Statistical techniques for data analysis in Cosmology

arXiv:0712.3028; arXiv:0911.3105
Numerical recipes (the “bible”)

Licia Verde

ICREA & ICC UB-IEEC

http://icc.ub.edu/~liciaverde

outline

- Lecture 1: Introduction Bayes vs Frequentists, priors, the importance of being Gaussian, modeling and statistical inference, some useful tools.
- Lecture 2: Monte Carlo methods. Different type of errors. Going beyond parameter fitting.
- Lecture 3: forecasting: Fisher matrix approach. Introduction to model selection.

Conclusions.

There is a BIG difference between χ^2 & reduced χ^2

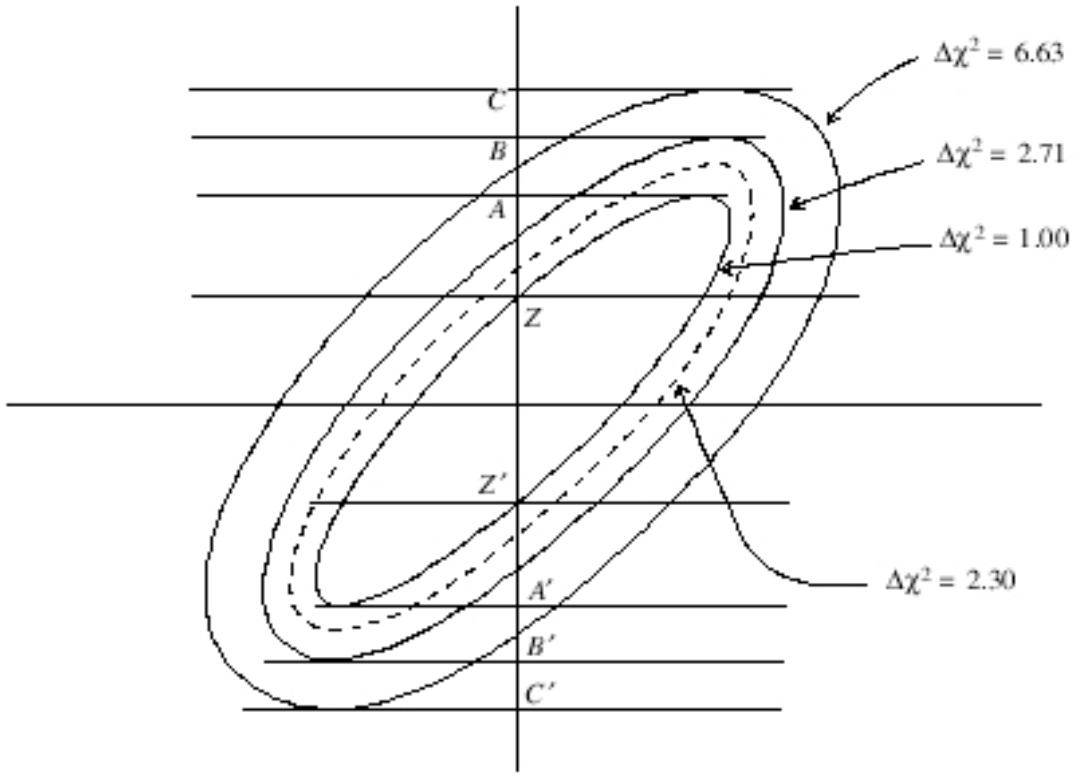
$\Delta\chi^2$ as a Function of Confidence Level and Number of parameters						
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

Only for multivariate Gaussian with constant covariance

$$-2\ln L = \chi^2$$

From: “Numerical recipes” Ch. 15

If likelihood is Gaussian and Covariance is constant



Example: for multi-variate Gaussian

example

Cash 1979

Observation of N clusters is Poisson

$$\mathcal{P} = \Pi_{i=1}^{N}[e_i^{n_i}\exp(-e_i)/n_i!]$$

n_i is the number of clusters observed in the $i-th$ experimental bin

$e_i = I(x)\delta x_i$ ..... expected .....

Experimental bin (mass, SZ decrement, X-ray lum, z…)

Define $C \equiv -2\ln\mathcal{P} = 2(E - \sum_{i=1}^{N}\ln I_i)$

Unbinned or small bins
occupancy 1 or 0 only

E is the total expected number of clusters in a given model

ΔC Between 2 different models is chisquared-distributed!

Marginalization

$$P(\alpha_1..\alpha_j|D) = \int d\alpha_{j+1}, ...d\alpha_m P(\vec{\alpha}|D)$$

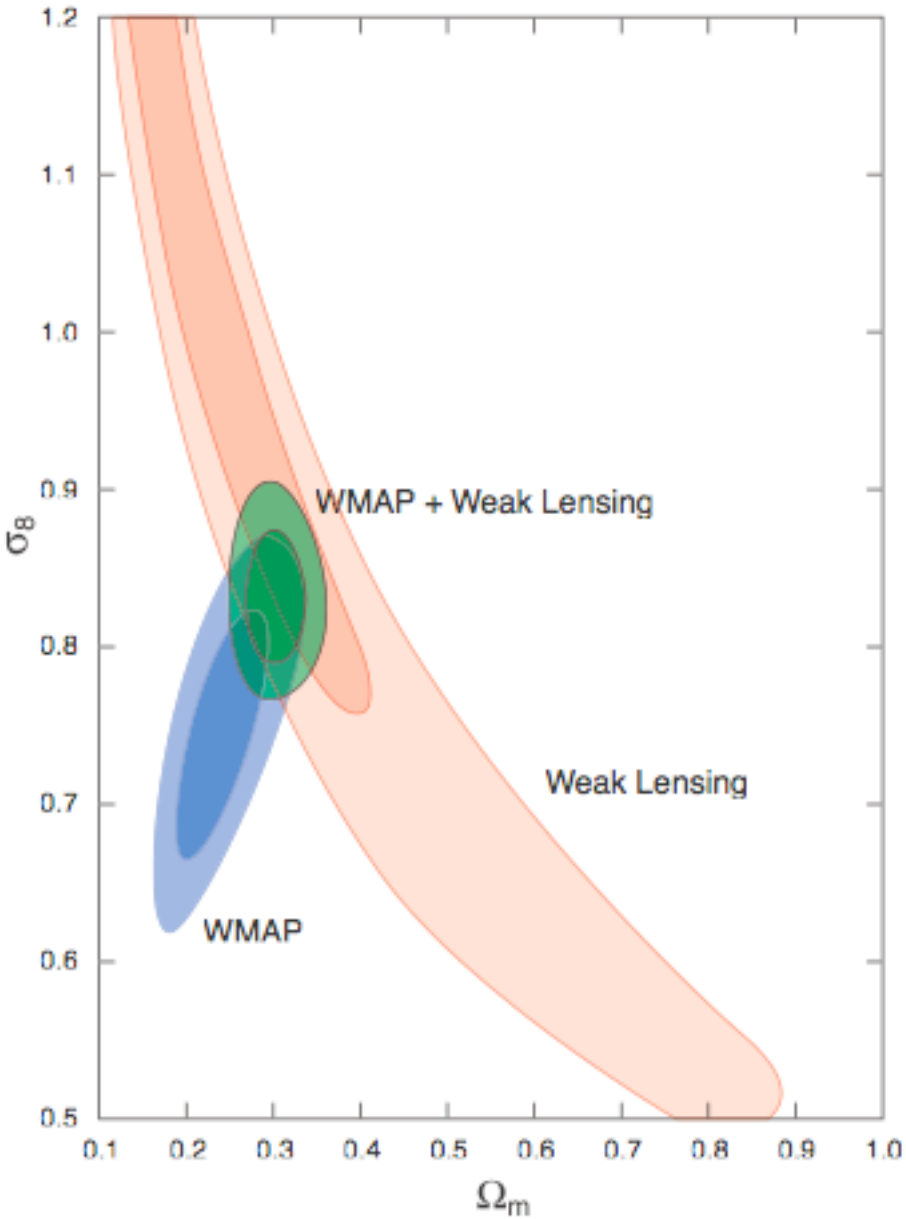
example

Other data sets

If independent, multiply the two likelihoods

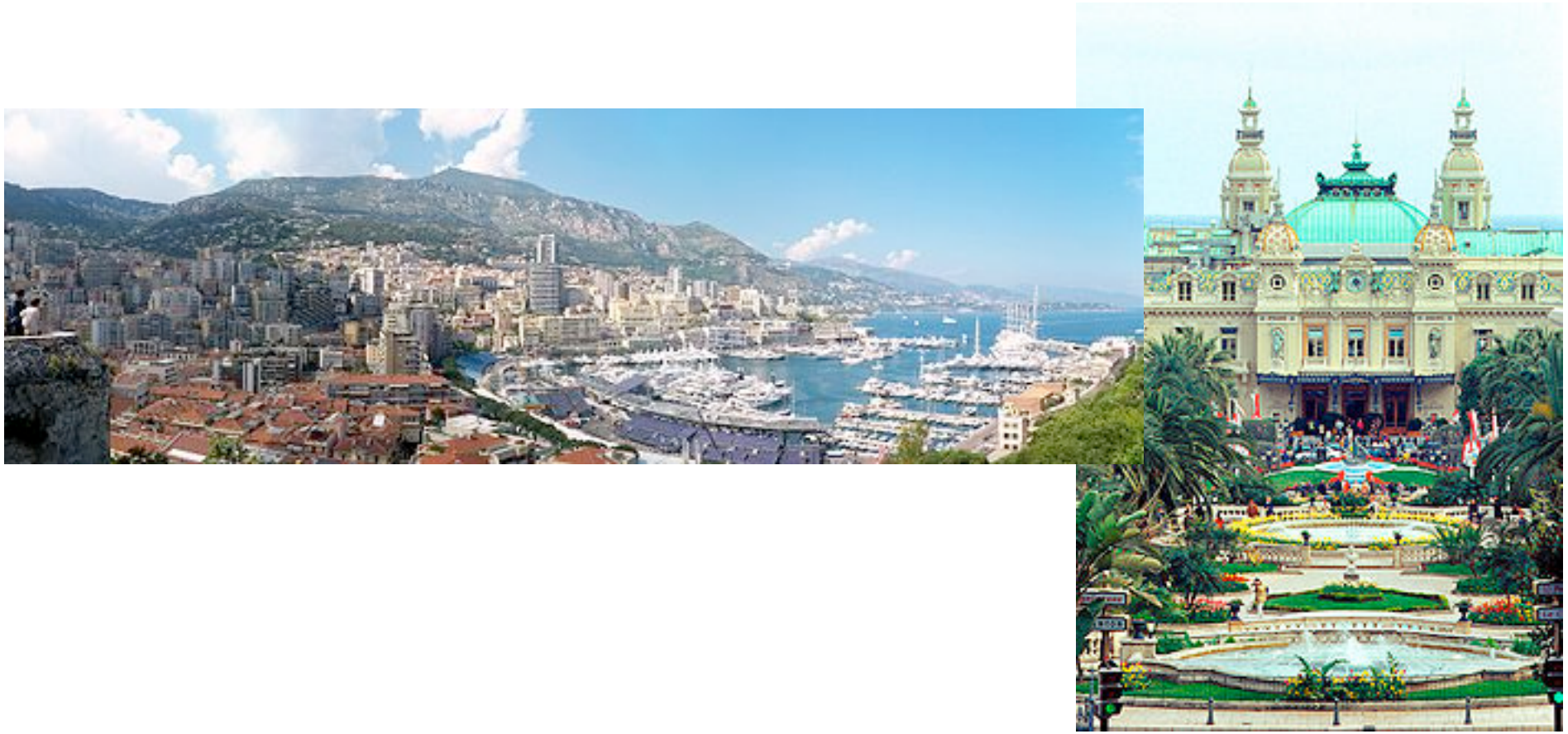
(can use some of them as “priors”)

