Likelihood analysis

Florian Beutler

June, 2015

Lawrence Berkeley National Lab

Outline

- Introduction, χ^2 and likelihood
- Covariance matrix
- Gaussian random field
- Marginalization
- MCMC
- Fitting

Introduction, χ^2 and likelihood

- We want to get the likelihood of a model given some observations $\mathcal{L}(\text{model}, \text{data})$.
- Bayes' theorem tells us that:

$$\mathcal{L}(\text{model}, \text{data}) \propto \mathcal{L}(\text{data}, \text{model})\pi(\text{model})$$

- $\pi(\text{model})$ is the prior information on the model.
- Bayes' theorem tells you how to update your likelihood in the light of new data.

Introduction, χ^2 and likelihood

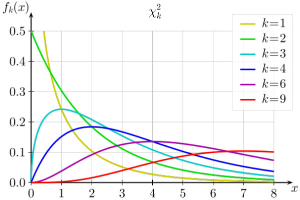
How can we get the likelihood?

$$\chi^2 = \sum_{ij} (D_i - M_i) C^{-1} (D_j - M_j)$$

and the likelihood is $\mathcal{L} \propto \exp(-\chi^2/2)$.

We can now make simple statements like "The model is a good description of the data if $\chi^2 \approx$ d.o.f." or more precisely the probability to get a certain χ^2 when having d degrees of freedom is

$$Q(\chi^2, d) = \left[2^{d/2}\Gamma(d/2)\right]^{-1} \int_{\chi^2}^{\infty} t^{d/2-1} \exp(-t/2) dt.$$



Covariance matrix

- To get the likelihood we need a covariance matrix for the observations.
- The easiest source of the covariance matrix is an analytic model... often such a model does not exist.
- If we have random realizations X_i of our dataset we can get the covariance matrix as

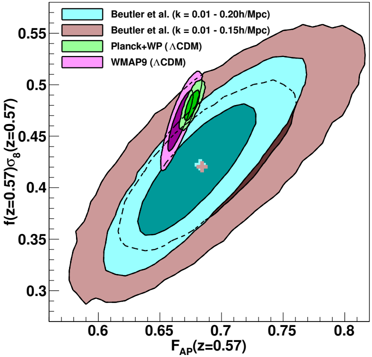
$$C = \frac{1}{N-1}\sum_{ij}(X_i - \overline{X})(X_j - \overline{X}).$$

- Otherwise we can get the covariance matrix directly from the data using jack-knife resampling, bootstrapping... see e.g. Norberg et al. (2008)

Marginelization

The likelihood contains the uncertainties of the model parameters. One can quote marginalized and maximum likelihood uncertainties. To marginalize over the parameter C means

$$\mathcal{L}(\theta) = \int \mathcal{L}(C,\theta)dC$$



MCMC

What is Monte-Carlo Markov chain?

- Monte-Carlo: Use random sampling to solve numerical problems.
- Markov-chain: The probability distribution of the next chain element depends only on the current chain element $P(x_i|x_{i-1})$.
- MCMC is a way to get a sample $X = x_1, x_2, x_3, ...$ from a distribution $\mathcal{L}$ without the need to know the normalization.
- Having a sample X allows us to derive anything we want from the likelihood distribution (confidence intervals, maximum, covariance, standard deviations).
- In the case of only a few parameters there are probably better ways to get the posterior. In the case of many parameters MCMC might be the only way.

MCMC

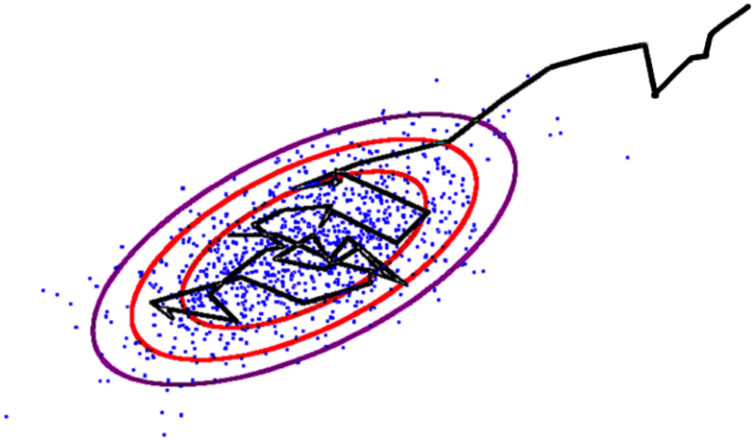
The Metropolis-Hastings algorithm:

1. Choose a start value x_0.
2. Propose a new value x_{new} using the proposal distribution $q(x_0, x_{\mathrm{new}})$.
3. Decide whether the value x_{new} is accepted by calculating the acceptance probability

$$A(x_0, x_{\mathrm{new}}) = min\left(1, \frac{f(x_{new})}{f(x_0)}\frac{q(x_0, x_{\mathrm{new}})}{q(x_{\mathrm{new}}, x_0)}\right).$$

4. If accepted set $x_1 = x_{\mathrm{new}}$, otherwise set $x_1 = x_0$.
5. Repeat this process as long as necessary.

MCMC



MCMC

- A key property of MCMC is that it only depends on likelihood ratios. The normalization of a likelihood function is difficult to evaluate.
- There are some moving parts in the MCMC algorithm... optimization problem
- We need to exclude some burn in phase. It does't tell you anything about the posterior distribution.
- What to choose for the proposal distribution?

