Lecture 6

- Fitting a line to data

$$\text{Data} = y_i$$
$$\text{Signal Model} = s_i = ax_i + b$$
$$\text{Noise} = r_i \quad (\text{or } ``n_i" \text{ -- doesn't matter})$$

$$\boxed{\text{Data} = \text{Signal} + \text{Noise}}$$

⇒ what makes this a probabilistic problem?
↳ noise!

$$p(r_i) = N(0, \sigma)$$
$$\therefore \quad p(y_i - s_i) = N(0, \sigma)$$
$$p(y_i) = N(s_i, \sigma)$$

$$\Rightarrow \quad p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - s_i)^2}{\sigma^2}}$$

BUT s_i is now a LINE RELATIONSHIP not just the single "μ" as in the last lecture.

$$\ln \mathcal{L} = \ln\left[\prod_i p(y_i)\right]$$

$$= \text{constant} - \frac{1}{2}\underbrace{\sum_i \frac{(y_i - ax_i - b)^2}{\sigma^2}}_{\chi^2}$$

$\boxed{\text{NOTE}}$ Also possible to modify the likelihood to reduce the impact of outliers (see the notebook).

- Goodness of Fit
  - MLE gives best-fit parameters of model.
  - How do we know if the model is any good?

$\Rightarrow$ Max-likelihood should not be an unlikely occurrence ... need to know $\mathcal{L}$ distribution.

$$z_i = \frac{(y_i - s_i)}{\sigma_i}$$

$$\ln \mathcal{L} = \text{constant} - \frac{1}{2}\sum_i z_i^2$$

$$= \text{constant} - \frac{1}{2}\chi^2$$

chi-squared

data-points

params

Mean of χ^2 distribution $= N - k$

Std of χ^2 distribution $= \sqrt{2(N-k)}$

χ^2 per degree of freedom

$$\chi^2_{dof} = \frac{1}{N-k} \sum_i z_i^2 \quad \Big\} \text{ GOOD FIT } = \chi^2_{dof} \simeq 1.$$

- Model Comparison — much more on this later in the course.

$$AIC_M = \underbrace{-2\ln[\mathcal{L}^{Max}(M)]}_{\propto \chi^2} + \underbrace{2k + \frac{2k(k+1)}{N-k-1}}_{\text{penalty on model complexity.}}$$

$\Rightarrow$ Model with lowest AIC wins!

- Confidence Estimation — Fisher matrix approach is an approximation

$\Rightarrow$ simulate $\boxed{\text{NEW}}$ data by resampling from the data you have ... NOT MAGIC!!

BOOTSTRAP $\Rightarrow$ resample data (with replacement) and compute statistics on each new data realization ... EASY.

JACKKNIFE $\Rightarrow$ remove one data point at a time, and re-compute statistics

$\hookrightarrow$ gives $(N-1)$ new data realizations ... EASY.