Beware of inconsistent experiments!



Spergel 2007

Monte Carlo methods



Monte Carlo methods

a) Monte Carlo error estimation

b) Monte Carlo Markov Chains

Your brain does it!

Spot the differences…

Intro to:
Monte Carlo

Simple problem: what's the mean of a large number of objects?

What's the mean height of people in PV?

$$\sum_{i=1}^{N} \frac{h_i}{N}$$

If N is very large this is untractable soo…

If n<<N but still a fair sample, great!

$$\sim \sum_{i=1}^{n} \frac{h_i}{n}$$

In probability:

$$\int f(x)P(x)dx \sim \frac{1}{S}\sum^{S} f(x^s) \qquad \text{if } x^s \sim P(x)$$

In Bayesian inference:

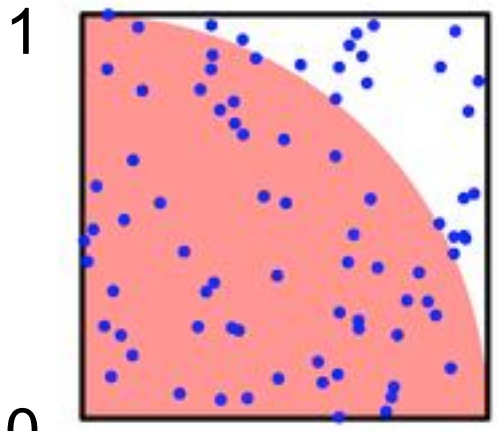
$$p(x|D) = \int P(x|\theta, D)P(\theta|D)d\theta \sim \frac{1}{S}\sum P(x|\theta^s, D) \quad \text{if } \theta^s \sim P(\theta|D)$$

You can show that:

The estimator is unbiased
and you can quantify the variance of the estimator:
The error shrinks like $S^{1/2}$

Very simple example:

A dumb approximation of π



4 times the red area

$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\, \mathrm{d}x\, \mathrm{d}y$$

```
octave:1> S=12; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.3333
octave:2> S=1e7; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.1418
```

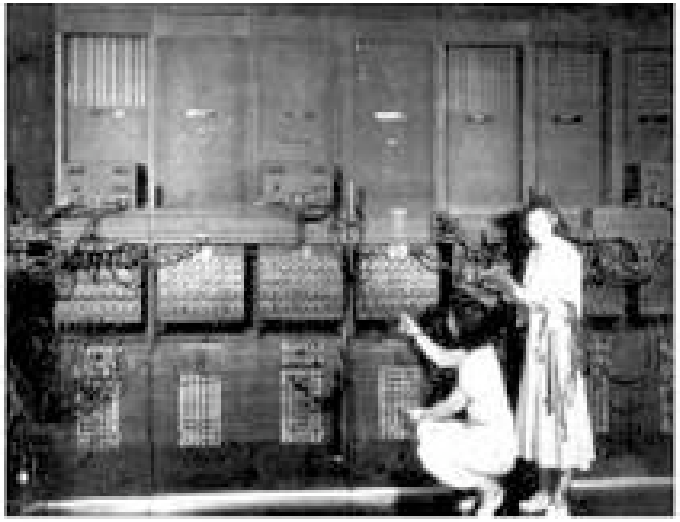
There are better ways to compute π, so use mcmc only when right to use…

Historical note

Enrico Fermi (1901–1954) took great delight in astonishing his colleagues with his remakably accurate predictions of experimental results. . . he revealed that his "guesses" were really derived from the statistical sampling techniques that he used to calculate with whenever insomnia struck in the wee morning hours!

—*The beginning of the Monte Carlo method*,
N. Metropolis

history



Monte Carlo methods

a) Monte Carlo error estimation

Back to parameter estimation and confidence regions

Conceptual interpretation in cosmology

α_{true} Set of parameters known only to Mother Nature

Statistically realized

α Observable universe

want

Measurement (with its errors)

Do Measured data $\xrightarrow{\text{analysis}}$ α_o

You (the experimenter)
Can see

NOT a unique realization of α_{true}

There could be infinitely many realizations (hypothetical data sets) $D_1, D_2, \ldots.$

Each one with best fit parameters $\alpha_1, \alpha_2, \ldots.$

Expect: $< \alpha_i >= \alpha_{\rm true}$

If I knew the distribution of $\alpha_i - \alpha_{\rm true}$ That'd be all I need

Trick: say that (hope) $\alpha_0 \sim \alpha_{\rm true}$

In many cases we can simulate the distribution of $\alpha_i - \alpha_0$

Make many synthetic realizations of universes where α_0 is the truth; mimic the observational process in all these mock universes, estimate the best fit parameters from each; map $\alpha_S - \alpha_0$

Very important tool

How to sample from the probability distribution?

- For some well known univariate probability distributions there are numerical routines
  http://cg.scs.carleton.ca/~luc/rnbookindex.html
- In other cases there may be numerical techniques to sample P(x) [more later]
- Importance sampling: (if you know how to sample from Q but not from P)

$$\int f(x)P(x)dx = \int f(x)\frac{P(x)}{Q(x)}Q(x)dx \sim \frac{1}{S}\sum_{s=1}^{S} f(x^s)\frac{P(x^s)}{Q(x^s)} \text{ if } x^s \sim Q(x)$$

Some Q are more suitable for P than others….

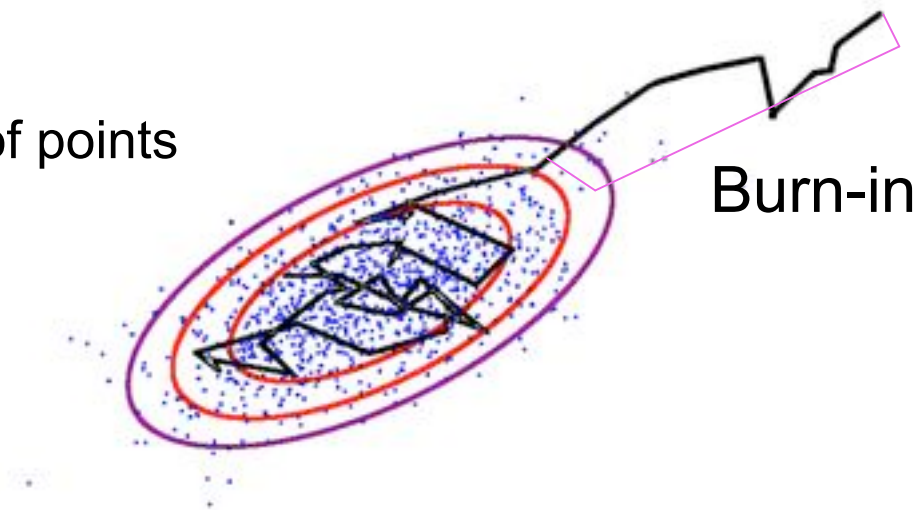
Monte Carlo Markov Chains

So you have a higher-dimensional probability distribution,
you want to sample in a way proportional to it ,
with a random walk

Start at an
arbitrary point

Take Markov steps

Goal: density of points
proportional to
the probability

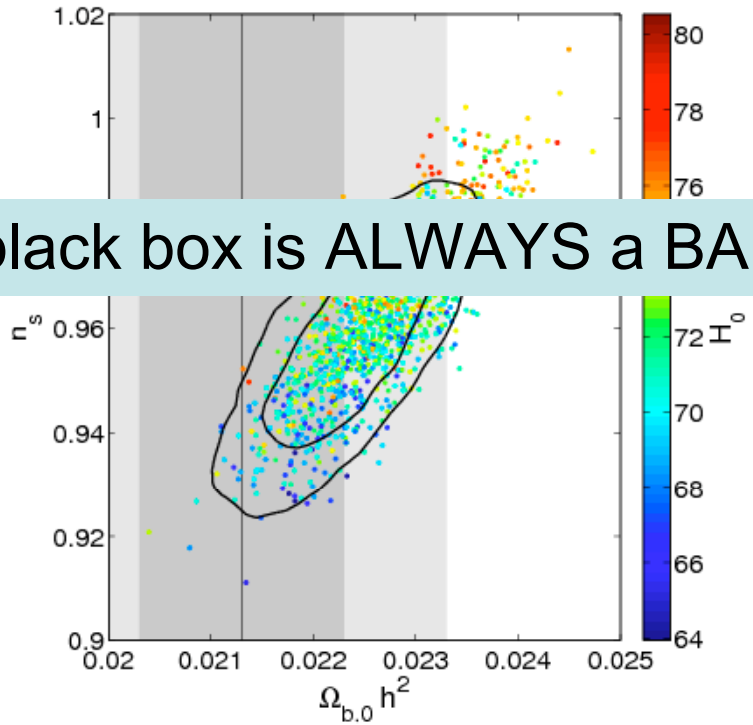


MCMC gives approximated, correlated samples from the target distribution

b) Monte Carlo Markov Chains

http://cosmologist.info/cosmomc/

Cosmological MonteCarlo



Using software as black box is ALWAYS a BAD idea

Samples from WMAP 5-yr likelihood combined with deuterium constraint (0805.0594)

