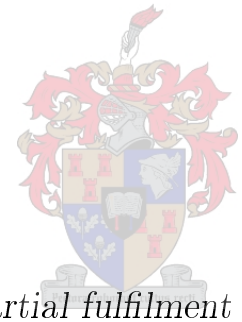
Bayesian parameter estimation for discrete data spectra

by

Li Wang



Thesis presented in partial fulfilment of the requirements for the degree of Master of Science (Theoretical Physics) in the Faculty of Science at Stellenbosch University

Supervisor: Prof. H.C. Eggers

December 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2017

Abstract

Bayesian parameter estimation for discrete data spectra

L. Wang

Department of Physics,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc (Theoretical Physics)

September 2017

Discrete spectra are ubiquitous in physics; for example nuclear physics, laser physics and experimental high energy physics measure integer counts in the form of particles in dependence of angle, wavelength, energy etc. Bayesian parameter estimation (fitting a function with free parameters to the data) is a sophisticated framework which can handle cases of sparse data as well as input of pertinent background information into the data analysis in the form of a prior probability. Bayesian comparison of competing models and functions takes into account all possible parameter values rather than just the best fit values. We first review the general statistical basis of data analysis, focusing in particular on the Poisson, Negative Binomial and associated distributions. After introducing the conceptual shift and basic relations of the Bayesian approach, we show how these distributions can be combined with arbitrary model functions and data counts to yield two general discrete likelihoods. While we keep an eye on the asymptotic behaviour as useful analytical checks, we then introduce and review the theoretical basis for Markov Chain Monte Carlo numerical methods and show how these are applied in practice in the Metropolis-Hastings and Nested Sampling algorithms. We proceed to apply these to a number of simple situations based on simulation of a background plus two or three Gaussian peaks with both Poisson and Negative Binomial likelihoods, and discuss how to select models based on numerical outputs.

Uittreksel

Bayesiese parameterberaming vir diskrete dataspektra

("Bayesian parameter estimation for discrete data spectra")

L. Wang

Departement Fisika,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc (Teoretiese Fisika)

September 2017

Diskrete spektra is 'n algemene verskynsel in fisika: kernfisika, laserfisika en eksperimentele hoë-energiefisika meet byvoorbeeld heelgetalle in die vorm van deeltjies as 'n funksie van hoek, golflengte, energie ens. Bayesiese parameterberaming (die passing van 'n funksie met vrye parameters op die data) is 'n gesofistikeerde raamwerk wat gevalle van lae tellings asook pertinente agtergrondinligting as inligting vir die data-analise in die vorm van prior-waarskynlikhede kan hanteer. Bayesiese vergelyking van kompeterende modelle en modelfunksies neem alle moontlike parameterwaardes in ag eerder as net die enkele beste waardes daarvan. Ons gee eerstens 'n oorsig van die algemene statistiese basis van data-analise met 'n besondere fokus op die Poisson-, Negative Binomial- en verwante verdelings. Die konseptuele omwenteling wat Bayes impliseer en die basiese vergelykings word bespreek, waarna ons wys hoe hierdie verdelings met willekeurige modelfunksies en datatellings gekombineer kan word om twee algemene diskrete likelihood-waarskynlikhede te skep. Terwyl ons 'n oog hou op die asimptotiese gedrag as nuttige analitiese verwysings, gee ons daarna 'n inleiding tot en sit ons die teoretiese basis van Markovketting Monte Carlo numeriese metodes uiteen en wys hoe hulle in die vorm van die Metropolis-Hastings en Nested Sampling algoritmes toegepas word. Ons pas hierdie algoritmes op 'n aantal eenvoudige situasies gebaseer op simulasies van 'n agtergrond plus twee of drie Gaussiese pieke toe met sowel Poisson asook Negative Binomial waarskynlikhede, en bespreek hoe om modelle te kies gebaseer op numeriese uitsette.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Eggers for his enthusiasm, insightful comments, and hard questions. His guidance helped me in all the time of research and writing of this thesis.

My thanks also goes to all my lecturers in the Physics Department for their patient guidance and generous help.

Dedications

This thesis is dedicated to my son Zijuan Guo.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background and motivation

The Least Squares method of fitting binned data

Figure 1.1 is a simple data spectrum from a counting experiment. We observe a peak centered at μ with standard deviation σ, and wish to fit the peak and background using a Gaussian function with amplitude β_1 and background noise β_2,

$$f(x_b, \boldsymbol{\beta}) = \beta_1 \exp\left\{ -\frac{(x_b - \mu)^2}{2\sigma^2} \right\} + \beta_2. \tag{1.1}$$

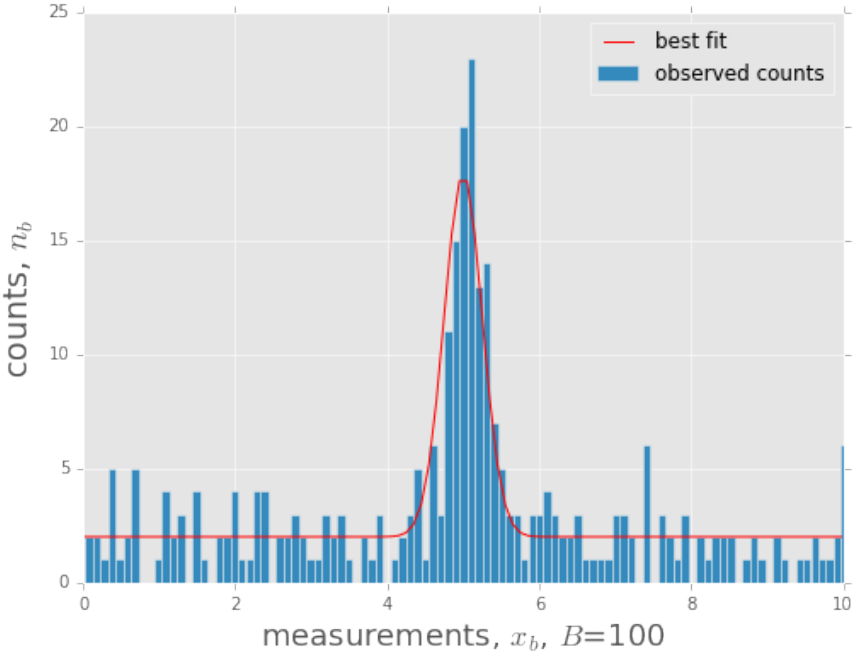


Figure 1.1: Toy model counts spectrum for 100 bins. The blue bars represent the number of observations (counts) in each bin centered on midpoints x_b, while the red line is a typical "best fit".

The best-fit parameters are conventionally obtained by minimizing Least Squares based on the assumption that data are Gaussian distributed around the fit function $f(x_b, \boldsymbol{\beta})$,

$$\chi^2 = \sum_{b=1}^{B} \frac{\big(n_b - f(x_b, \boldsymbol{\beta})\big)^2}{\sigma_b^2} \tag{1.2}$$

where $\big(n_b - f(x_b, \boldsymbol{\beta})\big)$ is the error in each channel and σ_b is the experimental standard deviation of n_b. β_1 and β_2 are solved by setting the partial derivatives with respect to β_1 and β_2

$$\frac{\partial \chi^2}{\partial \beta_1} = -2 \sum_{b=1}^{B} \frac{n_b - f(x_b, \boldsymbol{\beta})}{\sigma_b^2} \exp\left\{ - \frac{(x_b - \mu)^2}{2\sigma^2} \right\} = 0 \tag{1.3}$$

$$\frac{\partial \chi^2}{\partial \beta_2} = -2 \sum_{b=1}^{B} \frac{n_b - f(x_b, \boldsymbol{\beta})}{\sigma_b^2} = 0 \tag{1.4}$$

Where experimental standard deviations are not measured separately, one typically assumes that the counts are Poisson-distributed and that $\sigma_b^2 \approx n_b$, in which case

$$\chi^2 = \sum_{b=1}^{B} \frac{\big(n_b - f(x_b, \boldsymbol{\beta})\big)^2}{n_b} \tag{1.5}$$

The Least Squares method works well in many cases, however, when the number of counts is low, it encounters problems. We illustrate this in Figure 1.2.

Firstly, the number of counts is an integer, while the Gaussian variable x is real.

Secondly, the Gaussian covers the entire real line, and when its peak is near zero, a significant part of its lower tail may fall into the negative-x part, as shown in the upper left panel. Data counts, on the other hand, can never be negative, and so the Gaussian is clearly inappropriate. The problem persists to some degree even for an average number of counts of 5 as shown in the upper right panel.

Thirdly, the Gaussian probability function is symmetric i.e. the most probable value is at the center, while in counting experiments, especially rare events, the distribution is asymmetric. This is quite clear for small counts, but actually persists even for larger counts as shown in the lower two panels. Putting a Gaussian onto a Poisson count set even for an average of 50 counts represents a small but systematic bias.

Using the Least Squares method in such cases is bad, and assuming $\sigma_b^2 \approx n_b$ makes it even worse: as shown in Figure 1.1, there are zero counts in some bins, and by excluding zero bins from the fit, we are throwing away important information. As will be explained in Figure 3.2, a null observation does not necessarily mean a zero signal.

We therefore see that for small numbers of integer counts, there is a need to replace the Least Squares method by something based on the Poisson distribution.

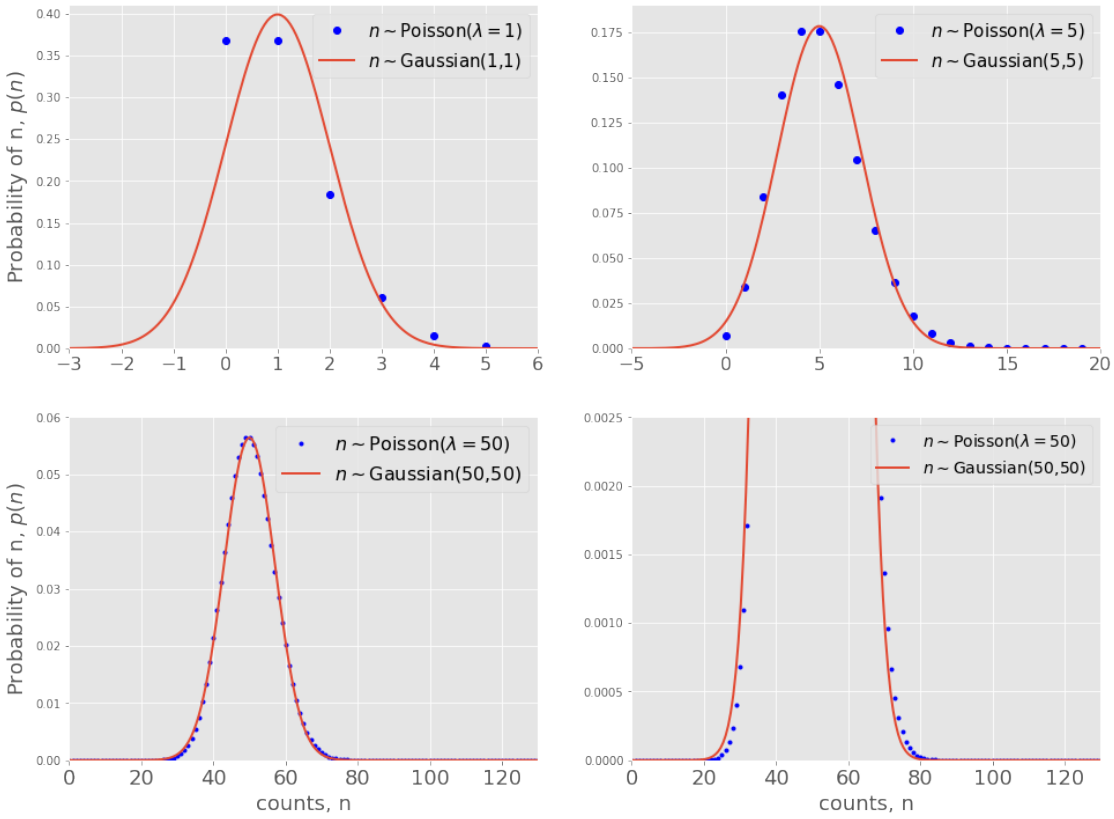


Figure 1.2: The upper two figures show that for small rate parameter λ, Gaussian and Poisson distributions differ significantly; counts are integers and cannot be negative. In the bottom two figures for $\lambda = 50$, the Poisson and Gaussian do look very similar, there is nevertheless a residual asymmetry in the Poisson distribution as the zoomed detail in the fourth panel shows: there is a systematic bias.

The Poisson model and its extensions

While the Poisson distribution

$$p(n \mid \lambda) = \frac{e^{-\lambda}\lambda^n}{n!} \qquad n = 0, 1, 2, \ldots, \quad \lambda > 0$$

is clearly the first option to describe discrete counts, it can easily be proved to be inadequate. The Poisson mean and variance both equal to the rate parameter λ, so for any given mean the variance is fixed. When the variance of observations is greater than the variance predicted by the Poisson equality, var(Observations)/var(Poisson model) > 1, there is no mechanism or parameter to describe this within the Poisson framework.

One possible reason for larger variance or "over-dispersion" is positive correlation between individual counts or "clustering", thus violating the assumption of independent observations which led to the Poisson distribution, since it arises from an underlying assumption that events are independent. Other reasons such as outliers in the data or other nonstandard features also contribute to additional variance.

We shall in this thesis explore the Negative Binomial distribution as a fairly obvious extension to the Poisson case. Since the Negative Binomial has two parameters, it is clear from the start that its variance and mean can be disentangled.

1.2 Probability as relative frequency and as degree of belief

Probability theory is nothing but common sense reduced to calculation.
Pierre Laplace, 1812

The primary purpose of probability theory rests in inferring patterns from seemingly random phenomena. This provides a platform to formulate a model concerning existing information in order to make conclusions and predictions.

