P_i = x_i - \hat{x}$

$$\hat{a} = \frac{\sum y_i\,(x_i - \hat{x})/\sigma_i^2}{\sum (x_i - \hat{x})^2/\sigma_i^2}, \quad \text{Var}\left[\hat{a}\right] = \frac{1}{\sum (x_i - \hat{x})^2/\sigma_i^2}$$

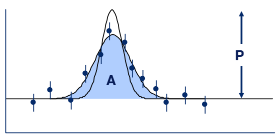
Choose Orthogonal Parameters

- Good practice (when possible).
- Results for any one parameter don't depend on values of other parameters.
- Example: fit a gaussian profile. 2 fit parameters:
  - Width, w
  - Area or peak value. Which is best?

Peak value depends on width – bad

$$f(x) = P e^{-\frac{1}{2}\left(\frac{x - x_0}{w}\right)^2}$$

$$g(x) = \frac{A}{w\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x - x_0}{w}\right)^2}$$



Area is (more nearly) independent of width – good