Get help: Search Google™ Custom Search

NEW: *(May 08)* Support for UNION supernovae, equal-likelihood limits, WMAP5-format chains, more confidence limits
(Mar/Apr 08) Support for WMAP5, CMB SZ templates, new reionization model
(Feb 08) Latest ACBAR data, CAMB update, option to use as a generic sampler

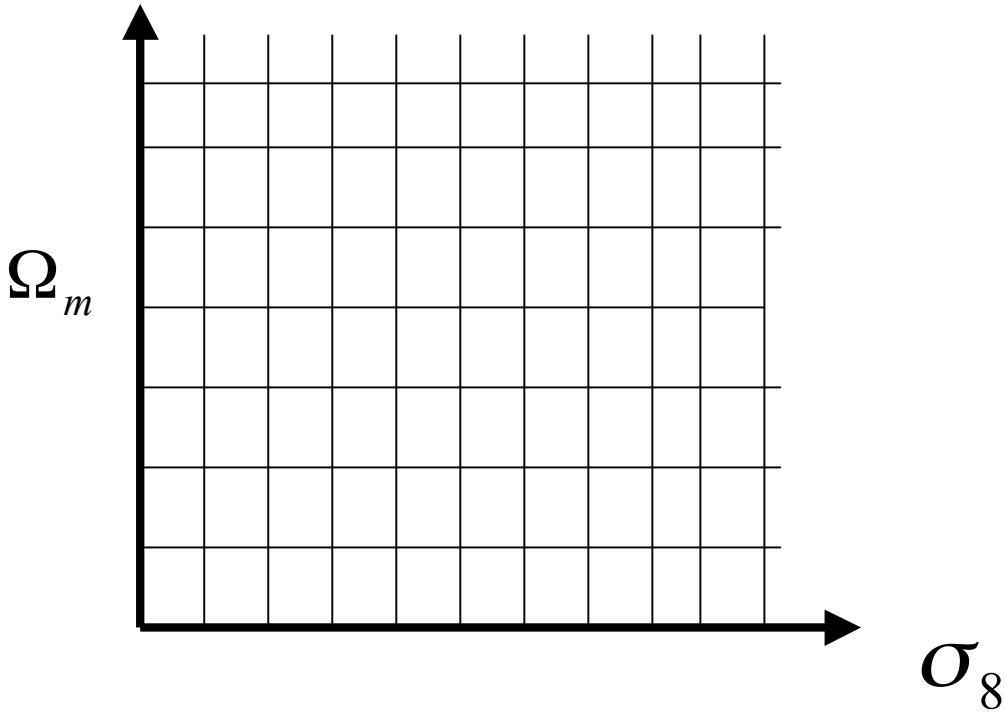
See the ReadMe file for program documentation and download. Also the CosmoloGUI documentation.

b) Monte Carlo Markov Chains

Explore likelihood surface

Grid-based approach

Operationally:



e.g., 2 params: 10 x 10

What if you have (say) 6 parameters?

6 params, 20 pixels/dim
= 6.7×10^7 evals

say 1.6 s/eval

~1200 days!

You've got a problem !

Markov Chain Monte Carlo (MCMC)

Standard in CMB analyses (publicly available COSMOMC)

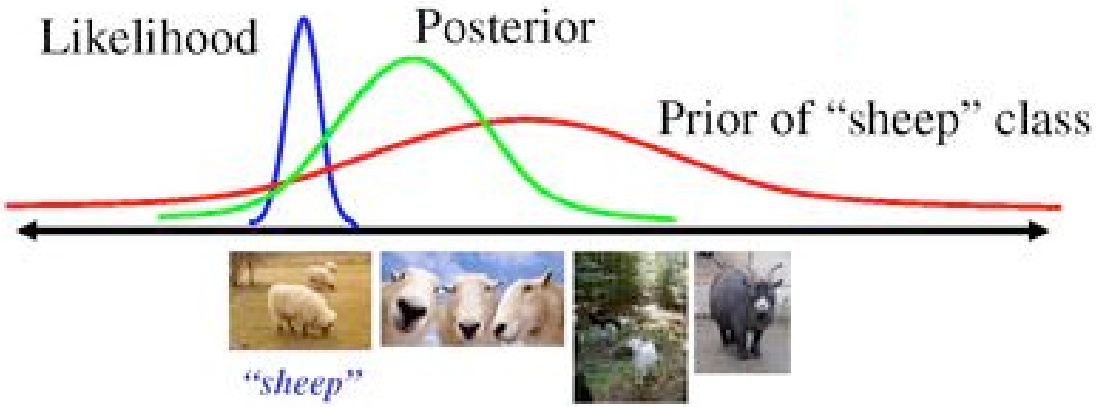
Simulat

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

cosm. params

$$C_\ell^{\mathrm{th}}$$

Bayes



Genera

are

a fair sample of the likelihood surface

Markov Chain Monte Carlo (MCMC)

Random walk in parameter space

At each step, sample one point in parameter space

The density of sampled points $\propto$ <u>posterior distribution</u>

FAST: before 10^7 likelihood evaluations, now< 10^5

marginalization is easy:
just project points and recompute their density

Adding external data sets is often very easy

Operationally (Metropolis-Hastings):

1. Start at a random location in parameter space: α_i^{old} L^{old}

2. Try to take a random step in parameter space: α_i^{new} L^{new}

3a. If $L^{new} \geq L^{old}$ Accept (take and save) the step, "new"--> "old" and go to 2.

3b. If $L^{new} < L^{old}$ Draw a random number x uniform in 0,1

If $x \geq \frac{L^{new}}{L^{old}}$ do not take the step (i.e. save "old") and go to 2.

If $x < \frac{L^{new}}{L^{old}}$ do as in 3a.

KEEP GOING….

"Take a random step"

The probability distribution of the step is the "**proposal distribution**", which you should not change once the chain has started.

The proposal distribution (the step-size) is crucial to the MCMC efficiency.

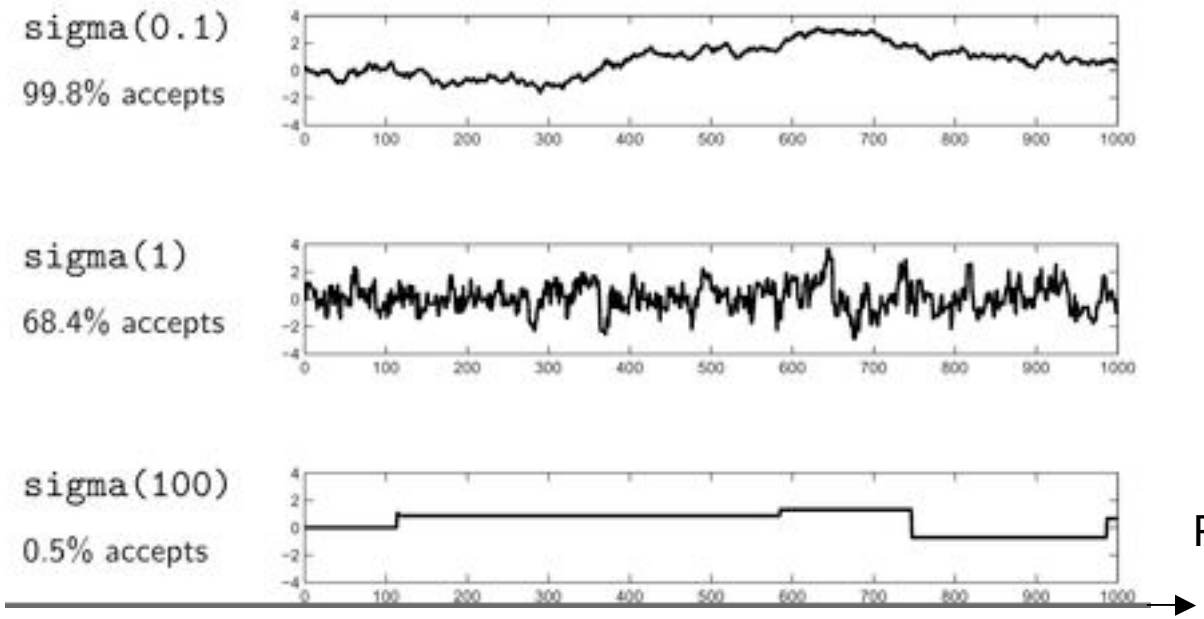
Steps too small step poor mixing

Steps too big step poor acceptance rate

"fair sample of the likelihood surface", remember?