In general, there are two approaches to defining "probability":

1. Frequentist definition:

   An event A can occur or not occur in a given situation. After many repetitions, commonly called *trials*, the probability for this event to occur is defined by the long-run *relative frequency*, the ratio of the number of occurrences n to the number of trials N [1],

   $$p(A) = \lim_{N\to\infty} \frac{n}{N}. \tag{1.6}$$

   For example, if we would like to know whether a coin is fair, the probability that this coin lands face up (head), θ, is obtained by flipping this coin in N times repeatedly, then count the number of heads n, thus θ is determined by its frequency of occurrence, which is n/N.

   In this view, the probability for an event A to occur is acquired by running an experiment of infinite repetitions. In other words, A must be a random variable, a quantity that fluctuates throughout repeated experiments or a physically meaningful ensemble [2].

   However, this understanding of probability restricts its scope. In experimental physics, many problems are of such a nature that many repetitions are not possible and limited measurements or trials are unavoidable. In cosmology and astrophysics, the number of observations is limited to the observed cases, and instrumental restrictions or budget or computer time often mean that $N \to \infty$ does not describe the experimental reality.

2. Bayesian probability theory:

   The Bayesian concept of probability does not rely on physical trials; rather, it quantifies the probability of *logical propositions* or logical statements A, given (for example) some background information $\mathcal{I}$ and the assumed truth of some other proposition B. The probability $p(A\,|\,B,\mathcal{I})$ then quantifies a reasonable degree of belief that A is true in a situation where lack of information or data means that it may or may not be true, or acquire different values. One can read it as "Given background information $\mathcal{I}$, and some other proposition B being true, we can make a predictive statement: My present degree of belief for proposition A to be true has value $p(A\,|\,B,\mathcal{I})$."

   In practice, $p(A\,|\,B,\mathcal{I})$ represents "The probability distribution of parameter values given observations and background information", or "The probability that a hypothesis $\mathcal{H}$ is the true description of data." A more complete account of Bayesian theory is provided in Section 3.1.

In Bayesian statistics, we assign probability distributions to parameters. Both the Poisson model and Negative Binomial model are studied in this thesis and some basic situations are compared in explicit examples. It will soon become clear that the conceptual beauty of Bayesian calculus comes at the price of a much expanded computational task. For that reason, the family of Monte Carlo methods is introduced and used to calculate numerically what is impossible to do analytically.

Chapter 2

Statistical basis

In this chapter, we set out the necessary mathematical and statistical basis for probability theory in general. The specifically Bayesian concepts and methods will appear in Section 3.1.

2.1 Basic mathematical relations

Mathematical concepts and formulae in this section are common to all approaches to probability theory. In Chapter 3, these are interpreted and extended within the Bayesian framework.

2.1.1 Probability basics

The concepts and relations set out here are well known and will be treated only very briefly.

- Any probabilistic system is defined by the threesome of the variable X, its outcome or sample space $\mathcal{A}(X)$ and the corresponding probability $p(X)$ for each possible X.
- The "*random*" *variable* X is any question or operation or situation which has more than one possible answer or outcome. We assume that X is a real number or else an integer.
- The *outcome space* $\mathcal{A}(X)$ is the set of all possible values of X, $x \in \mathcal{A}(X), \forall x$. It can be a discrete or continuous set, finite or infinite.
- The *probability* $p(X{=}x)$ is the probability of obtaining a particular outcome x of random variable X. The *probability density function* (pdf) is a function that assigns a probability to each outcome of continuous random variable X; the pdf is a density because $\int \mathrm{d}x\, p(x) = 1$ is dimensionless. For discrete variables we speak of the *probability mass function* (pmf) which is dimensionless.
- Probabilities are *normalised* in the sense that $\int_{\mathcal{A}(X)} p(x)\,\mathrm{d}x = 1$ for a pdf and $\sum_{\mathcal{A}(X)} p(x) = 1$ for a pmf.

- The *Cumulative Distribution Function* or cdf $F(x)$ is the sum (or integral) of all probabilities of outcomes up to a given maximum outcome x; it is a monotonically increasing function and is distributed over the interval $[0, 1]$,

$$F(x) = P(X \leq x) = \sum_{X \leq x} p(X), \quad \text{for the discrete case,}$$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(X)\,\mathrm{d}X, \quad \text{for the continuous case.}$$

  The sums are replaced by integrals for continuous random variables throughout the theory.

- When two variables X and Y are considered at the same time, we speak of the *joint probability* $p(X, Y)$ of both X and Y having some particular values.

- Given a joint probability mass function, the *marginal probabilities* for Y and for X respectively are

$$p(Y) = \sum_{X \in \mathcal{A}(X,Y)} p(X, Y), \tag{2.1}$$

$$p(X) = \sum_{Y \in \mathcal{A}(X,Y)} p(X, Y), \tag{2.2}$$

  with equivalent integrals for the pdf cases.

- The *conditional probability* of a variable X given any hypothesis $\mathcal{H}$ or variable Y or any other information or combination of such is denoted by the use of a vertical line, e.g. $p(X \mid \mathcal{H}), p(X \mid Y)$ where quantities to the right of the line are by hypothesis known and true, while those to the left are unknown and predicted.

- The *product rule* forms the basis of the calculus of probability. Given any two variables X and Y, the joint probability $p(X, Y)$ can always be expressed as the probability of X times the conditional probability of Y given that X is known,

$$p(Y, X) = p(Y \mid X)\, p(X), \tag{2.3}$$

  where $p(Y)$ is the marginal; of course the reverse is also true if X and Y are exchangable,

$$p(X, Y) = p(X \mid Y)\, p(Y). \tag{2.4}$$

  The product rule is sometimes also called Bayes' rule.

- Two variables are *independent* if the joint probability factorises or equivalently the conditional probability becomes independent of the known variable,

$$p(X, Y) \overset{\text{SI}}{=} p(X)\, p(Y) = p(Y)\, p(X), \tag{2.5}$$

$$p(X \mid Y) \overset{\text{SI}}{=} p(X), \tag{2.6}$$

  from which $p(Y \mid X) = p(Y)$ follows immediately.

- The combination of the product rule with marginalisation yields the *chain rule*,

$$p(X) = \sum_Y p(X, Y) = \sum_Y p(X \mid Y)\, p(Y). \tag{2.7}$$

- All of the above can be extended to three or more variables $X_1, X_2, \ldots X_N$ in many different combinations. One example would be N independent variables, for which the joint probability density factorises

$$\begin{aligned} p(X_1, X_2, X_3 \ldots X_N) &= p(X_1)\, p(X_2 \mid X_1) p(X_3 \mid X_2, X_1) \ldots p(X_N \mid X_{N-1}, X_{N-2} \ldots, X_1) \\ &\overset{\text{SI}}{=} \prod_{i=1}^{N} p_i(X_i), \end{aligned} \tag{2.8}$$

  where each $p_i(X_i)$ is the appropriate marginal. If in addition the N marginals are given by the same pdf or pmf function and the outcome spaces are the same for all X_i, we speak of *independent identically distributed* (i.i.d.) variables,

$$p(X_1, X_2, \ldots, X_N) \overset{\text{iid}}{=} \prod_{i=1}^{N} p(X_i). \tag{2.9}$$

- The *mode* x^* of a distribution is that point or outcome in $\mathcal{A}(X)$ at which $p(X)$ reaches a maximum, $x^* = \text{argmax}_X\, p(X)$ or in more conventional physics notation

$$\left.\frac{\partial p(X)}{\partial X}\right|_{X=x^*} = 0, \qquad \left.\frac{\partial^2 p(X)}{\partial X^2}\right|_{X=x^*} < 0. \tag{2.10}$$

  $p(X)$ is called *unimodal* if it has only one mode and *multimodal* otherwise. A pdf is called *uniform* if it is constant between finite limits,

$$U(X \mid x_{\min}, x_{\max}) = \frac{1}{x_{\max} - x_{\min}} \tag{2.11}$$

- Transformation properties: when X is transformed to a new variable U by some known invertible transformation $f(X)$, the outcome space transforms accordingly, $\mathcal{A}(U) = \{u \mid u = f(x), x \in \mathcal{A}(X)\}$. For discrete probabilities, the transformed pmf relates to the original one by

$$p(U) = \sum_X p(X)\, \delta(U, f(X)), \tag{2.12}$$

  where the Kronecker delta $\delta(a, b)$ is 1 if $a = b$ and 0 otherwise. For continuous X and U, the density implies an additional Jacobian factor

$$p(Y) = p(X) \left| \frac{\mathrm{d}X}{\mathrm{d}Y} \right| \tag{2.13}$$

  which follows the fact that the probability mass, not the density, must be invariant under change of variables. For transformations of more than one variable, the corresponding probability transformation involves the absolute value of the Jacobian determinant,

$$p(U_1, U_2, \ldots, U_N) = p(X_1, X_2, \ldots, X_N) \left| \frac{\partial(X_1, \ldots, X_N)}{\partial(U_1, \ldots U_N)} \right|. \tag{2.14}$$

2.1.2 Interval probabilities

Often we are interested not only in finding the mode of a probability but a measure of uncertainty around that maximum. The simplest measure of uncertainty is the *variance* or more specifically its square root, the standard deviation. These will be treated below. More generally, one would quote *quantiles* along with the *confidence interval* associated with the relevant interval probability.

The cumulative distribution for a particular value $x_{1-\alpha}$ is

$$F(x_{1-\alpha}) = p(X \leq x_{1-\alpha}) = 1 - \alpha \tag{2.15}$$

where $1-\alpha$ is the quantile, and $x_{1-\alpha} = F^{-1}(1-\alpha)$ is called quantile value. A *central interval* (x_1, x_2) is then defined with x_1 and x_2 the smallest and largest elements in outcome space $\mathcal{A}(X)$ subject to the constraints

$$F(x_1) = p(X \leq x_1) = \alpha/2 \tag{2.16}$$

$$1 - F(x_2) = p(X \geq x_2) = \alpha/2 \tag{2.17}$$

or

$$F(x_1 < x < x_2) = 1 - \alpha \tag{2.18}$$

so that the bounds (x_1, x_2) represent the $100 \times (1-\alpha)\%$ confidence interval of variable X.

2.1.3 Moments, cumulants and generating functions

While the full functional form of a pdf or pmf contains the ultimate information, it is for unimodal cases often helpful to describe it in terms of a few numbers to characterise it. By far the most common quantities are the mean and variance, which are the first two so-called *cumulants*. In addition to yielding the mean and variance, the associated *generating functions* provide further information and powerful problem-solving tools.

The various generating functions play an important role in many calculations; for example the Central Limit Theorem, convolutions and systems with additional constraints. Firstly, for integer variables n, the *probability generating function* (pgf) is defined as the expectation value of the n-th power of the dual variable z,

$$G(z) = E(z^n) = \sum_n p(n)\, z^n, \tag{2.19}$$

which, if the sum can be expressed in closed form, yields probabilities by taking derivatives,

$$p(n) = \frac{1}{n!} \left(\frac{d}{dz} \right)^n G(z) \bigg|_{z=0}. \tag{2.20}$$

Moments of order $q = 1, 2, 3, \ldots$ are defined as

$$\mu_q = \sum_{\mathcal{A}(x)} x^q p(x); \tag{2.21}$$

they can be calculated either directly by carrying out the sum (or the integral when X is continuous variables) or indirectly from the *moment generating function* (mgf),

$$M(t \,|\, p(x)) = E(e^{tx}) = \sum_{\mathcal{A}(x)} e^{tx} p(x), \tag{2.22}$$

which we often abbreviate to $M(t)$. If $M(t)$ is a closed function of t, successive derivatives of M yield the moments

$$\mu_q(x) = \left(\frac{d}{dt}\right)^q M\big(t \,|\, P(x)\big)\Big|_{t=0}. \tag{2.23}$$

The *cumulant generating function* (cgf) is related to the mgf by

$$K(t \,|\, P(x)) = \ln M(t|\,|\, P(x)) = \ln E(e^{tx}), \tag{2.24}$$

and cumulants of order $q = 1, 2, \ldots$ are also obtained by differentiation,

$$\kappa_q(x) = \left(\frac{d}{dt}\right)^q K\big(t \,|\, P(x)\big)\Big|_{t=0}. \tag{2.25}$$

The set of cumulants have important properties. The first two are the *mean* and the *variance*,

$$\kappa_1(x) = \mu(x) = E(x) \tag{2.26}$$
$$\kappa_2(x) = \mathrm{var}(x) = \mu_2 - \mu_1^2 = E(x^2) - E(x)^2 \tag{2.27}$$

where κ_1 measures the expected value or *location* of a random variable while κ_2 is the square of the width or "standard deviation" measuring the *scale* of the peak around κ_1. Beyond these first two, cumulants have some powerful properties:

1. Additivity under convolution of i.i.d variables: Given N i.i.d. variables, the cumulant of their convolution is N times the cumulant of the individual one,

$$\kappa_q\Big(\sum_{i=1}^{N} X_i\Big) = N\kappa_q(X) \tag{2.28}$$

2. Properties under linear transformation: Defining $U = aX + c$, where a, c are constant, the respective cumulants transform as

$$\kappa_1(U) = a\kappa_1(X) + c, \tag{2.29}$$
$$\kappa_q(U) = a^q\kappa_q(X), \quad \forall q = 2, 3, \ldots. \tag{2.30}$$

3. The third- and fourth-order cumulants characterise further properties of the distribution. Defining $K^{(q)}$ to be the qth derivative with respect to t, the *skewness*

$$\gamma_1(x) = \frac{\kappa_3}{\sqrt[3]{\kappa_2}} = \frac{K^{(3)}(t)\big|_{t=0}}{\sqrt[3]{K^{(2)}(t)\big|_{t=0}}} \tag{2.31}$$

measures the asymmetry of the probability distribution with respect to its mode. Symmetric distributions have zero skewness. A probability with negative skew is said to be left-tailed; the mass of the distribution is concentrated on the right of the probability's maximum (the "mode"). Positive skew is the opposite to negative skewness.