MCMC

- A key property of MCMC is that it only depends on likelihood ratios. The normalization of a likelihood function is difficult to evaluate.
- There are some moving parts in the MCMC algorithm... optimization problem
- We need to exclude some burn in phase. It does't tell you anything about the posterior distribution.
- What to choose for the proposal distribution?
- A common choice is a multi-variant Normal (Gaussian) distribution. But what to choose for the width of the distribution?

MCMC

- A key property of MCMC is that it only depends on likelihood ratios. The normalization of a likelihood function is difficult to evaluate.
- There are some moving parts in the MCMC algorithm... optimization problem
- We need to exclude some burn in phase. It does't tell you anything about the posterior distribution.
- What to choose for the proposal distribution?
- A common choice is a multi-variant Normal (Gaussian) distribution. But what to choose for the width of the distribution?
- Use the covariance matrix:
    1. Calculate the covariance matrix from the n-1 chain elements.
    2. Diagonalize the matrix.
    3. Step along the eigenvectors with the step size proportional to the eigenvalues.
- If the proposal distribution is symmetrical, the acceptance probability reduces to the likelihood ratios (Metropolis algorithm).

MCMC

When to stop the chain?

MCMC

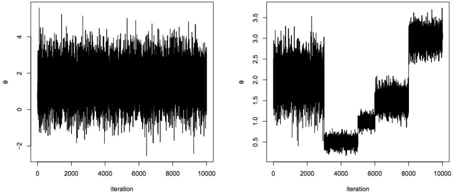
When to stop the chain?

- There is no method which can be proven to converge in a finite time.

MCMC

When to stop the chain?

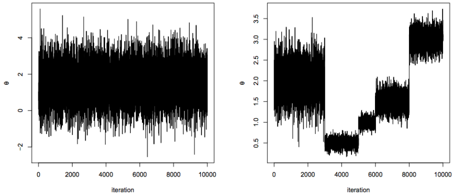
- There is no method which can be proven to converge in a finite time.
- One could visually inspect the chains... the variance and mean should be stable → Traceplot.



MCMC

When to stop the chain?

- There is no method which can be proven to converge in a finite time.
- One could visually inspect the chains... the variance and mean should be stable → Traceplot.



- One could run two chains in parallel and compare the chains. Convergence can be defined as the state when the two chains give the same likelihood (e.g. the mean dispersion of a parameter between the chains ≪ the mean dispersion of the parameter within the chain).

Gelman and Rubin convergence criterium

Steps (for each parameter):

1. Run $m \geq 2$ chains of length $2n$ from over-dispersed starting values.
2. Discard the first n draws in each chain.
3. Calculate the within-chain and between-chain variance.
4. Calculate the estimated variance of the parameter as a weighted sum of the within-chain and between-chain variance.
5. Calculate the potential scale reduction factor.

Gelman and Rubin convergence criterium

- Within chain variance:

$$s_j^2 = \frac{1}{n-1}\sum_i^n (\theta_{ij} - \overline{\theta_j})^2$$

- Mean of the chain variances:

$$W = \frac{1}{m}\sum_j^m s_j^2$$

- Between chain variance:

$$B = \frac{n}{m-1}\sum_{j=1}^m (\overline{\theta}_j - \hat{\theta})^2$$

with $\hat{\theta} = \frac{1}{m}\sum_{j=1}^m \overline{\theta}_j$. This is the variance of the chain means multiplied by n because each chain is based on n draws.

Gelman and Rubin convergence criterium

- We want that the variance of the means (B) is much smaller than the variance of the parameter (W).
- Calculate a weighted average of B and W

$$V(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$$

- The scale reduction factor is

$$R = \sqrt{\frac{V(\theta)}{W}}$$

- R should be $< 1 + \epsilon$
- Calculate R for each parameter
- We can then combine the *mn* total draws from our chains to produce one chain from the stationary distribution.

Summary

- MCMC is a way to sample the likelihood distribution.
- If you have a small parameter space there are simpler ways to get your likelihood, but in a high parameter space MCMC might be the only way.
- When to stop the chain is a difficult question, especially if your likelihood has many local minima.

Model selection

Put here the model selection slides...