The importance of stepsize

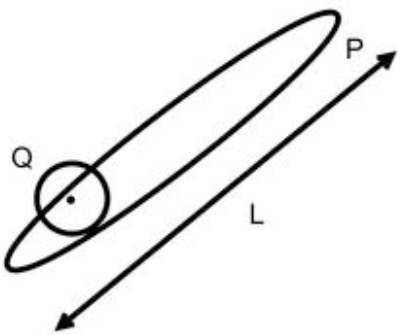
Likelihood



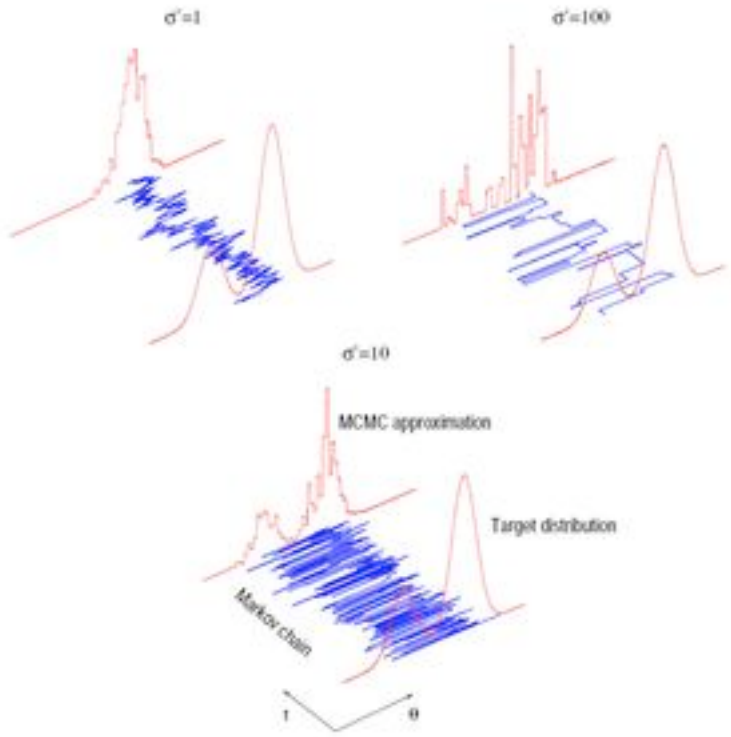
Poor exploration

Poor exploration

Step number



The importance of stepsize

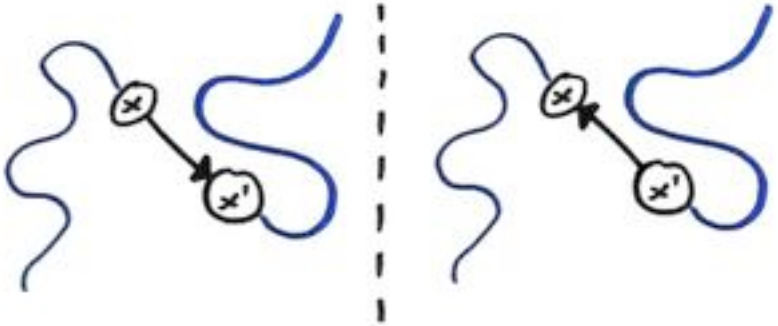


Take a random step

For statisticians: transition operators

Detailed balance: (beware of boundaries….)

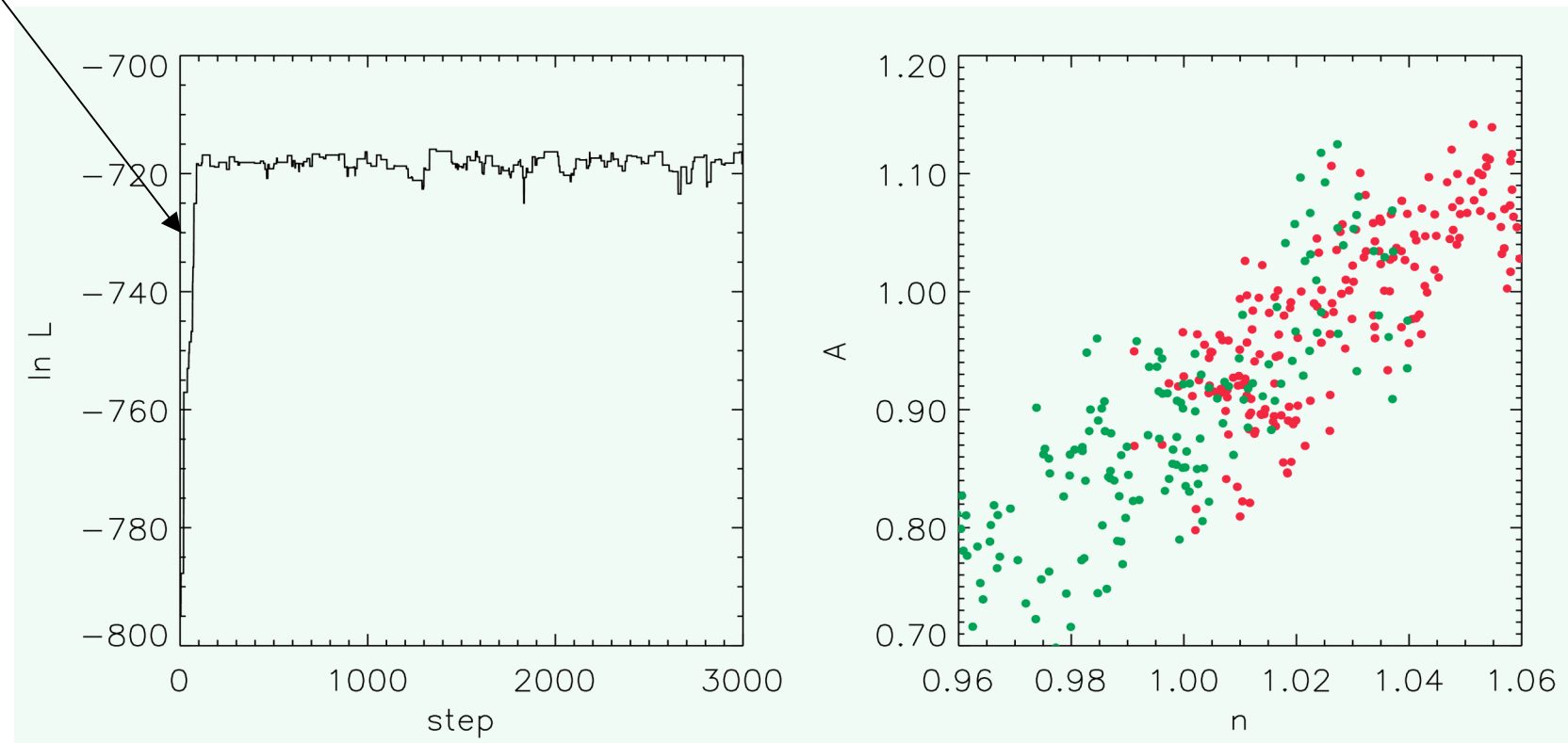
Detailed balance means $\rightarrow x \rightarrow x'$ and $\rightarrow x' \rightarrow x$ are equally probable:



When the MCMC has forgotten about the starting location and has well explored the parameter space you're ready to do parameter estimation.

USE a MIXING and CONVERGENCE criterion!!!

Burn-in



Gelmans and Rubin convergence

Recommended: start 4 to 8 chains at well separated points

M chains, N elements

Chain mean $$\bar{y}^j = \frac{1}{N}\sum_{i=1}^{N} y_i^j,$$

Vector with parameters value

Mean of distrib. $$\bar{y} = \frac{1}{NM}\sum_{ij=1}^{NM} y_i^j\,.$$

Variance between chains $$B_n = \frac{1}{M-1}\sum_{j=1}^{M}(\bar{y}^j - \bar{y})^2$$

And within $$W = \frac{1}{M(N-1)}\sum_{ij}(y_i^j - \bar{y}^j)^2$$

$$\boxed{\hat{R} = \frac{\frac{N-1}{N}W + B_n\left(1+\frac{1}{M}\right)}{W}}$$

Always >1 by construction

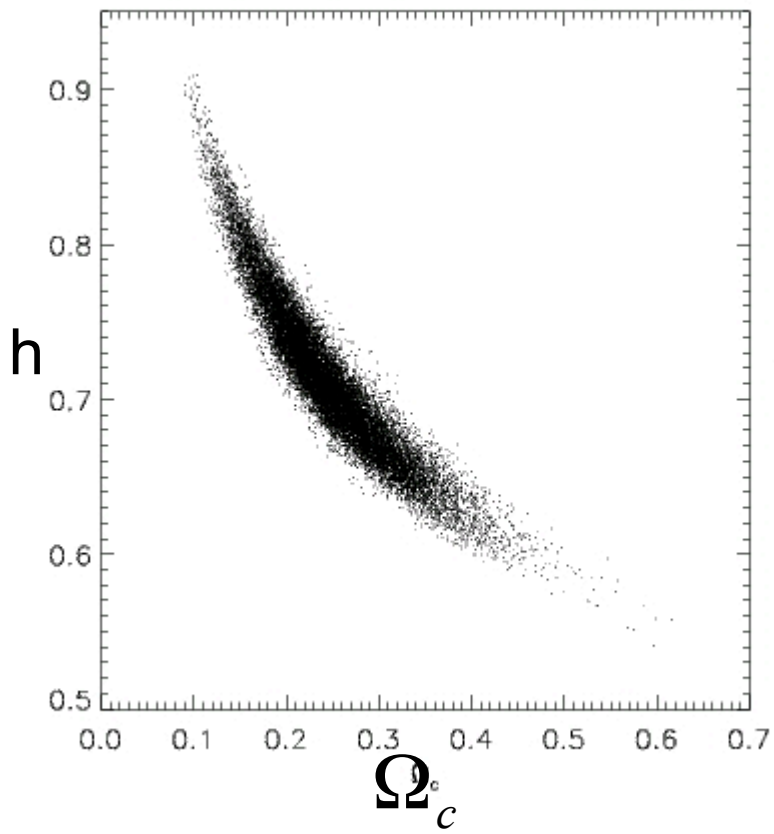
Require <1.03

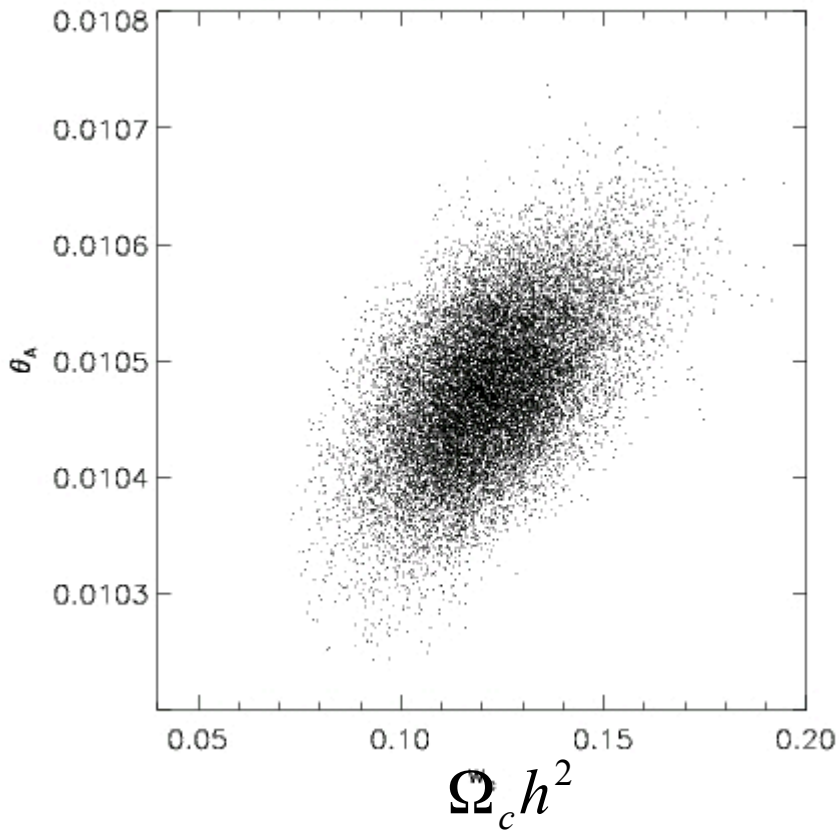
Unconverged chains are just nonsense

Metropolis-Hastings is NOT the only implementation,

Other options are:
Gibbs Sampler
Rejection method
Hamiltonian Monte-Carlo
Simulated annealing (though you do not get an MCMC)