The *kurtosis* is defined by

$$\gamma_2(x) = \frac{\kappa_4}{\kappa_2^2} = \frac{K^{(4)}(t)\big|_{t=0}}{\big(K^{(2)}(t)\big|_{t=0}\big)^2} \tag{2.32}$$

High kurtosis means heavy tails, or outliers, while a distribution with low kurtosis tends to have light tails and few outliers.

2.2 Specific probability distributions

Different situations and needs require different tools. In this section, we summarise those probability distributions which we shall need later. For those with integer outcomes, the random variable will be termed n, while real random variables are termed x. Parameters are often specified by Greek letters, but there are some exceptions.

2.2.1 Binomial distribution

Suppose that a fisherman has N hooks in the sea to catch fish; for each specific hook, the probability is θ that a fish is caught and $(1-\theta)$ that no fish is caught. The single-fish case is called a Bernoulli distribution with catch (success) probability θ and failure probability $(1-\theta)$. Defining n as the number of hooks which did catch a fish, then $(N-n)$ is the number of hooks which failed. Since we are not interested in the identity of the hook but only the total number n of catches, the $\binom{N}{n}$ possible combinations of hooks which could have caught them, so that the probability for n follows a binomial distribution

$$p(n|N,\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}, \quad n = 0, 1, 2, \cdots, N; \tag{2.33}$$

it generally describes the probability of n successes given fixed N independent trials each with success probability θ.

The convention is to characterise each distribution with a two-letter code. For the Binomial, we therefore write

$$\mathrm{Be}(n \,|\, N,\theta) = p(n|N,\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}, \quad n = 0, 1, 2, \cdots, N; \tag{2.34}$$

which is sometimes shortened to $\mathrm{Be}(N,\theta)$.

Properties

The moment generating function for the binomial distribution is

$$\begin{aligned} M(t) = E(e^{tn}) &= \sum_{n=0}^{\infty} e^{tn}\binom{N}{n}\theta^n(1-\theta)^{N-n} = \sum_{n=0}^{\infty}\binom{N}{n}(\theta e^t)^n(1-\theta)^{N-n} \\ &= \left(\theta e^t + (1-\theta)\right)^N \end{aligned} \tag{2.35}$$

and the cumulant generating function follows as

$$K(t) = \ln M(t) = N\ln\left(\theta e^t + (1-\theta)\right) \tag{2.36}$$

resulting in the mean, variance, skewness and kurtosis,

$$\mu(n) = N\theta \tag{2.37}$$

$$\mathrm{var}(n) = N\theta(1-\theta) \tag{2.38}$$

$$\gamma_1 = \frac{1-2\theta}{\sqrt{N\theta(1-\theta)}} \tag{2.39}$$

$$\gamma_2 = \frac{1-6\theta(1-\theta)}{N\theta(1-\theta)}. \tag{2.40}$$

The probability generating function is given by

$$G(z) = E(z^n) = [\theta z + (1-\theta)]^N \tag{2.41}$$

2.2.2 Poisson distribution

We pay particular attention to the Poisson distribution as it is used extensively in later chapters.

If data consists of a set of discrete events distributed in space, time, energy, angle or some other informative variables, it is common to describe their probability in terms of Poisson distributions. As shown below, the Poisson distribution has a particular simple and general form and rivals the Gaussian distribution in its generality and many ways it is used.

Given the detected counts n, the Poisson distribution is characterized in terms of a parameter λ as

$$\mathrm{Pn}(\lambda) = p(n \,|\, \lambda) = \frac{e^{-\lambda}\lambda^n}{n!}, \qquad n = 0, 1, 2, \ldots, \qquad \lambda > 0, \tag{2.42}$$

One property unique to the Poisson distribution is called *equidispersion*. The term *statistical dispersion* used in some environments is really just the variance, or its square root, the standard deviation, which measures the width of a distribution. As shown below, the parameter λ determines both the expectation value of counts n and its variance.

$$\kappa_1 = E(n) = \sum_{n=0}^{\infty} n\, p(n \,|\, \lambda) = \lambda, \tag{2.43}$$

$$\kappa_2 = \mathrm{var}(n) = E(n^2) - E(n)^2 = \lambda. \tag{2.44}$$

The equality of mean and variance defines equidispersion. Overdispersion, where $\mathrm{var}(n) > E(n)$ or underdispersion with $\mathrm{var}(n) < E(n)$ necessarily imply that the relevant probability must have two or more parameters.

Derivation

The Poisson distribution can be derived in at least two ways:

1. **Limit of Binomial Distribution:** If we take the limit $N \gg 1$ while at the same time making θ tends to zero such that $\lambda = N\theta$ remains constant [3], the binomial distribution equation (2.33) becomes

$$p(n|N,\lambda) = \frac{N!}{n!(N-n)!}\left(\frac{\lambda}{N}\right)^n \left(1 - \frac{\lambda}{N}\right)^{N-n}. \tag{2.45}$$

Since

$$\lim_{N\gg 1} \frac{N!}{(N-n)!} = \lim_{N\gg 1} N(N-1)\cdots(N-n+1) \simeq N^n \tag{2.46}$$

and

$$\lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N-n} = \lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N} = e^{-\lambda} \tag{2.47}$$

we obtain the expression of Poisson distribution

$$p(n \,|\, \lambda) = \frac{e^{-\lambda}\lambda^n}{n!}. \tag{2.48}$$

2. **Waiting-time derivation:** The *Poisson process* is fundamental in the field of stochastic time series. It is based on the two assumptions that there is a constant *rate* ω of events occurring and that the events are independent, no matter how close in time they occur. As we are not concerned with Poisson processes as such, the arguments below are not rigorous but only an intuitive motivation.

   Suppose that we would like to model the occurrence of random events that happen completely independently. The probability of an event occurring at any time t after starting observations at $t = 0$ is independent of the probability of another event occuring at any other time t', no matter how small the interval $(t' - t)$ is; in other words, each of the interval $\mathrm{d}t$ is a Bernoulli process.

   A typical example of the cumulative number of events as a function of time is illustrated in Figure 2.1.

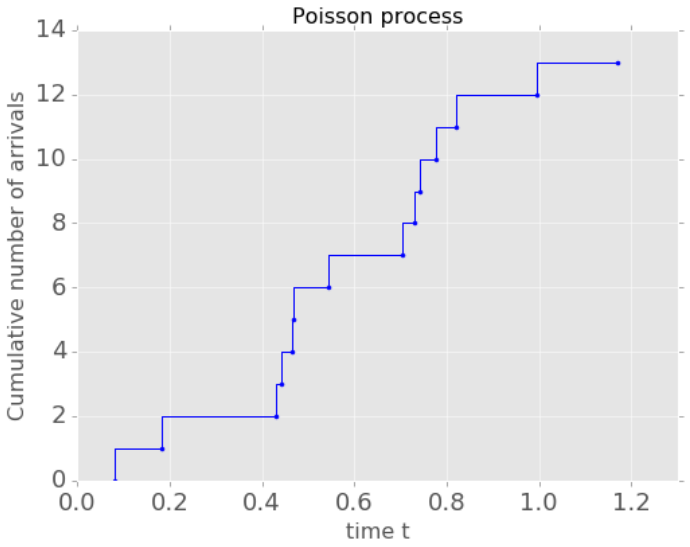


Figure 2.1: Example of a Poisson Process represented as a stochastic cumulative distribution. Given a constant rate ω, events happen randomly in time. The times of occurrence shown as the jump times on the x-axis and the cumulative number of events on the y-axis.

   Assuming the rate of arrival is ω, we assign the following probability values during an infinitesimal time interval $\mathrm{d}t$:

   (a) $p(n = 1|\mathrm{d}t, \omega) = \omega\,\mathrm{d}t$, the probability that one event takes place;

   (b) $p(n = 0|\mathrm{d}t, \omega) = 1 - \omega\,\mathrm{d}t$, the probability that no event occurs;

   (c) $p(n > 1|\mathrm{d}t, \omega) = 0$, the probability that more than one event is observed.

   Let $p(t)$ be the probability of no events occurring in time interval $(0, t)$, then the probability that zero counts are observed in time interval $(0, t+\mathrm{d}t)$ is, by using product

rule (2.5),

$$\begin{aligned} p(t+\mathrm{d}t) &= p(\text{null event in } (0,t) \text{ and null event in } (t,\mathrm{d}t)) \\ &= p(t)\Big(1-w\mathrm{d}t\Big) \quad \text{or} \quad \frac{dp(t)}{\mathrm{d}t} = -wp(t). \end{aligned} \tag{2.49}$$

With initial conditon $p(0) = 1$, the solution for no counts is an exponential [4],

$$p(n=0\,|\,t,w) = e^{-wt}. \tag{2.50}$$

The probability that exactly one event will occur at a time t_1 is a product of three components [5]: the probability that nothing happens between $(0,t_1)$, the probability one event occurs in the infinitesimal interval $(t_1, t_1+\mathrm{d}t_1)$, and the probability that there is no count in the final interval $(t_1+\mathrm{d}t_1, t)$,

$$\begin{aligned} p(n=1\,|\,t{=}t_1,\omega) &= p(0\,|\,(0,t_1),\omega)\; p(1\,|\,(t_1,t_1{+}dt_1),\omega)\; p(0\,|\,(t_1{+}dt_1,t),\omega) \\ &= e^{-\omega t_1}\,(\omega\,dt_1)\,e^{-\omega(t-t_1-dt)} \;=\; e^{-\omega t}e^{\omega dt_1}\omega\,dt_1. \end{aligned} \tag{2.51}$$

Hence the probability that this one event occurs at any time t_1 in the interval $(0,t)$ is

$$\begin{aligned} p(n=1|t,\omega) &= \int_0^t e^{-\omega t}e^{\omega dt_1}\omega\,dt_1 \\ &= \omega e^{-\omega t}\int_0^t e^{-\omega \mathrm{d}t_1}\,\mathrm{d}t_1 \;\approx\; \omega e^{-\omega t}\int_0^t \mathrm{d}t_1 \\ &= \omega t e^{-\omega t} \end{aligned} \tag{2.52}$$

In the case of n events occurring during time interval $(0,t)$, we must take into account time ordering. Using an abbreviated notation,

$$\begin{aligned} p(n|t,\omega) &= \int_0^t\int_0^{t_n}\cdots\int_0^{t_3}\int_0^{t_2} P(t_1)\,P(t_2-t_1)\,P(t_3-t_2)\cdots P(t-t_n)w^n\,\mathrm{d}t_n\cdots\mathrm{d}t_2\,\mathrm{d}t_1 \\ &= \omega^n e^{-\omega t}\int_0^t \mathrm{d}t_n\cdots\int_0^{t_3}\mathrm{d}t_2\int_0^{t_2}\mathrm{d}t_1 \end{aligned}$$

so that the result is a Poisson distribution with $\lambda = \omega t$,

$$p(n|\lambda) = \frac{e^{-\omega t}(\omega t)^n}{n!} = \frac{e^{-\lambda}\lambda^n}{n!} \tag{2.53}$$

Properties

The moment generating function of the Poisson distribution is

$$M(t) = E(e^{tn}) \;=\; \sum_{n=0}^{\infty} e^{tn}\frac{e^{-\lambda}\lambda^n}{n!} \;=\; e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^t)^n}{n!}$$

and summing up the series,

$$M(t) = \exp[\lambda(e^t - 1)] \tag{2.54}$$

so that the cumulant generating function is very simple,

$$K(t \,|\, n) = \ln M(t) = \lambda(e^t - 1). \tag{2.55}$$

so that the cumulants to all order are the same, $\kappa_q = \lambda, \forall\, q$. In particular, the mean and variance are equal,

$$\mu(n) = \mathrm{var}(n) = \lambda, \tag{2.56}$$

while skewness and kurtosis are suppressed by powers of $\lambda^{-1/2}$,

$$\gamma_1 = \lambda^{-1/2} \tag{2.57}$$
$$\gamma_2 = \lambda^{-1}. \tag{2.58}$$

The probability generating function is even simpler,

$$G(z) = E(z^n) = \sum_n \frac{e^{-\lambda}\lambda^n}{n!} z^n \;=\; e^{-\lambda} \sum_n \frac{(z\lambda)^n}{n!} \;=\; e^{\lambda z - \lambda}. \tag{2.59}$$

Properties of the Poisson with respect to Bayesian calculations will be treated in Section 3.2.2.1.

2.2.3 Negative Binomial

The Negative Binomial distribution (NBD) of integer counts n is a generalisation of the Poisson distribution and is obtained in a variety of different situations and scenarios. It is, for example, the probability of n successes given r failures,

$$\mathrm{Nb}(r, \theta) = p(n \,|\, r, \theta) = \binom{n + r - 1}{n} \theta^n (1 - \theta)^r \qquad n = 0, 1, 2, \ldots. \tag{2.60}$$

It has two parameters, a Bernoulli-type parameter θ with $0 \leq \theta \leq 1$ and secondly the parameter $r \geq 0$ which does not necessarily have to be an integer when interpreted as a shape parameter in Section 3.2.2.2, so the more general form is

$$p(n \,|\, r, \theta) = \frac{\Gamma(n + r)}{n!\Gamma(r)} \theta^n (1 - \theta)^r. \tag{2.61}$$

The NBD can be written in several different ways which has created much confusion. The Poisson-Gamma mixture will be treated in Chapter 3.