Beware of DEGENERACIES





Reparameterization. e.g., Kososwsky et al. 2002

$$\theta_A = \frac{r_s(a_{dec})}{D_A(a_{dec})}$$

Even “better”:

Cosmomc has the option of computing the covariance for the parameters
Find the axis of the multi dim. degeneracies
perform a rotation and re-scaling to obtain
azimutally symmetric contours

An improve MCMC efficiency by factor of up to 10

It is still a linear operation

Where’s the prior ?

$$D_{KL} \equiv \int p(\Theta|D) \ln \frac{p(\Theta|D)}{p(\Theta)} d\Theta.$$

Once you have the MCMC output:

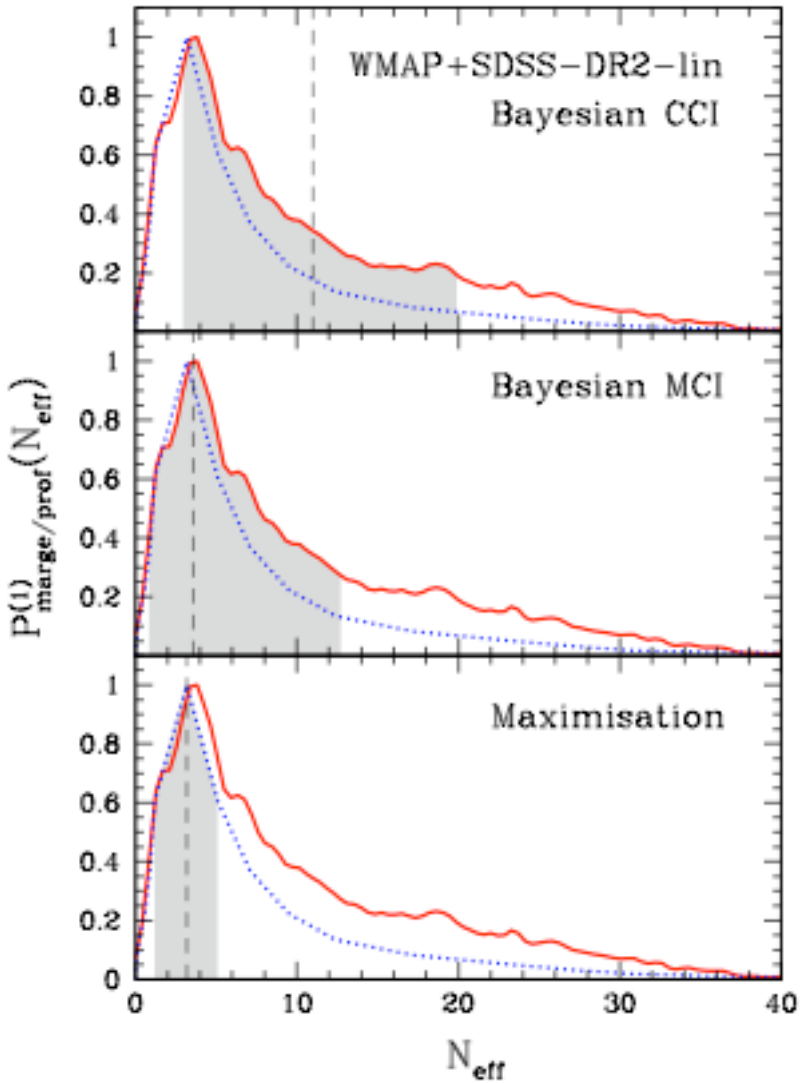
- The density of points in parameter space gives you the posterior distribution
- To obtain the marginalized distribution, just project the points
- To obtain confidence intervals, - integrate the “likelihood” surface

  -compute where e.g. 68.3% of points lie
- To each point in parameter space sampled by the MCMC give a weight proportional to the number of times it was saved in the chain
- To add to the analysis another dataset (that does not require extra parameters) renormalize the weight by the “likelihood” of the new data set.

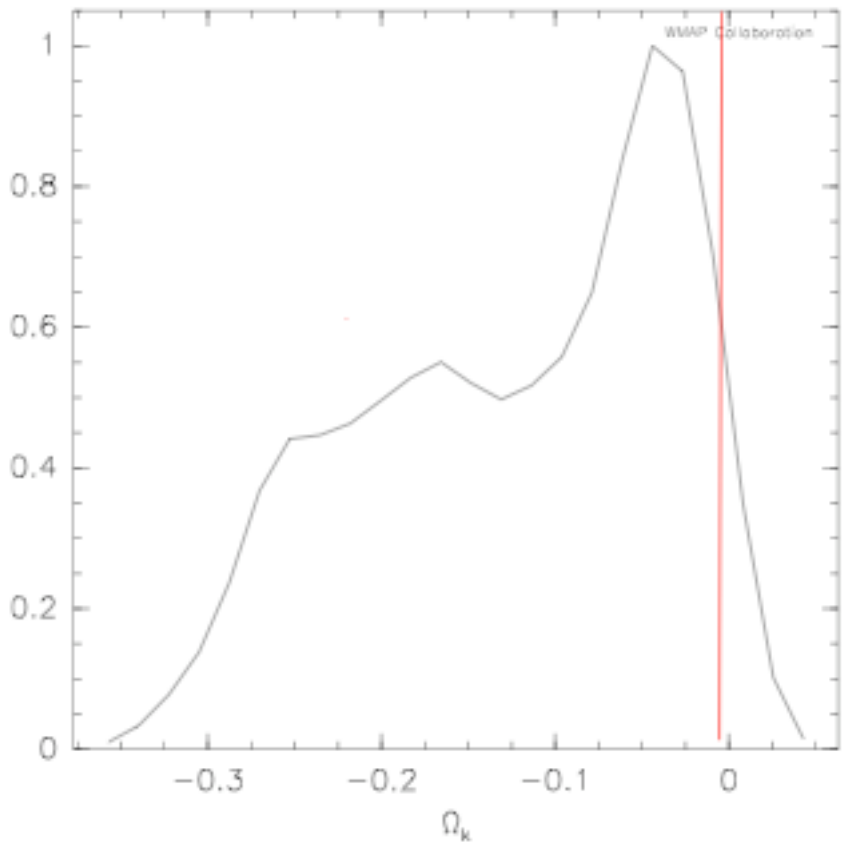
  No need to re-run!

warning: if new data set is not consistent with the old one--> nonsense

Errors, what errors?



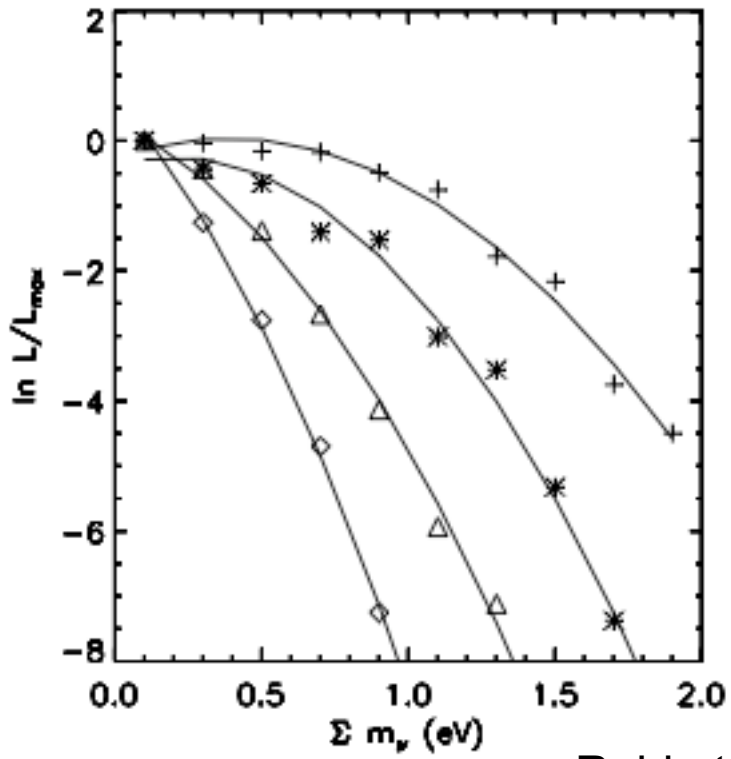
Hamann et al. arXiv:0705.0440


WMAP Collaboration
1
0.8
0.6
0.4
0.2
0
−0.3
−0.2
−0.1
0
Ω_k

Prior-independence?

Once you and an MCMC output what you can do is to look at the likelihood value not the weight.

Say you have n uninteresting parameters and one that you are interested in e.g. mν. For each value of mν find the maximum likelihood L_m regardless of the values assumed by the other parameters. Then consider L_m/L_{max} as a function of mν.



Reid et al '10

Figure 4. The profile likelihood defined in Section 2.3 in bins of $\Delta(\sum m_\nu) = 0.2$ eV for the ΛCDM model with WMAP5 only (crosses), WMAP5+maxBCG (stars), WMAP5+H_0 (triangles), and WMAP5+maxBCG+H_0 (diamonds). The black curves overlay a quadratic fit to these points, illustrating that a Gaussian curve provides a good fit to this one-dimensional distribution.

You can do it yourself! http://lambda.gsfc.nasa.gov

In particular:

http://lambda.gsfc.nasa.gov/product/map/dr4/parameters.cfm

National Aeronautics and Space Administration

RSS LAMBDA News

HOME | PRODUCTS | TOOLBOX | LINKS | NEWS | SITE INFO

LEGACY ARCHIVE FOR MICROWAVE BACKGROUND DATA ANALYSIS

"One Stop Shopping for CMB Researchers"

Data Products

- Mission Data
  - WMAP
    - Overview
    - Products
    - Documents
    - Software
    - Images
    - Education
  - COBE
  - Relikt
  - IRAS
  - SWAS
- CMB Related Data
  - Space Missions
  - Suborbital CMB
  - Foreground
  - LSS Links

WMAP Cosmological Parameters Model/Dataset Matrix

		WMAP7+									
Model [all are +SZ+LENS]	WMAP7	WMAP7.1	BAO+ H0	BAO+ SNSALT	SNCONST	BAO+ SNCONST	BAO+ H0+ TDEL	LRG+ H0	LRG+ H0+ SNCONST	LRG+ H0+ CMB	CMB
ΛCDM	●		●	◆							●
ΛCDM+DELZ	●										
ΛCDM+RUN	●		●			◆					●
ΛCDM+TENS	●		●	◆		◆					●
ΛCDM+RUN+TENS	●		●								
ΛCDM+ISO1	●		●			◆					
ΛCDM+ISO2	●		●			◆					
ΛCDM+MNU	●	▲	●			◆		●	◆		
ΛCDM+YHE	●							●		●	
WCDM+MNU	●		●			●		●	●		
ΛCDM+NREL	●	▲	●					●	◆		
ΛCDM+NREL>3	◆		◆					◆			
OΛCDM	●		●	◆		◆					
WCDM	●	▲	◆		●	◆	●				
OWCDM	●		◆			●	●				

The icons indicate what data is available for a model/dataset pair:

- ▲ Filled Red Triangles — Parameters with Markov chains (WMAP version 4.1, RECFAST version 1.5)
- ◆ Filled Green Diamonds — Post Processed Parameters with spectra and/or Markov chains (WMAP version 4.0, RECFAST version 1.4.2)
- ● Filled Green Circles — Parameters with spectra and/or Markov chains (WMAP version 4.0, RECFAST version 1.4.2)
- □ Hollow Blue Squares — Parameters only

Beyond parameter fitting: model testing

Akaike Information criterion (Akaike 1974; Liddle 04)

$$\mathrm{AIC} \equiv -2 \ln \mathcal{L}_{\max} + 2k \,,$$

k=Number of parameters

Bayesian Information criterion (Schwarz 78, Liddle 04)

$$\mathrm{BIC} \equiv -2 \ln \mathcal{L}_{\max} + k \ln N$$

N=Number of data points

Bayesian Evidence

$$E = \int \mathcal{L}(\theta) \Pr(\theta) d\theta \,.$$

it does not focus on the best-fitting parameters of the model, but rather asks “of all the parameter values you thought were viable before the data came along, how well on average did they fit the data?”

Computationally expensive! (there are packages to help out there e.g. cosmonest)

$$\mathrm{P}(H \mid D) = \frac{\mathrm{P}(D \mid H)\mathrm{P}(H)}{\mathrm{P}(D)}$$

Bayes

$$\mathrm{P}(\boldsymbol{\theta} \mid D, H) = \frac{\mathrm{P}(D \mid \boldsymbol{\theta}, H)\mathrm{P}(\boldsymbol{\theta} \mid H)}{\mathrm{P}(D \mid H)}$$

Bayes, for parameter fitting

$$D_{KL} \equiv \int p(\Theta|D) \ln \frac{p(\Theta|D)}{p(\Theta)} d\Theta.$$

$$\mathcal{E} \equiv \mathrm{P}(D \mid H) = \int \mathrm{d}^n \boldsymbol{\theta} \mathrm{P}(D \mid \boldsymbol{\theta}, H)\mathrm{P}(\boldsymbol{\theta} \mid H)$$

Bayes for the MODEL itself

Suggested exercises

- Go and download the H(z) data from table 2 of the link from http://icc.ub.edu/~liciaverde/clocks.html make a plot in the Ω_m -$\Omega_\Lambda$plane marginalizing over Ho. You can do that using a grid or using a MCMC approach.
- Add a prior given by the measurement of Ho of Riess et al. http://arxiv.org/pdf/0905.0695
- Download one of the WMAP chains, plot confidence limits for a few parameters and for an example of couple of parameters.
- Importance-sample it to add information from e.g. the Ho measurement or the H(z) measurements.
- Or try to compute profile likelihood for one of the parameters and compare the results with the standard MCMC error.

If you are familiar with numerical integrals you can try he SNeIA sample e.g., http://supernova.lbl.gov/Union/

If you are a wizard with computers you can try to install cosmomc and run chains.

Key concepts today

Recap: Likelihoods and chisquared

Confidence levels; confidence regions

Monte Carlo methods

Monte-Carlo errors

MCMC

What confidence intervals

Beyond parameter fitting (intro)

For discussion: statistical errors vs. systematic errors

Interesting literature start appearing…

Beyond parameter fitting

Going minimally parametric

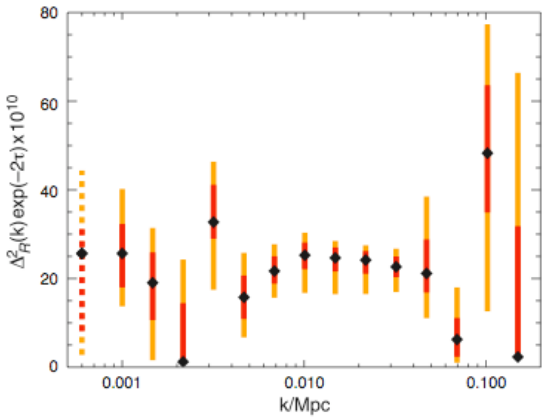
Working example: the shape of the primordial P(k)

Parameter fitting e.g., : $P(k)= A\ (k/k0)^{n-1}$

Other popular options are:

Instead of fitting a function to the data, use a basis function (wavelets, principal components etc…)

Use bins



Spergel et al 07

Piecewise linear

How do you know you are not “fitting the noise”?

How do you know the model (e.g. power law, running) is OK?

Minimally parametric technique
Based on smoothing splines

Splines: Piecewise polynomial (usually cubic) fit. Describe P(k) with splines

Smoothing: Suitable for looking for smooth deviations from power laws

Knots: Discrete values of k,ki. P(ki) will be “free” parameters.
Do spline for the knots

Sealfon et al (2005); Verde, Peris (2008); Peiris, Verde (2010); Bird et al (2010)

How do you know you are not “fitting the noise”?

How do you know the <u>model</u> (e.g. power law, running) is OK?

Minimally parametric technique (in 3 “easy” steps):

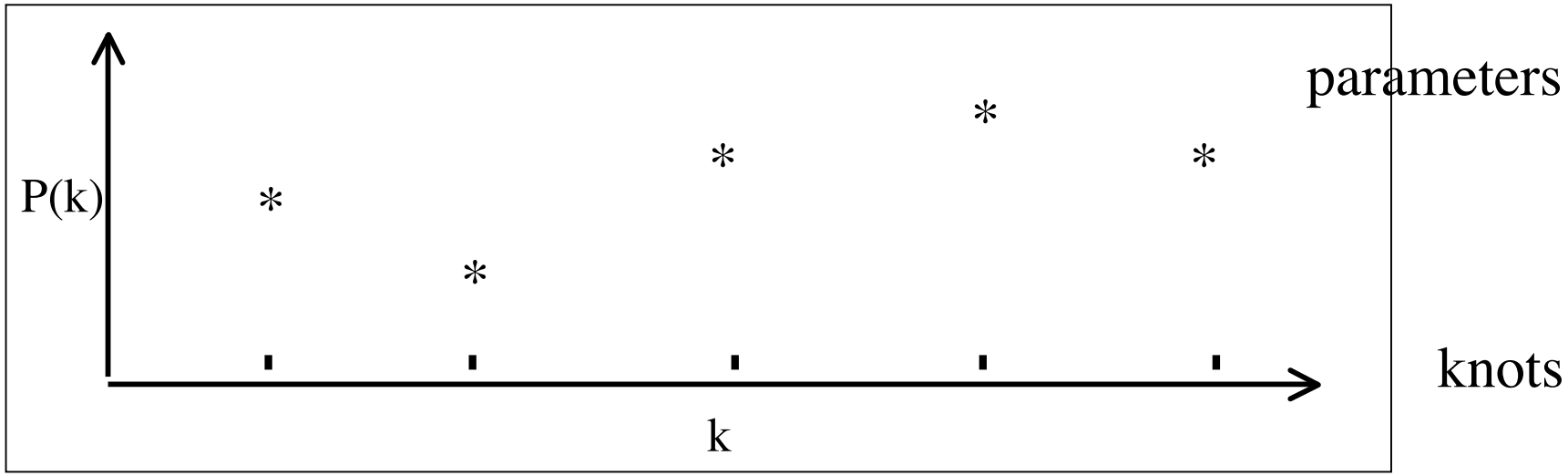
1)Select # knots and use a piecewise cubic spline

2)Penalize the likelihood for the “wiggliness”

3)Use CROSS VALIDATION to chose optimal penalty

full analysis is computationally expensive!

1)



2) $$\log \mathcal{L} = \log \mathcal{L}(\text{Data}|\alpha, P(k)) + \lambda \int_k (P(k)^{''})^2 dk$$

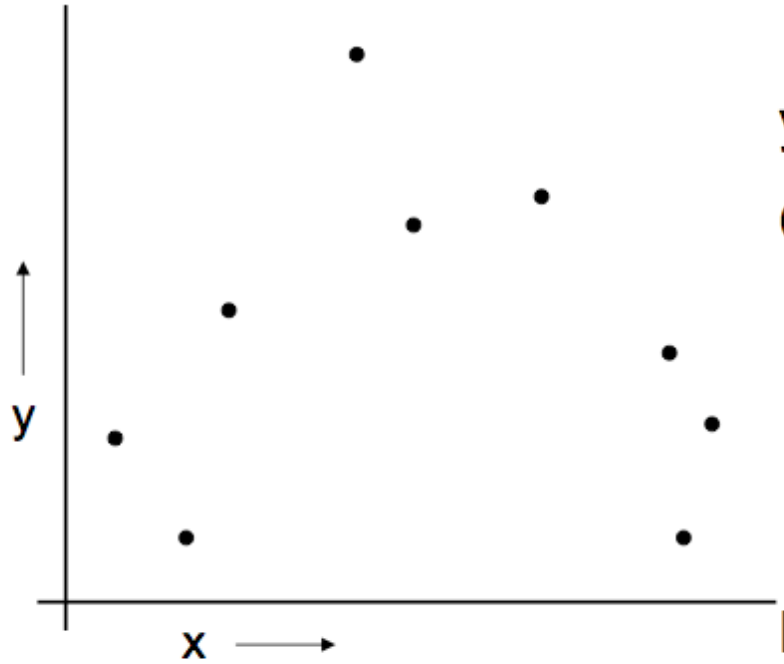
HOW TO SELECT THE BEST PENALTY?