From Negative Binomial to Poisson Transforming from θ to a new parameter λ by the transformation $\theta = \lambda/(\lambda + r)$, in the limit $r \to \infty$ the NBD yields the Poisson distribution as shown below and in Figure 2.2:

$$\begin{aligned}
\lim_{r\to\infty} p(n \,|\, r, \lambda) &= \lim_{r\to\infty} \frac{\lambda^n}{n!} \frac{\Gamma(n+r)}{\Gamma(r)(r+\lambda)^n} \frac{1}{(1+\frac{\lambda}{r})^r} \\
&= \lim_{r\to\infty} \frac{\lambda^n}{n!} \frac{\Gamma(r)\cdot r\cdot (r+1)\cdots(r+n-1)}{\Gamma(r)(r+\lambda)^n} \frac{1}{(1+\frac{\lambda}{r})^r} \\
&\approx \frac{\lambda^n}{n!} \cdot 1 \cdot \frac{1}{e^\lambda} \\
&= \frac{e^{-\lambda}\lambda^n}{n!}
\end{aligned} \tag{2.62}$$

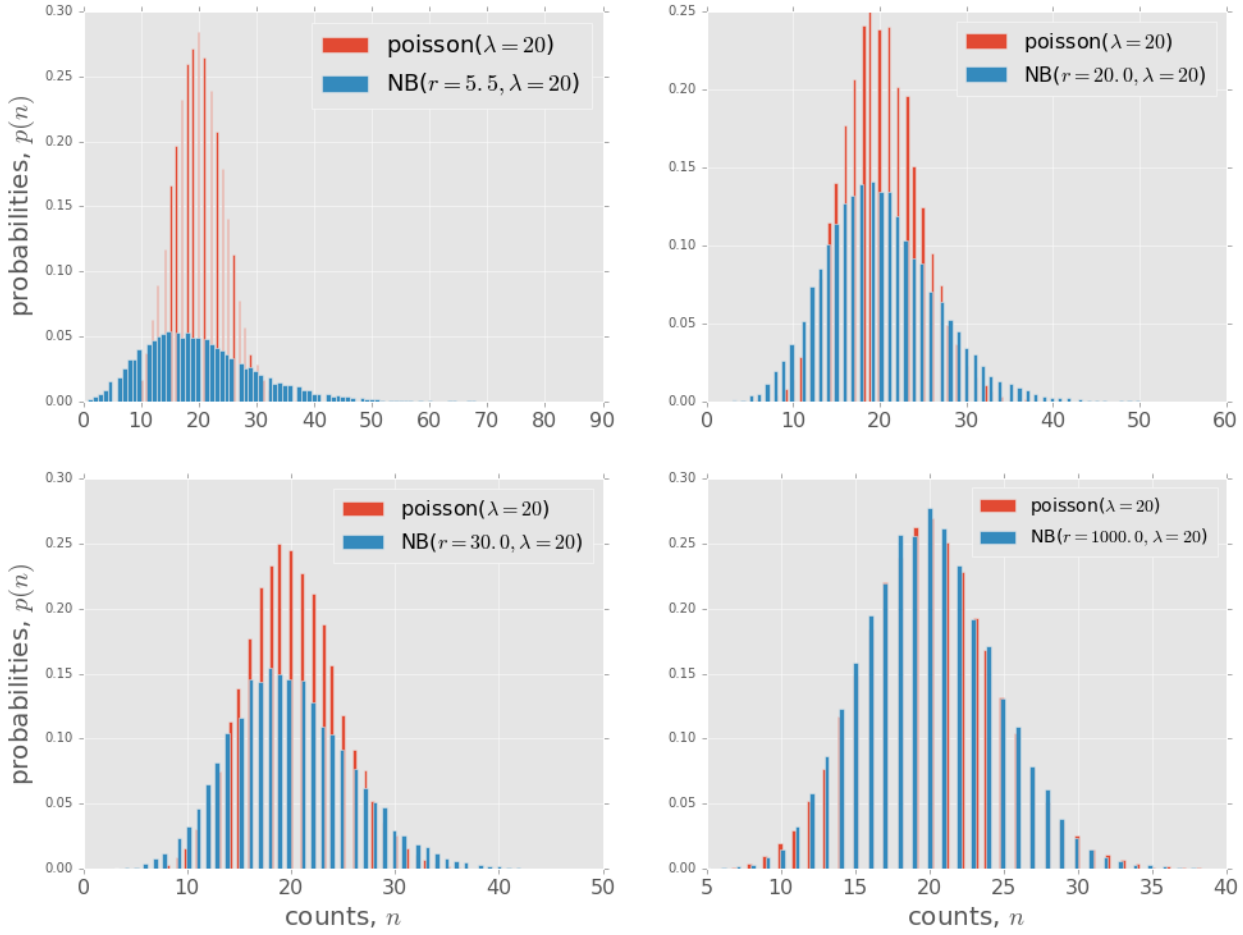


Figure 2.2: Comparison of the NBD and Poisson distributions for fixed λ and increasing r; note the different scales on the x-axis (actually the n-axis). For small r, the NBD has a much larger tail and hence variance and "dispersion" are greater than the Poisson distribution. As r increases, the tail shrinks until the two are indistinguishable.

Properties

The moment generating function is

$$M(t) = E(e^{tn}) = \sum_{n=0}^{\infty} e^{tn} \binom{n+r-1}{n} \theta^n (1-\theta)^r = (1-\theta)^r \sum_{n=0}^{\infty} \binom{n+r-1}{n} (\theta e^t)^n.$$

Using the negative binomial series expansion

$$(1-x)^{-r} = \sum_{n=0}^{\infty} \binom{n+r-1}{n} x^n \tag{2.63}$$

we obtain

$$M(t) = \big(1-\theta\big)^r \big(1-\theta e^t\big)^{-r} = \left(\frac{1-\theta}{1-\theta e^t}\right)^r. \tag{2.64}$$

From the cumulant generating function

$$K\big(t \,|\, P(n)\big) = \ln M\big(t \,|\, P(n)\big) = r \ln\left(\frac{1-\theta}{1-\theta e^t}\right) \tag{2.65}$$

we can derive the mean, variance, skewness and kurtosis,

$$\kappa_1 = \frac{r\theta}{1-\theta}, \tag{2.66}$$

$$\kappa_2 = \sigma^2 = \frac{r\theta}{(1-\theta)^2}, \tag{2.67}$$

$$\gamma_1 = \frac{1+\theta}{\sqrt{r\theta}}, \tag{2.68}$$

$$\gamma_2 = \frac{\theta^2 + 4\theta + 1}{r\theta}. \tag{2.69}$$

The variance of NB distribution provides a simpler way to show that the NBD approximates to the Poisson distribution in the limit of large r,

$$\mu(n) = \frac{r\theta}{1-\theta} = \lambda, \tag{2.70}$$

$$\text{var}(n) = \frac{r\theta}{(1-\theta)^2} = \lambda + \frac{\lambda^2}{r} \approx \lambda. \tag{2.71}$$

Following the same steps as for the mgf, we obtain the probability generating function

$$G(z) = E(z^n) = \left(\frac{1-\theta}{1-\theta z}\right)^r \tag{2.72}$$

for radius of convergence $|z| \leq 1/\theta$.

2.2.4 Gamma distribution

A gamma distributed random variable x has two free parameters, the shape parameter a and the inverse scale parameter c,

$$\mathrm{Ga}(a,c) \;=\; p(x\,|\,a,c) = \frac{c^a x^{a-1} e^{-cx}}{\Gamma(a)}, \quad x>0 \;\text{ and }\; a,c>0, \tag{2.73}$$

where $\Gamma(a)$ is the gamma function defined as

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\,\mathrm{d}t. \tag{2.74}$$

Normally these parameters are called (α,β) but we call them (a,c) to avoid notational confusion later.

Properties

The Gamma distribution results from the convolution of exponentially-distributed variables x, $p(x\,|\,c) = ce^{-cx}$. Since x is nonnegative, the Gamma distribution is suitable as a prior for the Poisson rate parameter; see Section 3.2.2.1.

The moment generating function is [6]

$$M(t) = \int_0^\infty e^{tx} \frac{c^a x^{a-1} e^{-cx}}{\Gamma(a)}\,\mathrm{d}x \;=\; \frac{1}{(1-t/c)^a}, \tag{2.75}$$

and the cumulant generating fuction is hence

$$K(t) = \ln M(t) = -a\ln(1-t/c). \tag{2.76}$$

The mean, variance, skewness and kurtosis are then,

$$\kappa_1 = \mu(x) = \frac{a}{c}, \tag{2.77}$$

$$\kappa_2 = \sigma^2 = \frac{a}{c^2}, \tag{2.78}$$

$$\gamma_1 = \frac{2}{\sqrt{a}}, \qquad \gamma_2 = \frac{6}{a}. \tag{2.79}$$

2.2.5 Gaussian

The Gaussian probability distribution, also called the normal distribution, is far and away the most widely used distribution for real variables, mainly because it results from the limiting form of most other distributions (including discrete ones) and also because it forms the basis for linear calculations done in physics and elsewhere.

The Gaussian contains two parameters, a location variable $\mu \in \mathbb{R}$ and a width or scale variable $\sigma > 0, \sigma \in \mathbb{R}$,

$$\mathcal{N}(\mu,\sigma^2) = p(x\,|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{2.80}$$

Properties

There is no scope to set out the many important properties of the Gaussian except to point out that, given the mgf,

$$M(t) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} + tx \right] \mathrm{d}x = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2 \right] \tag{2.81}$$

the cgf is particularly simple, being a quadratic polynomial with respect to t

$$K(t) = \ln M(t) = \mu t + \frac{1}{2}\sigma^2 t^2 \tag{2.82}$$

The mean, variance, skewness and kurtosis are then,

$$\kappa_1 = \mu, \tag{2.83}$$
$$\kappa_2 = \sigma^2, \tag{2.84}$$
$$\kappa_q = 0, \quad \text{for } q \geq 3 \tag{2.85}$$

The fact that all cumulants of higher order are identically zero forms the basis of many expansions, including the Laplace approximation treated below. The Central Limit Theorem set out in Section 2.3.2 shows how these expansions work.

2.3 Two important limits

These two limits are basic for most of statistics. Very importantly, they assume that the first and second cumulants exist; if they do not, then the entire structure of maths and statistics built on them collapses and other methods must be used. The Cauchy distribution is a prime example.

2.3.1 Law of Large Numbers

Let $\{X_i\}_{i=1}^N$ be a set of i.i.d. random variables. If X_i has finite mean $E(X_i) = \mu$ and variance $\mathrm{var}(X_i) = \sigma^2 < \infty, \forall i$, then as N approaches infinity, the sample average $\langle X \rangle$ converges to μ with probability 1. [7] [8]

$$\lim_{N\to\infty} p(\langle X \rangle \to \mu) = 1 \tag{2.86}$$

The proof runs something like this. The expected value and variance of average $\langle X \rangle$ are given as

$$E(\langle X \rangle) = E\left(\frac{1}{N} \sum_{i=1}^{N} X_i \right) = \frac{1}{N} E\left(\sum_{i=1}^{N} X_i \right) = \mu \tag{2.87}$$

$$\mathrm{var}(\langle X \rangle) = \mathrm{var}\left(\frac{1}{N} \sum_{i=1}^{N} X_i \right) = \frac{1}{N^2} \mathrm{var}\left(\sum_{i=1}^{N} X_i \right) = \frac{\sigma^2}{N}. \tag{2.88}$$

Chebychev's inequality for $\langle X \rangle$ states that

$$p\left(\left| \langle X \rangle - \mu \right| \geq \lambda \frac{\sigma}{\sqrt{N}} \right) \leq \frac{1}{\lambda^2} \tag{2.89}$$

where $\lambda > 1$. It is equivalent to say that the probability of the discrepancy betwee $\langle X \rangle$ and μ which is greater than λ units of the standard deviation $\sigma_{\langle X \rangle}$ is less than $\frac{1}{\lambda^2}$ [3] . In the limit $N \to \infty$, $\sigma_{\langle X \rangle}$ approaches zero and so $\langle X \rangle$ must converge to μ.

2.3.2 Central Limit Theorem

If mutually independent variables $\{X_i\}_{i=1}^N$ follow a common distribution $p(X_i)$ with population mean $E(X_i) = \mu$ and finite variance $\text{var}(X_i) = \sigma^2$, sample average $\langle X \rangle$ and variance of sample average $\text{var}(\langle X \rangle) = \sigma^2_{\langle X \rangle} = \sigma^2/N < \infty$, then the probability density function of standardised variable $X^* = \frac{\langle X \rangle - \mu}{\sigma_{\langle X \rangle}}$ tends to a standard Gaussian distribution

$$p(X^*) \to \mathcal{N}(0, 1) \tag{2.90}$$

Proof: The proof originates from [9]. The mgf and cgf of the standardised variable X^* are

$$M(t) = M(e^{tX^*}) = \exp\left(- \frac{\mu t}{\sigma_{\langle X \rangle}} \right) M\left(\frac{t}{\sigma_{\langle X \rangle}} \right) \tag{2.91}$$

$$K(t) = \ln M(t) = -\frac{\mu t}{\sigma_{\langle X \rangle}} + K\left(\frac{t}{\sigma_{\langle X \rangle}} \right) \tag{2.92}$$

then the first two cumulants are

$$\kappa_1(X^*) = -\frac{\mu}{\sigma_{\langle X \rangle}} + \frac{\kappa_1(\langle X \rangle)}{\sigma_{\langle X \rangle}} = 0 \tag{2.93}$$

$$\kappa_2(X^*) = \frac{\kappa_2(\langle X \rangle)}{\sigma^2_{\langle X \rangle}} = 1 \tag{2.94}$$

since $\kappa_1(\langle X \rangle) = E(\langle X \rangle) = \mu$ and $\kappa_2(\langle X \rangle) = \sigma^2_{\langle X \rangle}$, while cumulants of higher order $q > 2$ are

$$\kappa_q(X^*) = \frac{\kappa_q(\langle X \rangle)}{\sigma^q_{\langle X \rangle}} \tag{2.95}$$

Expanding the sample average,

$$\begin{aligned} \kappa_q(X^*) &= \frac{N^{-q}\kappa_q(\sum_i X_i)}{(\sigma/\sqrt{N})^q} \\ &= \frac{N^{1-q}\kappa_q(X)}{N^{-q/2}\kappa_2(X)} \\ &= \frac{1}{N^{q/2-1}} \frac{\kappa_q(X)}{\kappa_2(X)^{q/2}}, \quad \text{for } q = 2, 3, \cdots \end{aligned} \tag{2.96}$$

In the limit $N \to \infty$, all cumulants $\kappa_q(X^*)$ for $q \geq 3$ therefore tend to zero, and since the Gaussian higher-order cumulants are exactly zero, we conclude the probability distribution for the standandarised variable $p(X^*)$ approaches $\mathcal{N}(0, 1)$.