3) Beware of overfitting:

Cross Validation is a powerful technique to make sure one is not fitting the noise

Cross Validation is a powerful technique to make sure one is not fitting the noise

A Regression Problem



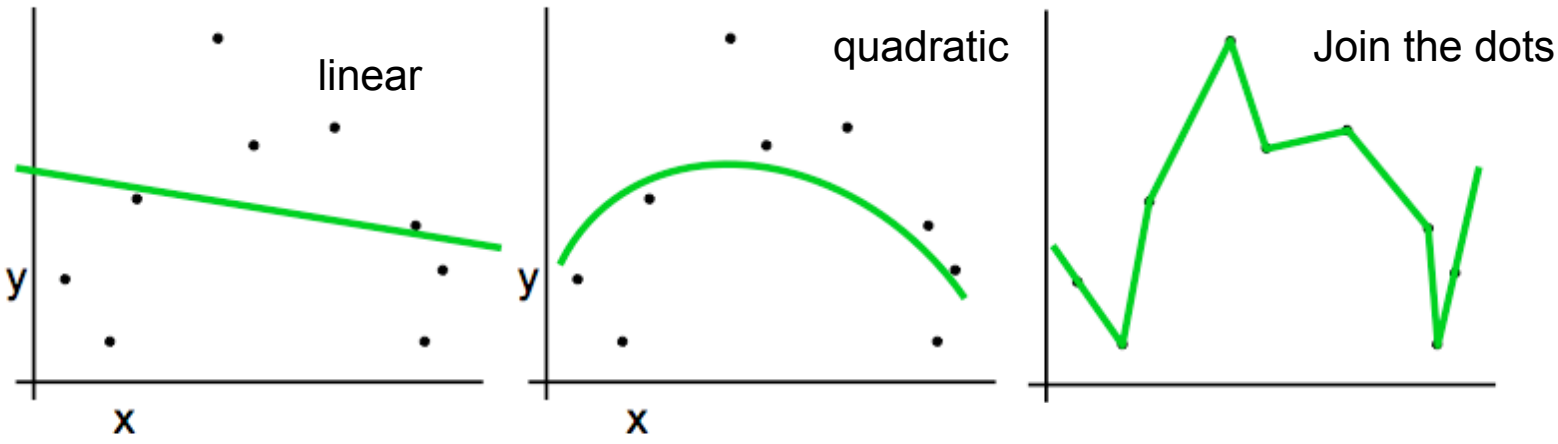
y = f(x) + noise

Can we learn f from this data?

Let's consider three methods…

Copyright ©Andrew W. Moore

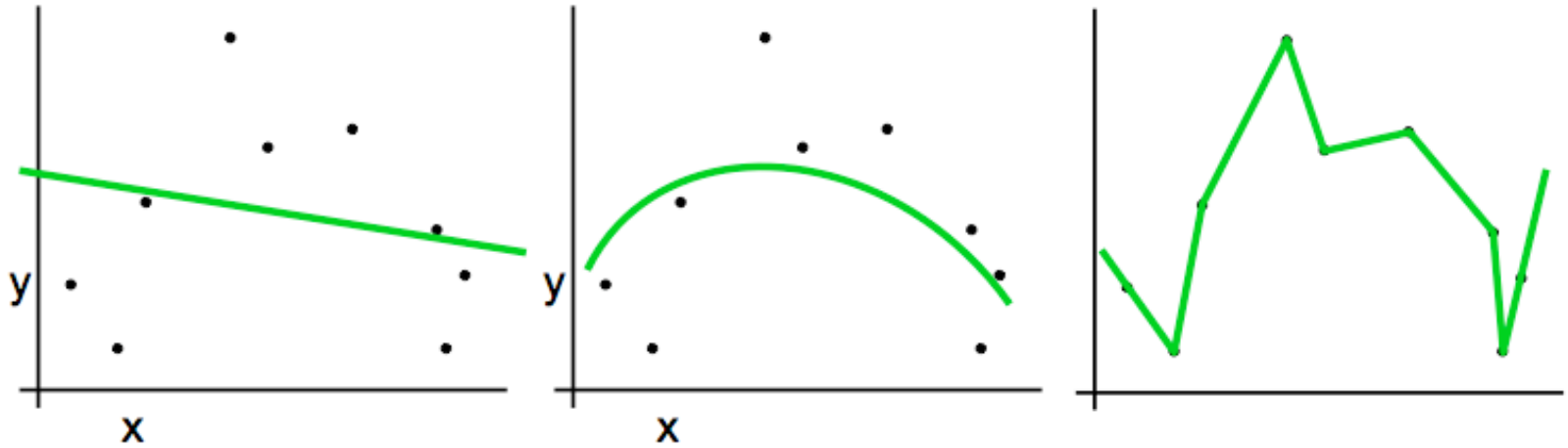
Which is best?



Why not choose the method with the best fit to the data?

Copyright ©Andrew W. Moore

What do we really want?

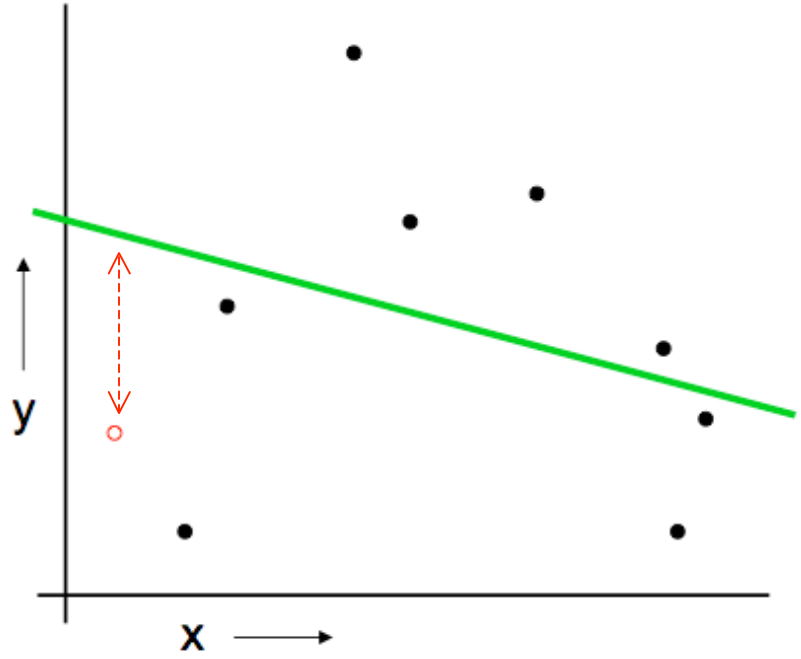


Why not choose the method with the best fit to the data?

"How well are you going to predict future data drawn from the same distribution?"

Copyright ©Andrew W. Moore

Leave one out cross validation:



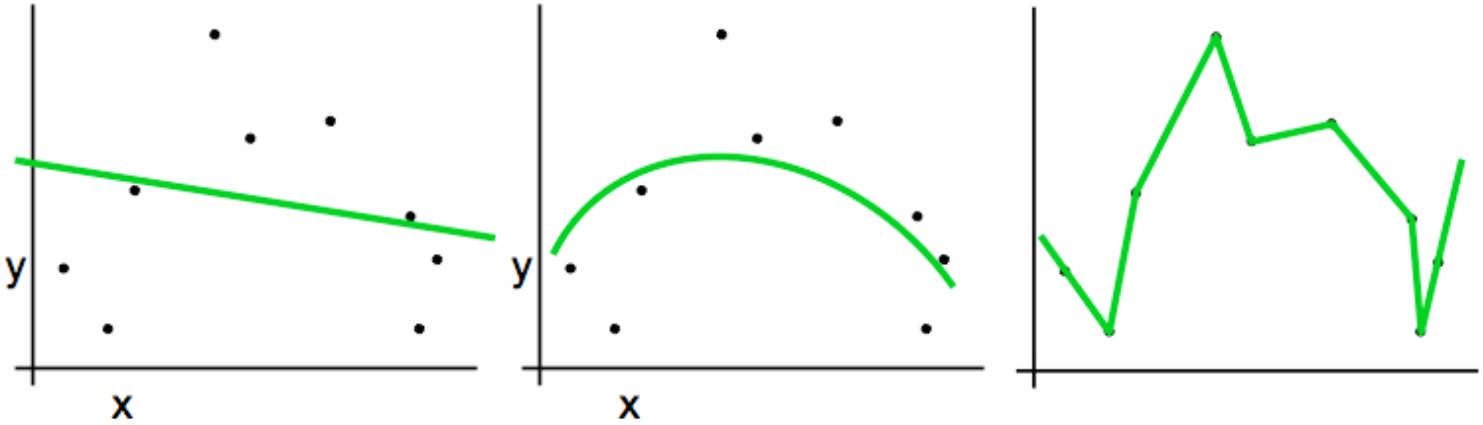
For k=1 to R

1. Let (x_k, y_k) be the kth record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints (example shown for linear model)
4. Note your error (x_k, y_k)

When you've done all points, report the mean error(CV score)

Copyright ©Andrew W. Moore

In this example:



$MSE_{LOOCV} = 2.12$

$MSE_{LOOCV} = 0.962$

$MSE_{LOOCV} = 3.33$

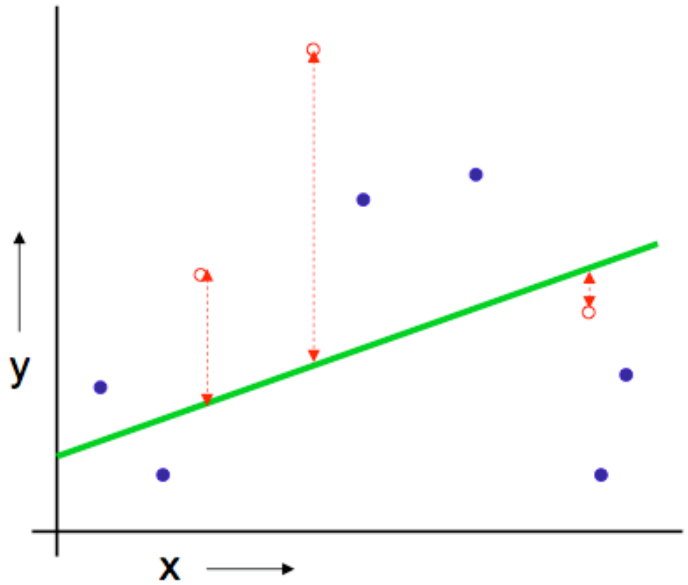
CV score

Leave one out CV is the ideal: does not waste much data but it is very expensive

Copyright ©Andrew W. Moore

The training/test set approach is similar to leave one out CV.