2.4 Laplace Approximations

Integrals are the bread and butter of inference and model comparison problems, but as the dimensionality of these integrals has grown, the so-called *curse of dimensionality* [10] has become a problem for numerical evaluation. The Laplace approximation offers an elegant method to calculate a first estimate of such integrals. With a view to later use we here define $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ as the variable. Assuming the nonnegative multivariate function $f(\boldsymbol{\theta})$ is strongly peaked at a global maximum $\boldsymbol{\theta}^*$ [11] determined by the first derivatives,

$$\left.\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*} = 0 \quad \text{or} \quad \left.\frac{\partial \log f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*} = 0.$$

Writing $f(\boldsymbol{\theta}) = e^{\log f(\boldsymbol{\theta})}$, and expanding the exponent up to the quadratic term, the argument goes that higher-order terms can be neglected due to the Central Limit Theorem,

$$\begin{aligned}\log f(\boldsymbol{\theta}) &= \log f(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\left.\frac{\partial \log f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*} + \frac{1}{2}\sum_i \sum_j \left.\frac{\partial^2 \log f(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right|_{\boldsymbol{\theta}^*} (\theta_i - \theta_i^*)(\theta_j - \theta_j^*) + \cdots \\ &\approx \log f(\boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T H(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \qquad (2.97)\end{aligned}$$

where $H(\boldsymbol{\theta}^*)$ is the $K \times K$ negative Hessian Matrix with $[i, j]$th element

$$H_{i,j}(\boldsymbol{\theta}^*) = -\left.\frac{\partial^2 \log f(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right|_{\boldsymbol{\theta}^*}$$

The function $f(\boldsymbol{\theta})$ is thereby approximated by a multivariate Gaussian and, extending the integration limits to infinity, we obtain the Laplace approximation for the integral of $f(\boldsymbol{\theta})$,

$$\begin{aligned}Z &= \int f(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = \int e^{\log f(\boldsymbol{\theta})}\, \mathrm{d}\boldsymbol{\theta} \\ &= f(\boldsymbol{\theta}^*) \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T H(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\} \mathrm{d}\boldsymbol{\theta} \\ &= f(\boldsymbol{\theta}^*) \sqrt{\frac{(2\pi)^K}{|\det H(\boldsymbol{\theta}^*)|}}. \qquad (2.98)\end{aligned}$$

Chapter 3

Bayesian parameter estimation and model comparison

3.1 Introduction to Bayesian probability theory

3.1.1 Probability of logic

The main difference between traditional probability theory and Bayesian probability is the concept of probability itself. In the 20th century in most fields, including physics, probability was understood as limited to predictions of "random variables" X or n given some fixed parameters, and since the variable was assumed to model only "objective reality" quantities, it was not allowed to think about probabilities of other quantities such as parameters, models etc. The Bayesian framework, on the other hand, allows one to think of the probability of any logical proposition whatsoever, including of course all "objective" measurements and quantities, but also any other logical proposition such as "It will rain tomorrow" or "The probability that parameter λ has some value" etc.

Allowing not just conventional "random variables" to appear in probability but any logical proposition results in a massive increase in the scope of the theory to range from the most common daily questions to finding the probability that the mass of the Higgs particle has some value.

At the basis of the Bayesian framework is therefore the *logical proposition* which is typically denoted by a capital letter. For example "A = *The moon is shining*" and its opposite "$\bar{A}$ = *The moon is not shining*" can each be assigned a probability such that $p(A)+p(\bar{A}) = 1$. Where there are more than two answers to a given logical proposition, or where two or more logical propositions are put simultaneously, the normal rules of probability theory apply, but they have a different or wider meaning.

A *logical proposition* is a declarative sentence that is either true or false, and the mathematics of such propositions is well developed. Most books on logic will provide a list of properties one can call *Boolean Algebra*, including

1. definition: "A" means *the logical proposition A is true*,
2. conjunction: "A AND B", denoted as A, B or just AB,
3. disjunction: "A OR B", written as $A + B$,

4. implication: "if A then B", written as $A \Rightarrow B$

5. negation: "A is false", is written as $\bar{A}$.

Basic boolean identities include

1. commutativity:
$$A, B = B, A$$
$$A + B = B + A$$

2. associativity:
$$A, (B, C) = (A, B), C = A, B, C$$
$$A + (B + C) = (A + B) + C = A + B + C$$

3. distributivity:
$$A, (B + C) = A, B + B, C$$
$$A + (B, C) = (A + B), (B + C)$$

4. duality:
$$\text{If } C = A, B, \quad \text{then } \bar{C} = \bar{A} + \bar{B}$$
$$\text{If } D = A + B, \quad \text{then } \bar{D} = \bar{A}, \bar{B}$$

There are many important issues to consider here at the level of logic, causality etc. For the purposes of this thesis, we only note that a logical proposition appearing to the *right* of the vertical line in a conditional probability is considered *true by hypothesis* becoming a "logical statement", meaning that for the purposes of the calculation it is considered true, while a logical proposition appearing to the *left* of the vertical line may be true or not. In other words $p(A \,|\, C)$ is the probability that proposition A is true while assuming for the moment that C is true.

Following the redefinition of probability in terms of logical proposition, the mathematics of probability follows from two cornerstones, namely the product rule and the sum rule. Let $p(A, B \,|\, C)$ be the probability of A and B being true, given an assumption that C is true.

Product Rule: The joint probability of A and B being true given C is the product of the probability that A is true (given C) times the probability that B is true (given A and C),

$$p(A, B \,|\, C) = p(A \,|\, C)\, p(B \,|\, A, C) \tag{3.1}$$

and of course the reverse product rule also holds

$$p(A, B \,|\, C) = p(B \,|\, C)\, p(A \,|\, B, C) \tag{3.2}$$

which shows that there is no "time" ordering involved but rather just an ordering in the logic sequence of arguments.

Sum Rule: While it may seem self-evident to most, the derivation of the sum rule

$$p(A \,|\, C) + p(\bar{A} \,|\, C) = 1 \tag{3.3}$$

requires a thorough discussion; see for example [12]. From this, the Product Rule and the rules of Boolean algebra follows the *generalised sum rule*

$$p(A + B \,|\, C) = p(A \,|\, C) + p(B \,|\, C) - p(A, B \,|\, C) \tag{3.4}$$

and almost all of the usual probability theory.

Information: A crucial consequence of widening the scope of probability was that *information* $\mathcal{I}$ about a given situation or background or experiment could be taken into account. Given a logical proposition A, two observers may assign different probabilities to it if their information on the situation differs. For example, the proposition *the moon is shining* would have a different probability given information $\mathcal{I}_1 =$ *the night is cloudy* and information $\mathcal{I}_2 =$ *there are no clouds*. It hence becomes necessary to state such information explicitly by including it in the probability, $p(A \,|\, \mathcal{I}_1) \neq p(A \,|\, \mathcal{I}_2)$.

Hypotheses: Besides the explicit specification of available information, the use of any hypothesis $\mathcal{H}$ used must be indicated in the probability notation. A hypothesis is a set of statements which are taken to be true only for the time being, while the probability is being calculated, without having to be true or false. This opens up the way to make quantitative comparison of different hypotheses using Bayesian probability theory; see for example Sections 3.1.6 and 3.3.

3.1.2 Bayesian Inference

Inverting conditional probabilities of logical propositions: Combining the two ways of using the product rule for logical statements in Eqs. (3.1) and (3.2) and setting $C = \mathcal{H}$, the two conditional probabilities $p(A \,|\, B, \mathcal{H})$ and $p(B \,|\, A, \mathcal{H})$ are related by

$$p(A \,|\, B, \mathcal{H}) = \frac{p(B \,|\, A, \mathcal{H})\, p(A \,|\, \mathcal{H})}{p(B \,|\, \mathcal{H})} \tag{3.5}$$

The denominator can be written to resemble the numerator. If A has only two possible answers A and $\bar{A}$, then

$$\begin{aligned} p(B \,|\, \mathcal{H}) &= p(A, B \,|\, \mathcal{H}) + p(\bar{A}, B \,|\, \mathcal{H}) && (3.6)\\ &= p(B \,|\, A, \mathcal{H})\, p(A \,|\, \mathcal{H}) + p(B \,|\, \bar{A}, \mathcal{H})\, p(\bar{A} \,|\, \mathcal{H}), && (3.7) \end{aligned}$$

and so

$$p(A \,|\, B, \mathcal{H}) = \frac{p(B \,|\, A, \mathcal{H})\, p(A \,|\, \mathcal{H})}{p(B \,|\, A, \mathcal{H})\, p(A \,|\, \mathcal{H}) + p(B \,|\, \bar{A}, \mathcal{H})\, p(\bar{A} \,|\, \mathcal{H})}, \tag{3.8}$$

By extension, if A has K different possible answers A_k which are mutually exclusive then

$$p(A_k \,|\, B, \mathcal{H}) = \frac{p(B \,|\, A_k, \mathcal{H})\, p(A_k \,|\, \mathcal{H})}{\sum_k p(B \,|\, A_k, \mathcal{H})\, p(A_k \,|\, \mathcal{H})}. \tag{3.9}$$

Application to parameters and data: We note again that the above relations are completely general and can be applied to any logical propositions. For the specific work in this thesis, we specialise to say that A becomes the statement *The parameter θ has a particular (real) value*, and identify B with the *experimental data* which we call D for the moment and which will later be the set of counts n_b. Rewriting Eq. (3.5) in terms of $A = \theta$ and $B = D$, we have Bayes' Theorem

$$p(\theta \,|\, D, \mathcal{H}) = \frac{p(D \,|\, \theta, \mathcal{H})\, p(\theta \,|\, \mathcal{H})}{p(D \,|\, \mathcal{H})}. \tag{3.10}$$

This is so important that each of the four probabilities has its own name:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}. \tag{3.11}$$

We discuss each one:

- The *prior* $p(\theta \,|\, \mathcal{H})$ is the probability density of parameter θ before any data is taken into account. It is based on pre-existing information or studies beforehand.
- The *likelihood* $p(D \,|\, \theta, \mathcal{H})$ is the probability usually associated with probability theory; it predicts the probability that a particular set of data D would be obtained under the hypothesis $\mathcal{H}$ and for a particular value of the parameter within that hypothesis.
- The *evidence* or *marginal likelihood* $p(D \,|\, \mathcal{H})$ is at first sight just a normalisation constant, which for continuous θ is just the integral over θ of the numerator,

  $$p(D \,|\, \mathcal{H}) = \int p(D, \theta \,|\, \mathcal{H})\, \mathrm{d}\theta = \int p(D \,|\, \theta, \mathcal{H})\, p(\theta \,|\, \mathcal{H})\, \mathrm{d}\theta \tag{3.12}$$

  in analogy with Eq.(3.9). In discrete systems, the integral may be replaced by a summation. For continuous θ, Bayes' Theorem can hence also be written as

  $$p(\theta \,|\, D, \mathcal{H}) = \frac{p(D \,|\, \theta, \mathcal{H})\, p(\theta \,|\, \mathcal{H})}{\int p(D \,|\, \theta, \mathcal{H})\, p(\theta \,|\, \mathcal{H})\, \mathrm{d}\theta}. \tag{3.13}$$

  Beyond being a normalisation constant, the evidence is a stepping stone towards comparing different models or hypotheses; see Section 3.1.6.
- The *posterior* $p(\theta \,|\, D, \mathcal{H})$ is the conditional probability distribution of θ based both on the hypothesis and the data obtained. The posterior is obviously proportional to the prior times likelihood by neglecting the constant denominator

  $$p(\theta \,|\, D, \mathcal{H}) \propto p(D \,|\, \theta, \mathcal{H})\, p(\theta \,|\, \mathcal{H}) \tag{3.14}$$

  which means that if we are interested only in sampling from the posterior, we do not have to calculate the evidence. This will be important in Chapter 4.

Bayes' Theorem can therefore be read as a process where the *prior* probability of θ is updated by information based on the data to give the *posterior*. This provides a powerful framework for continuously using new data to update the probability of θ.

Clearly the above definitions and relations apply not just to one parameter θ but to as many as we want.