Train on subset of the data



(Linear regression example)

1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

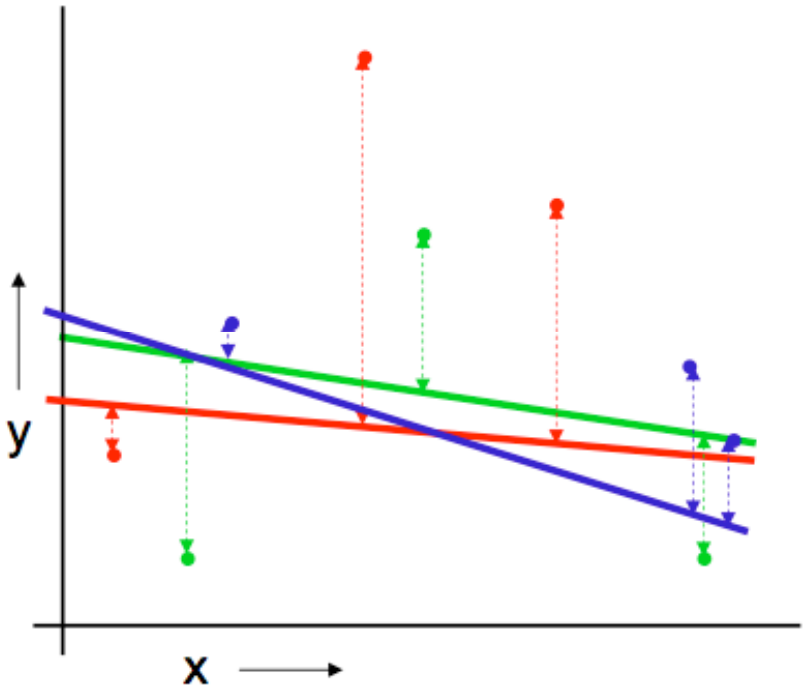
3. Perform your regression on the training set

4. Estimate your future performance with the test set

(this may remind you of training sets for photo-z)

Statisticians prefer:

k-fold Cross Validation



Linear Regression

Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue)

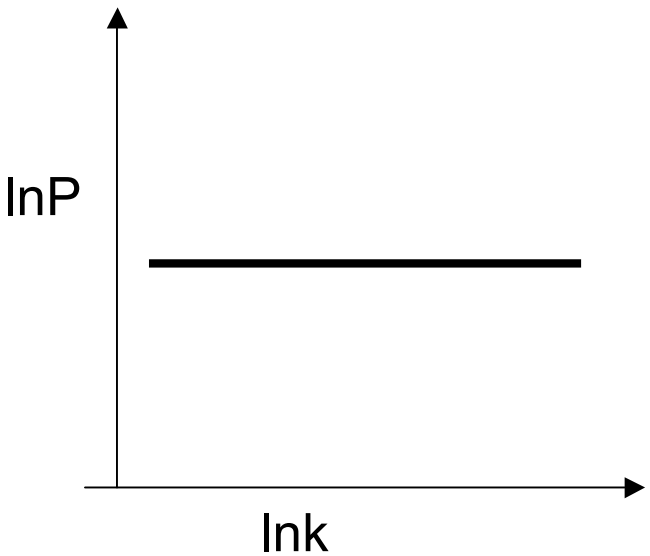
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

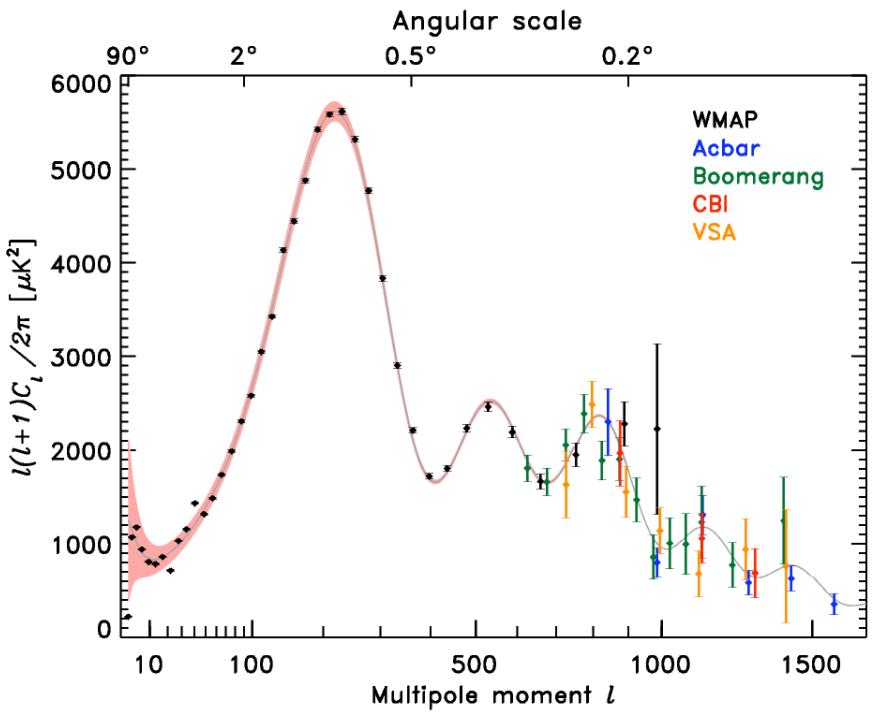
For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error
& compare different models

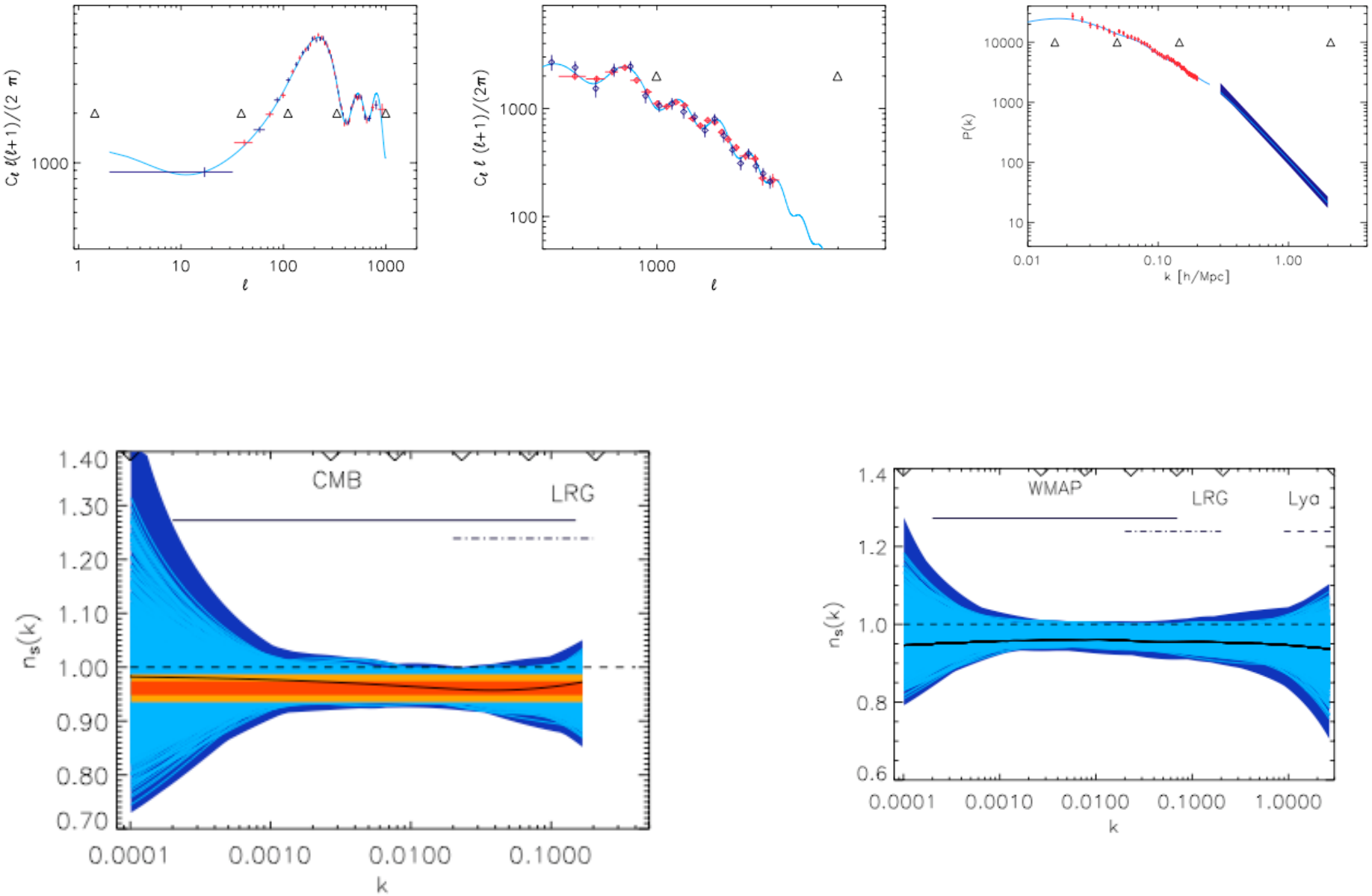
While “leave 1 out” CV would be ideal, it is too computationally intensive; we do 2-fold CV.



Non-analytic Transfer function



Split the data in 2 samples (CV1, CV2)
for each penalty value do a MCMC.
Compute the likelihood for the best fit model from CV1 and data of CV2 and viceversa.
The sum of these two log likelihoods give the CV score.
The optimal penalty is the one that minimizes the CV score.



Peiris, Verde 2010