Coin flip example: To illustrate the Bayesian approach, consider the example of repeated coin flips set out in many books such as [12]. Given the number of flips N and a known single-flip probability for heads θ, the likelihood for n heads is clearly a Binomial distribution

$$p(D\,|\,\theta,\mathcal{H}) = p(n\,|\,\theta,N,\mathcal{H}) = \binom{N}{n}\theta^n(1-\theta)^{N-n} = \frac{N!}{n!(N-n)!}\theta^n(1-\theta)^{N-n} \tag{3.15}$$

where θ is a parameter while N is a constant. Given more specific information, it is reasonable to be completely unbiased about the prior choice for θ; we give any value between 0 and 1 the same prior probability so that the prior is a uniform distribution,

$$p(\theta\,|\,N,\mathcal{H}_U) = U(0,1) = 1, \qquad 0<\theta<1. \tag{3.16}$$

Inserting this into Bayes' Theorem results in a posterior

$$p(\theta\,|\,n,N,\mathcal{H}_U) \;=\; \frac{\binom{N}{n}\theta^n\,(1-\theta)^{N-n}\cdot 1}{\int_0^1 d\theta\,\binom{N}{n}\theta^n\,(1-\theta)^{N-n}\cdot 1} \;=\; \frac{(N+1)!}{n!(N-n)!}\,\theta^n\,(1-\theta)^{N-n}. \tag{3.17}$$

Note that in this posterior n is constant while θ is variable, while in the likelihood (3.15) n is the variable while θ is constant. The posterior has the form of a Beta distribution for a variable $0\le x\le 1$,

$$\mathrm{Be}(x\,|\,\alpha,\beta) = p(x\,|\,\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{3.18}$$

$$\text{with}\qquad B(\alpha,\beta) = \int_0^1 dx\,x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!}\;; \tag{3.19}$$

clearly (3.18) is identical with (3.17) on identifying $\alpha = n+1$ and $\beta = N-n+1$ and $x=\theta$,

$$p(\theta\,|\,n,N,\mathcal{H}_U) = \mathrm{Be}(\theta\,|\,n{+}1,N{-}n{+}1). \tag{3.20}$$

While the uniform prior (3.16) would be appropriate when we know nothing about θ beforehand, the same problem can be treated with a different prior when we have, for example, information on previous results. In that case we would use the Beta distribution also as a prior,

$$p(\theta\,|\,\alpha,\beta,\mathcal{H}_B) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

but with α and β now treated as *hyperparameters* (i.e. pre-parameters whose value changes the shape of the prior if not its functional form). Inserting this prior into Bayes' Theorem along with the Binomial likelihood (3.15), we obtain a different posterior

$$p(\theta\,|\,n,N,\mathcal{H}_B) = \frac{\theta^{n+\alpha-1}(1-\theta)^{N-n+\beta-1}}{B(n+\alpha,N-n+\beta)} = \mathrm{Be}(n+\alpha,N-n+\beta)$$

which is the same as the uniform prior case for $\alpha=\beta=1$.

Given two different priors, we obtain two different posteriors reflecting the different information. This is as it should be. When data (in this case the head counts summing to n)

are independent, the effect of the prior tends to zero for large N, n, so that in the large limit the posteriors are the same for all priors.

The dependence on priors is one reason why Laplace's probability theory was ignored for a century, because in the frequentist world view, probabilities are not assigned but emerge "naturally" from many repetitions of an experiment. In the Bayesian view, however, the question of prior information is made explicit and hence explicit thought must be given to it. For independent data, the probability also "naturally" converges to the correct one.

In Figure 3.1, inspired by [13], we show the results of a simulation in which data n was created with a simulation parameter $\theta = 0.5$. With increasing number of trials N and head counts n, the posterior reflects that one has information to pinpoint the correct value $\theta = 0.5$ with more and more accuracy.

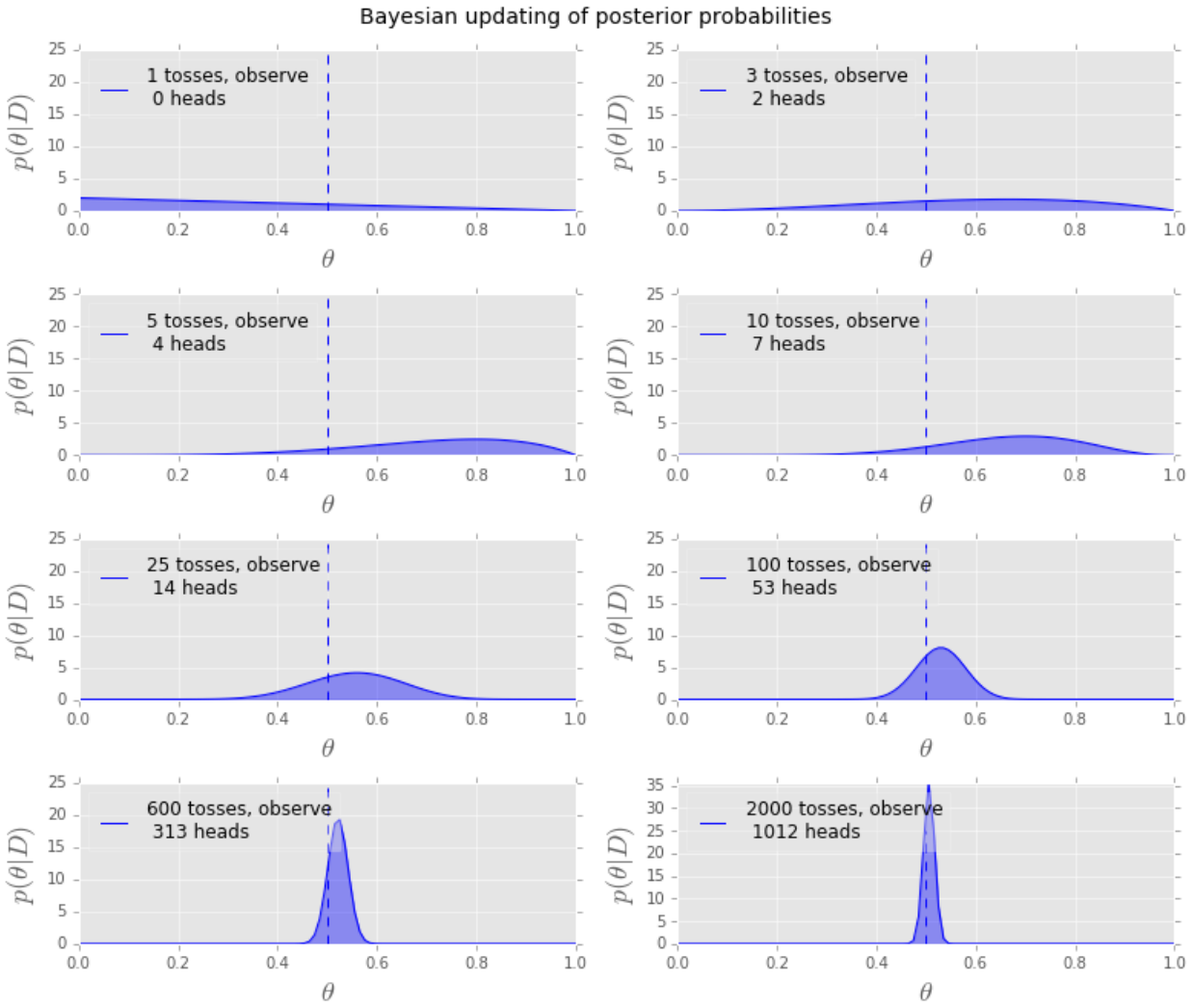


Figure 3.1: Change of the posterior for θ with increasing N using a uniform prior. As N becomes large, the likelihood "wipes out" the effect of the prior. The width of the posterior decreases with N and correctly has its mode closer and closer to 0.5.

3.1.3 Parameter Estimation

Obtaining an analytical answer for the posterior as in the above coin flip example is the ideal case; the posterior contains all relevant information on the parameters. As already indicated in Chapter 2, it is often convenient to characterise the posterior (or the likelihood in some cases) by the mean, variance and interval probabilities. In this section, we set out a few issues relating to parameter estimation.

Marginal posteriors and nuisance parameters: For a K-dimensional posterior distribution, if we are concerned only about parameter θ_1, the marginal posterior distribution $p(\theta_1 \,|\, D, \mathcal{H})$ is obtained through integration over the remaining $K-1$ parameters,

$$p(\theta_1 \,|\, D, \mathcal{H}) = \int p(\theta_1, \theta_2, \ldots, \theta_K \,|\, D, \mathcal{H}) \, \mathrm{d}\theta_2 \, \mathrm{d}\theta_3 \, \ldots \, \mathrm{d}\theta_K \tag{3.21}$$

Parameters which are integrated out are usually called *nuisance parameters*.

MAP: The best value of a parameter θ is usually taken to be its mode value θ^* at the maximum of the posterior. This is often called the *Maximum A Posteriori* (MAP) estimation, $\underset{\theta}{\operatorname{argmax}} P(\theta \,|\, D, \mathcal{H})$ or the logarithm $\underset{\theta}{\operatorname{argmax}} \log P(\theta \,|\, D, \mathcal{H})$. Frequentist statistics has only the likelihood and would calculate the *Maximum Likelihood* (ML) estimator, $\underset{\theta}{\operatorname{argmax}} P(D \,|\, \theta, \mathcal{H})$ or $\underset{\theta}{\operatorname{argmax}} \log P(D \,|\, \theta, \mathcal{H})$. Whenever the prior distribution is uniform, MAP is identical to ML.

Predictions: To make predictions for future observations, based on a model (likelihood, prior) and data, we can compute the *posterior predictive distribution* of new outcome x' conditional on previous observed data,

$$\begin{aligned} p(x' \,|\, D, \mathcal{H}) &= \int p(x', \theta \,|\, D, \mathcal{H}) \, \mathrm{d}\theta = \int p(x' \,|\, \theta, D, \mathcal{H}) p(\theta \,|\, D, \mathcal{H}) \, \mathrm{d}\theta \\ &= \int p(x' \,|\, \theta, \mathcal{H}) p(\theta \,|\, D, \mathcal{H}) \, \mathrm{d}\theta, \end{aligned} \tag{3.22}$$

i.e. Bayesian prediction is an integral over the likelihood for x' weighted by the posterior of the parameters. For example, in the coin toss problem, the prediction that the next flip is a head is, with likelihood $p(\text{head} \,|\, \theta) = \theta$, given by

$$p(\text{head} \,|\, D, \mathcal{H}) = \int \theta \, p(\theta \,|\, n, N, \mathcal{H}_B) \, \mathrm{d}\theta = \frac{n + \alpha}{N + \alpha + \beta}$$

Credible interval: This corresponds to the generic central interval in Chapter 2.1.2. Interval estimation is made by investigating a credible set $\mathcal{A}$ for a given quantile $1 - \alpha$,

$$\int_{\mathcal{A}(\theta)} p(\theta \,|\, D, \mathcal{H}) \, \mathrm{d}\theta = 1 - \alpha$$

The appropriate credible interval is found by scanning the posterior from left to right, with $\alpha/2$ on each tail.

Highest posterior density (HPD) interval: The Highest Posterior Density or HPD interval can be viewed as traversing the posterior distribution by a horizontal line decreasing progressively from top to bottom, such that the minimum density of any point within the HPD interval is never lower than the density of any point outside. Given a threshold probability p^*, the HPD interval is defined as the set of all points θ whose probability exceeds p^*, $\mathcal{A}(\theta \,|\, p^*) = \{\theta \,|\, p(\theta \,|\, D, \mathcal{H}) > p^*\}$. The total probability mass contained in the HPD interval is then

$$\int_{\mathcal{A}(\theta \,|\, p^*)} p(\theta \,|\, D, \mathcal{H}) \, \mathrm{d}\theta = 1 - \alpha(p^*). \tag{3.23}$$

with α dependent on p^*. Unlike the central interval, the HPD would not necessarily contain $\alpha/2$ probability in the tails of a single mode distribution. It is more robust so that it can handle two or more peaks.

3.1.4 Assigning Probabilities

Deciding which distribution to use for the likelihood and the prior is part of the process of model-building summarised in the symbol $\mathcal{H}$. There are different guidelines to do so.

Principle of Indifference: The "Principle of Indifference" is the simplest way to assign probabilities; it states that information which is invariant under any permutation of propositions implies assigning equal probability to all. If there is only information that there are B possible outcomes to a particular discrete variable n, then the prior should be uniform, $p(n \,|\, B, \mathcal{H}_U) = 1/B$; similarly when θ can take on all values in an interval $[\theta_{min}, \theta_{max}]$, the prior should be the uniform distribution (2.11),

$$p(\theta \,|\, \mathcal{H}) = U(\theta \,|\, \theta_{\min}, \theta_{\max}) = \frac{1}{\theta_{\max} - \theta_{\min}}.$$

The Principle of Indifference is fundamental; however, it is insufficient when additional constraints beyond the minimum and maximum are involved.

Conjugate priors: A second method for assigning probabilities involves *conjugate priors*. Given some likelihood $p(x \,|\, \theta, \mathcal{H})$, a prior $p(\theta \,|\, \alpha, \mathcal{H})$ is conjugate to $p(x \,|\, \theta, \mathcal{H})$ if the posterior follows the same distribution as the prior but with an updated parameter α',

$$p_c(\theta \,|\, x, \alpha', \mathcal{H}) \propto p_c(\theta \,|\, \alpha, \mathcal{H}) \, p(x \,|\, \theta, \mathcal{H}) \tag{3.24}$$

with p_c the same distribution. Conjugate prior-likelihood pairs are very convenient because updates just change the values of the parameters.

There is a known list of conjugate pairs, see e.g. [14]. We already saw that the Beta distribution is a conjugate prior for the Binomial distribution; later we will see that the Gamma distribution is the conjugate prior for the Poisson likelihood etc.

Maximum entropy The third method called *Principle of Maximum Entropy* makes some progress towards incorporating constraints as E.T.Jaynes suggested in 1963 [15]. For discrete

variables θ, Shannon's entropy for the prior is defined as

$$H(p) = -\sum_{\theta \in \mathcal{A}_\theta} p(\theta \,|\, \mathcal{H}) \log p(\theta \,|\, \mathcal{H}). \tag{3.25}$$

It measures uncertainty or information content of the discrete probability distribution of proposition θ, where $\mathcal{H}$ here includes background information which constrains the assignment of probabilities. Since $H(p)$ is the measurement of ignorance in the probability distribution, maximum entropy provides the probability distribution solution which is least informative while remaining consistent with constraints introduced by the information in $\mathcal{H}$ [16]. The principle of maximum entropy is implemented by introducing Lagrange multipliers to satisfy constraints set by information $\mathcal{H}$ and applying variational calculus to the constrained system

$$H = -\sum_{\theta \in \mathcal{A}_\theta} p(\theta \,|\, \mathcal{H}) \log p(\theta \,|\, \mathcal{H}) + (\lambda_0 - 1) \sum_{\theta \in \mathcal{A}_\theta} p(\theta \,|\, \mathcal{H}) + \sum_\alpha \lambda_\alpha \phi_\alpha(p), \tag{3.26}$$

where the second term is the normalization constraint and the third term $\phi_\alpha(p)$ represents constraints due to information [1] in the form $\phi_\alpha(p) = 0, (\alpha = 1, 2 \ldots)$. The maximum entropy solution is given by solving the partial derivatives with respect to $p(\theta \,|\, \mathcal{H})$,

$$\frac{\partial H}{\partial p(\theta \,|\, \mathcal{H})} = -\log p(\theta \,|\, \mathcal{H}) - 1 + \lambda_0 + \frac{\partial}{\partial p(\theta \,|\, \mathcal{H})} \sum_\alpha \lambda_\alpha \phi_\alpha(p) = 0 \tag{3.27}$$

where λ_0 and λ_α subject to normalization of $p(\theta \,|\, \mathcal{H})$.

3.1.5 Asymptotic Analysis

When there is no data available, the only probability available to the observer is the prior. However, as data accumulates, the application of Bayes' theorem again and again results in posteriors which become more and more peaked, as shown in the above Binomial example. This means that, unless the prior is somehow very restrictive or unrealistic, the information gained from data overrides any prior beliefs, the influence of a particular prior on Bayesian inference diminishes, and Bayesian parameter probabilities become dependent mostly on the likelihood only. In this sense and under the assumptions that the accumulating data is well-behaved (e.g. obeys the Law of Large Numbers) and that the prior is not too restrictive, the Bayesian and frequentist results converge.

Moreover, due to the Central Limit Theorem, the analysis of asymptotic distributions simplifies to consideration of only the quadratic approximation around an extremum (maximum or minimum).

In this section, we briefly consider two asymptotic analysis results for the posterior, namely an expression in terms of the Fisher Information and a Laplace Approximation of expectations.

Posterior in terms of Fisher Information matrix: The following derivation is based on [14]. For a sequence of observables $\{x_m\}_{m=1}^N$, the posterior can be approximated as

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) &\propto p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \,|\, \mathcal{H}) \\ &\propto \exp\Big\{ \log p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) + \log p(\boldsymbol{\theta} \,|\, \mathcal{H}) \Big\} \end{aligned} \tag{3.28}$$

Given the modes of the prior and likelihood (where "MLE" stands for "Maximum Likelihood Estimator" in the literature),

$$\nabla \log p(\boldsymbol{\theta}^*_{\text{prior}} \,|\, \mathcal{H}) = 0, \quad \nabla \log p(D \,|\, \boldsymbol{\theta}^*_{\text{MLE}}, \mathcal{H}) = 0 \tag{3.29}$$

we expand the two logarithms around the modes,

$$\log p(\boldsymbol{\theta} \,|\, \mathcal{H}) = \log p(\boldsymbol{\theta}^*_{\text{prior}} \,|\, \mathcal{H}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{prior}})^T H^*_{\text{prior}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{prior}}) + \cdots \tag{3.30}$$

$$\log p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) = \log p(D \,|\, \boldsymbol{\theta}^*_{\text{MLE}}, \mathcal{H}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{MLE}})^T H^*_{\text{MLE}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{MLE}}) + \cdots \tag{3.31}$$

The $[i, j]$th element of prior and likelihood negative Hessian matrices are

$$H^*_{ij,\text{prior}} = \left(- \frac{\partial^2 \log p(\boldsymbol{\theta} \,|\, \mathcal{H})}{\partial \theta_i \partial \theta_j} \right) \Bigg|_{\boldsymbol{\theta}^*_{\text{prior}}} \tag{3.32}$$

$$H^*_{ij,\text{MLE}} = \left(- \frac{\partial^2 \log p(D \,|\, \boldsymbol{\theta}, \mathcal{H})}{\partial \theta_i \partial \theta_j} \right) \Bigg|_{\boldsymbol{\theta}^*_{\text{MLE}}} \tag{3.33}$$

therefore, the posterior can be approximated as [14]

$$p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) \propto \exp \left\{ - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{prior}})^T H^*_{\text{prior}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{prior}}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{MLE}})^T H^*_{\text{MLE}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\text{MLE}}) \right\} \tag{3.34}$$

$$\propto \exp \left\{ - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T H^* (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\} \tag{3.35}$$

with

$$H^* = H^*_{\text{prior}} + H^*_{\text{MLE}} \tag{3.36}$$

$$\boldsymbol{\theta}^* = H^{*-1} \left(H^*_{\text{prior}} \boldsymbol{\theta}^*_{\text{prior}} + H^*_{\text{MLE}} \boldsymbol{\theta}^*_{\text{MLE}} \right) \tag{3.37}$$

The posterior of $\boldsymbol{\theta}$ approximates a multivariate Gaussian distribution

$$p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) \sim \mathcal{N}(\boldsymbol{\theta}^*, H^{*-1}) \tag{3.38}$$

where $\mu = \boldsymbol{\theta}^*$ and the covariance matrix $\boldsymbol{\Sigma}$ is H^{*-1}. As shown, the likelihood becomes more and more peaked with increasing N while the prior remains the same. The magnitude of elements of the Hessian H^*_{MLE} will hence also be much larger than those of H^*_{prior}, so that one can neglect the latter.

When the data is made up of N i.i.d. measurements, $D = \{x_m\}_{m=1}^N$, the H^*_{MLE}/N tends to converge to a constant, so that the large-N average of each matrix element becomes

$$\lim_{N\to\infty} \frac{(H^*_{\text{MLE}})_{ij}}{N} = \lim_{N\to\infty} \left\{ \frac{1}{N} \left(- \frac{\partial^2 \log p(D \,|\, \boldsymbol{\theta}, \mathcal{H})}{\partial \theta_i \partial \theta_j} \right) \right\} = \lim_{N\to\infty} \left\{ \frac{1}{N} \sum_{m=1}^{N} \left(- \frac{\partial^2 \log p(x_m \,|\, \boldsymbol{\theta}, \mathcal{H})}{\partial \theta_i \partial \theta_j} \right) \right\}$$

Under the assumption that the sample average converges to the expectation value 2.86, this converges to the expectation value

$$\lim_{N\to\infty} \frac{(H^*_{\mathrm{MLE}})_{ij}}{N} = \int p(x \,|\, \boldsymbol{\theta}, \mathcal{H}) \left(- \frac{\partial^2 \log p(x \,|\, \boldsymbol{\theta}, \mathcal{H})}{\partial\theta_i \partial\theta_j} \right) \mathrm{d}x \tag{3.39}$$

and so H^*_{MLE} converges to N times the *Fisher information matrix*

$$\lim_{N\to\infty} H^*_{\mathrm{MLE}} = N\, I(\boldsymbol{\theta}) \tag{3.40}$$

$$\text{with} \quad I(\boldsymbol{\theta})_{ij} = \int p(x \,|\, \boldsymbol{\theta}, \mathcal{H}) \left(- \frac{\partial^2 \log p(x \,|\, \boldsymbol{\theta}, \mathcal{H})}{\partial\theta_i \partial\theta_j} \right) \mathrm{d}x, \tag{3.41}$$

and the posterior distribution converges to a multivariate Gaussian

$$p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) \sim \mathcal{N}_K(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^*, [H^*_{\mathrm{MLE}}]^{-1}) = \mathcal{N}_K(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^*, [N\, I(\boldsymbol{\theta})]^{-1}) \tag{3.42}$$

Asymptotic expectation values: Apart from the standard mode, mean, variance and interval characteristics of the posterior, it is sometimes necessary to calculate expectation values of some function $f(\boldsymbol{\theta})$ of the parameters,

$$E[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, D, \mathcal{H})\, \mathrm{d}\boldsymbol{\theta}. \tag{3.43}$$

Using the Bayes' Theorem expression

$$p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) = \frac{p(D \,|\, \boldsymbol{\theta}, \mathcal{H})\, p(\boldsymbol{\theta} \,|\, \mathcal{H})}{\int p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \,|\, \mathcal{H})\, \mathrm{d}\boldsymbol{\theta}}$$

the expectation value of function $f(\boldsymbol{\theta})$ can be rewritten asymptotically as the ratio of two Laplace approximation integrals

$$E[f(\boldsymbol{\theta})] = \frac{\int f(\boldsymbol{\theta}) p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \,|\, \mathcal{H})\, \mathrm{d}\boldsymbol{\theta}}{\int p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \,|\, \mathcal{H})\, \mathrm{d}\boldsymbol{\theta}} = \frac{\int \exp\left\{A_1(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}}{\int \exp\left\{A_2(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}} \tag{3.44}$$

with exponents expanded in Taylor series around the mode,

$$\begin{aligned} A_1(\boldsymbol{\theta}) &= \log f(\boldsymbol{\theta}) + \log p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) + \log p(\boldsymbol{\theta} \,|\, \mathcal{H}) \\ &= A_1(\boldsymbol{\theta}^*_1) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_1)^T H_1(\boldsymbol{\theta}^*_1)(\boldsymbol{\theta} - \boldsymbol{\theta}^*_1) + \cdots \end{aligned} \tag{3.45}$$

$$\begin{aligned} A_2(\boldsymbol{\theta}) &= \log p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) + \log p(\boldsymbol{\theta} \,|\, \mathcal{H}) \\ &= A_2(\boldsymbol{\theta}^*_2) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_2)^T H_2(\boldsymbol{\theta}^*_2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*_2) + \cdots \end{aligned} \tag{3.46}$$

where $H_1 = -\nabla\nabla A_1(\boldsymbol{\theta}^*_1)$, $H_2 = -\nabla\nabla A_2(\boldsymbol{\theta}^*_2)$ are the negative $K \times K$ Hessian matrices. Under the Laplace approximation assumptions, we then obtain

$$E[f(\boldsymbol{\theta})] \simeq \frac{\exp(A_1(\boldsymbol{\theta}^*_1)) \sqrt{\frac{(2\pi)^K}{|\det H_1(\boldsymbol{\theta}^*_1)|}}}{\exp(A_2(\boldsymbol{\theta}^*_2)) \sqrt{\frac{(2\pi)^K}{|\det H_2(\boldsymbol{\theta}^*_2)|}}} = \exp[A_1(\boldsymbol{\theta}^*_1) - A_2(\boldsymbol{\theta}^*_2)] \sqrt{\frac{|\det H_2(\boldsymbol{\theta}^*_2)|}{|\det H_1(\boldsymbol{\theta}^*_2)|}} \tag{3.47}$$

Asymptotic Evidence: Using the same arguments and techniques, the Laplace approximation for the evidence yields a closed form,

$$p(D \mid \mathcal{H}) = \int p(D \mid \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \mid \mathcal{H}) \, \mathrm{d}\boldsymbol{\theta} \tag{3.48}$$

$$\simeq \exp(A_2(\boldsymbol{\theta}_2^*)) \sqrt{\frac{(2\pi)^K}{|\det H_2(\boldsymbol{\theta}_2^*)|}}$$

$$= p(D \mid \boldsymbol{\theta}^*, \mathcal{H}) \, p(\boldsymbol{\theta}^* \mid \mathcal{H}) \sqrt{\frac{(2\pi)^K}{|\det H_2(\boldsymbol{\theta}_2^*)|}} \tag{3.49}$$

While this looks like a quick and easy solution, it also makes apparent that the evidence depends explicitly on the prior. The posterior can be shown to become asymptotically independent of the prior, but the evidence does not. In the results of Chapter 5, we shall see how the choice of prior influences the numerical values of various evidences.

3.1.6 Model comparison

The purpose of model comparison is to choose the most probable model in the light of the data and background information, i.e. finding the largest probability among $p(\mathcal{H}_1 \mid D)$, $p(\mathcal{H}_2 \mid D)$ etc.

Frequentist statistics does not permit writing $p(\mathcal{H} \mid D)$ but calculates everything based on the likelihood $p(D \mid \theta, \mathcal{H})$. Therefore, the application of Bayes' theorem to calculate probabilities of hypotheses is rejected [2]. In the Bayesian framework, on the other hand, this becomes a simple matter. Given two competing hypotheses, one would take ratios of their respective probabilities, also called *odds*. For any pair of models $\mathcal{H}_i$ and $\mathcal{H}_j$, the odds

$$\mathcal{O}_{ij} = \frac{p(\mathcal{H}_i \mid D)}{p(\mathcal{H}_j \mid D)} \tag{3.50}$$

can, with the help of Bayes' Theorem applied to $\mathcal{H}$ and D, be translated into ratios of hypothesis priors and evidences:

$$\mathcal{O}_{ij} = \frac{p(D \mid \mathcal{H}_i) \, p(\mathcal{H}_i)}{p(D)} \Bigg/ \frac{p(D \mid \mathcal{H}_j) \, p(\mathcal{H}_j)}{p(D)}$$

$$= \frac{p(D \mid \mathcal{H}_i)}{p(D \mid \mathcal{H}_j)} \cdot \frac{p(\mathcal{H}_i)}{p(\mathcal{H}_j)} \tag{3.51}$$

If there is no prior reason to prefer one or the other hypothesis, we set $p(\mathcal{H}_i) = p(\mathcal{H}_j) = 1/2$ and the odds becomes just the ratio of evidences, also called the Bayes Factor $\mathcal{B}_{ij}$,

$$\mathcal{O}_{ij} = \mathcal{B}_{ij} = \frac{p(D \mid \mathcal{H}_i)}{p(D \mid \mathcal{H}_j)} \tag{3.52}$$

These can of course be inverted; if there are exactly two exhaustive and exclusive models, the following two equations hold

$$p(\mathcal{H}_i \mid D) = \frac{\mathcal{O}_{ij}}{1 + \mathcal{O}_{ij}} \tag{3.53}$$

$$p(\mathcal{H}_j \mid D) = \frac{1}{1 + \mathcal{O}_{ij}} \tag{3.54}$$

Odds $\mathcal{O}_{ij}$ larger than 1 favours model $\mathcal{H}_i$ and vice versa. If two models fit data equally well, then the odds will be close to 1. Some authors [1] state that in that case the simple models (with fewer parameters) should be preferred, but that remains contentious. Jeffreys [17] provides qualitive descriptions of its significance for the purposes of model comparison based on the $\log_{10}$ scale, then Kass and Raftery [18] modified Jeffreys' categories to natural logarithm scale, as the table shown below. Corresponding conclusions hold for negative log odds in favour of $\mathcal{H}_j$.

$2\log_e(\mathcal{B}_{ij})$	Results
>10	Decisively in favour of $\mathcal{H}_i$
6 to 10	Strongly favours $\mathcal{H}_i$
2 to 6	Positively favours $\mathcal{H}_i$
0 to 2	inconclusive
-2 to 0	inconclusive
-6 to -2	Positively against $\mathcal{H}_i$
-10 to -6	Strongly against $\mathcal{H}_i$
$<$ -10	Decisively against of $\mathcal{H}_i$

All these relations hold for the Bayes Factor $\mathcal{B}_{ij}$ when the hypotheses are given equal priors of $1/2$.

One big advantage of the odds is that, when the two models have the same number of parameters, it is possible to make their priors $p(\boldsymbol{\theta}_i \,|\, \mathcal{H}_i)$ and $p(\boldsymbol{\theta}_j \,|\, \mathcal{H}_j)$ quite similar and their effects can then be minimised because they cancel to some degree in the odds.

When there are more than two competing hypotheses, one can make a Bayes Factor matrix M, where $M_{ij} = \mathcal{H}_{ij}$ and $M_{ij} = 1/M_{ji}$, and designate one hypothesis as a "reference" and calculate all Bayes Factors with respect to this reference to find the relative probable hypotheses, then update the reference recursively until we find the most probable hypothesis.

3.2 Application to discrete data spectra

3.2.1 Overview

We now want to apply the generic theory of Chapter 2 and Section 3.1 to the specific problem of discrete data spectra. The general situation is that we have a raw data set of N counts as a function of some variable such as space, time, energy, angle etc. The data are pre-processed by grouping individual counts into measurement intervals, commonly called *bins*, with bin index $b = 1, \ldots, B$, bin midpoints x_b, and bin width ϵ. Usually the bin width is a constant,

$$\epsilon = \frac{x_{max} - x_{min}}{B} \tag{3.55}$$

but that is not a necessity. The binned data then has the form of bin counts n_b so that $D = \{n_b\}_{b=1}^{B}$ with $\sum_b n_b = N$. An example of binned data is shown in Figure 1.1. The height of each bar n_b is the sum of all raw counts n_i falling into bin b.

Our intent is to describe the data in terms of a model function (also called a parametrization or fit function) $f(x, \boldsymbol{\beta})$ of the coordinate x with K parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$. A

"good" description will be a case where a small number of parameters K describes a large number B of data points n_b well.

We set ourselves two tasks:

> Firstly, given a particular function $f(x, \boldsymbol{\beta})$, to find those values for $\boldsymbol{\beta}$ which "best" describe the data, where "best" will be defined in terms of posteriors for $\boldsymbol{\beta}$, i.e. parameter estimation,
>
> Secondly, to apply Bayesian evidence to the problem of deciding whether fit function f_1 or a different function f_2 is a better description of the data, i.e. model comparison between $\mathcal{H}_1$ and $\mathcal{H}_2$.

The first task will be covered in Sections 3.2.3 and 3.2.4 and the second in Section 3.3.

If, as we assume, the counts n_b in different bins b are independent, then the likelihood will factorise as usual,

$$p(D \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_B) \;=\; p(n_1, n_2, \ldots, n_B \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_B) = \prod_{b=1}^{B} p(n_b \,|\, \boldsymbol{\theta}_b), \tag{3.56}$$

where the symbols $\boldsymbol{\theta}_b$ represent the one or more parameters of the sort covered in Section 2.2. If for example we make the likelihood in each bin a Negative Binomial, then $\boldsymbol{\theta}_b$ represents the two NBD parameters (r_b, θ_b) in that bin.

In addition, there will be *hyperparameters*, i.e. parameters governing the distribution of the $\boldsymbol{\theta}_b$'s. There are two different kinds of hyperparameters:

1. the hyperparameters governing the priors for $\boldsymbol{\theta}_b$, one set for each bin b, and
2. the function parameters $\boldsymbol{\beta}$ which enter into the likelihoods of *all* bins because the same function f with the same parameters $\boldsymbol{\beta}$ is used in every bin in the form $f(x_b, \boldsymbol{\beta})$.

It is important to distinguish the function parameters $\boldsymbol{\beta}$ from the parameters entering the likelihood and the various priors.

In finding posteriors for $\boldsymbol{\beta}$, we are not interested in those prior hyperparameters on which $\boldsymbol{\beta}$ does not depend. These are therefore *nuisance parameters* which must either remain part of the answer or be integrated out. The process in Sections 3.2.3 and 3.2.4 will therefore be to first link the likelihood parameters $\boldsymbol{\theta}_b$ to the function parameters $\boldsymbol{\beta}$, and then to introduce the prior with its hyperparameters. Those hyperparameters which do not affect the $\boldsymbol{\beta}$ must then be integrated out.

3.2.2 Construction of Poisson and Negative Binomial likelihoods

Before we go to the full likelihood calculation, we must explain the connections between the Poisson, Gamma and Negative Binomial distributions in a single bin. In this section we therefore leave out the subscript b.

3.2.2.1 From Poisson to Gamma

As explained in Section 2.2.2, the Poisson distribution is the limit of a Binomial distribution and the result of a waiting time scenario. Due to its simplicity, it is commonly used to

model the likelihood. In each bin, the number of counts is described in terms of the Poisson parameter λ, as in Eq. (2.42),

$$p(n \,|\, \lambda) = \frac{e^{-\lambda}\lambda^n}{n!}. \tag{3.57}$$

The posterior for λ is found by applying Bayes' theorem, Eq. (3.10),

$$p(\lambda \,|\, n, \mathcal{H}) = \frac{p(n \,|\, \lambda, \mathcal{H}) p(\lambda \,|\, \mathcal{H})}{\int_0^\infty p(n \,|\, \lambda, \mathcal{H}) p(\lambda \,|\, \mathcal{H})\, \mathrm{d}\lambda} \tag{3.58}$$

For a uniform prior $p(\lambda) = \text{constant}$, we immediately obtain after cancellation a function which looks like Poisson

$$p(\lambda \,|\, n, \mathcal{H}) = \frac{e^{-\lambda}\lambda^n}{n!}, \tag{3.59}$$

but the variable is not n but λ, i.e. $p(\lambda \,|\, n, \mathcal{H})$ is the Gamma distribution of Eq. (2.73),

$$p(x \,|\, a, c) = \frac{c^a x^{a-1} e^{-cx}}{\Gamma(a)}$$

with $\lambda \equiv x$, $n \equiv a{-}1$ and $c \equiv 1$ and is normalised according to

$$\int_0^\infty d\lambda\, p(\lambda \,|\, n) = 1, \qquad \text{for all } n.$$

Here we use the notation

$$\lambda \sim \mathrm{Ga}(\lambda \,|\, n{+}1, 1) \tag{3.60}$$

to indicate that λ is gamma-distributed with the parameters as shown. Although it has the same funtion form of the likelihood function $\mathrm{p}(n \,|\, \lambda, \mathcal{H})$, unlike n, which can only be *positive integers*, λ has *positive continuous values*, so $\mathrm{p}(\lambda|n, \mathcal{H})$ is a probability density function.

For a Gamma prior

$$p(\lambda \,|\, a, c) = \mathrm{Ga}(\lambda \,|\, a, c) = \frac{c^a \lambda^{a-1} e^{-c\lambda}}{\Gamma(a)} \tag{3.61}$$

the posterior for λ is

$$p(\lambda \,|\, n, a, c) = \frac{p(n \,|\, \lambda)\, p(\lambda \,|\, a, c)}{\int p(n \,|\, \lambda)\, p(\lambda \,|\, a, c)\, \mathrm{d}\lambda} \tag{3.62}$$

given n is again a Gamma distribution but with shifted parameters,

$$p(\lambda \,|\, n, a, c) = \mathrm{Ga}(\lambda \,|\, n{+}a, c{+}1). \tag{3.63}$$

where setting $a = 1, c = 0$ corresponds to the uniform prior result Eq. (3.60).

In Figure 3.2, we plot $\mathrm{Ga}(\lambda \,|\, n{+}1, 1)$ for various n. Note that a measurement of $n = 0$ does not imply $\lambda = 0$ but rather that λ follows an exponential distribution.

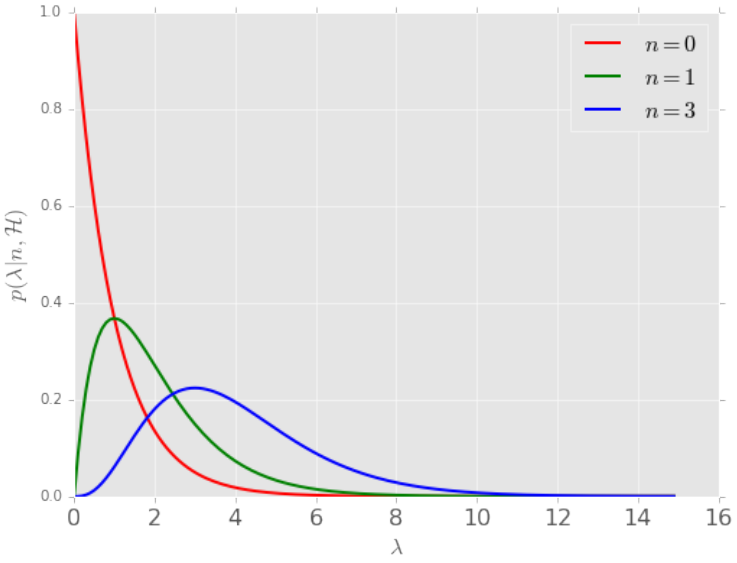


Figure 3.2: Posterior distribution of λ for different observations(flat prior). Noting that in case of null observation, the posterior distribution for λ is not necessarily zero.

The cumulant generating function for the posterior is $K(t|P(\lambda\,|\,n)) = -(n+1)\ln(1-t)$, so that the mean, variance, skewness and kurtosis are given as

$$\begin{aligned} E[\lambda\,|\,n,\mathcal{H}] &= n+1, \\ \mathrm{var}[\lambda\,|\,n,\mathcal{H}] &= n+1, \\ \gamma_1[\lambda\,|\,n,\mathcal{H}] &= 2/\sqrt{n+1}, \\ \gamma_2[\lambda\,|\,n,\mathcal{H}] &= 6/(n+1). \end{aligned}$$

If $n \gg 1$, then

$$\begin{aligned} E[\lambda\,|\,n,\mathcal{H}] = \mathrm{var}[\lambda\,|\,n,\mathcal{H}] &= n \\ \gamma_1 = \gamma_2 &= 0, \end{aligned}$$

in other words for large n, the posterior distribution of λ approaches a Gaussian (n, n).

3.2.2.2 From Gamma to Negative Binomial

In the case of a Poisson likelihood $p(n\,|\,\lambda)$, a fixed λ is assumed. There is however no reason why it should be fixed to a single value; there often are processes which make λ fluctuate. The likelihood for n for all values of the fluctuating λ is then an integral over $p(n\,|\,\lambda)$ weighted by whatever distribution λ itself follows. This situation is called a *mixture* or *compound process* where the parameter itself varies.

From the previous section, it is clear that it is a natural assumption that the Poisson parameter λ can be Gamma-distributed. It is therefore reasonable to use a Gamma prior for λ. In anticipation of later notation, we rewrite the hyperparameters for this Gamma prior as $a \equiv r$ and c. The likelihood for n is then the integral over all λ of the Poisson $p(n\,|\,\lambda)$, weighted by the prior $p(\lambda\,|\,r, c)$,

For this choice, the compound-process likelihood for n is

$$p(n \,|\, r, c, \mathcal{H}) = \int_0^\infty p(n \,|\, \lambda)\, p(\lambda \,|\, r, c)\, \mathrm{d}\lambda \tag{3.64}$$

which results in a likelihood

$$\begin{aligned} p(n \,|\, r, c, \mathcal{H}) &= \int \frac{e^{-\lambda}\lambda^n}{n!} \cdot \frac{c^r \lambda^{r-1} e^{-c\lambda}}{\Gamma(r)}\, \mathrm{d}\lambda \\ &= \frac{c^r}{n!\Gamma(r)} \int e^{-(c+1)\lambda} \lambda^{n+r-1}\, \mathrm{d}\lambda \\ &= \frac{\Gamma(n+r)}{n!\Gamma(r)} \left(\frac{1}{c+1}\right)^n \left(\frac{c}{c+1}\right)^r \\ &= \frac{\Gamma(n+r)}{n!\Gamma(r)} \left(\frac{1}{c+1}\right)^n \left(1 - \frac{1}{c+1}\right)^r \end{aligned} \tag{3.65}$$

Defining [19]

$$\theta = \frac{1}{c+1} \tag{3.66}$$

we recognise it as a Negative Binomial distribution as in Eq. (2.61),

$$p(n \,|\, r, \theta) = \frac{\Gamma(n+r)}{n!\Gamma(r)} \theta^n (1-\theta)^r \qquad \text{or} \qquad n \sim \mathrm{NBD}(n \,|\, r, (1{+}c)^{-1}).$$

From the above derivation, we see that $p(n \,|\, r, c)$, which is a likelihood, is at the same time also the denominator which entered into the posterior for λ in Eq. (3.62) and played the role of "evidence".

3.2.3 Fit function parameter estimation: Poisson case

In this section, we consider the data $D = \{n_b\}_{b=1}^B$ in all B bins simultaneously with a view to finding a posterior for the function parameters, $p(\boldsymbol{\beta} \,|\, D)$, omitting $\mathcal{H}$ in the arguments. As usual we assume that bin contents are independent and that the likelihood factorises,

$$p(D \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B) = \prod_{b=1}^B p(n_b \,|\, \boldsymbol{\theta}_b). \tag{3.67}$$

For the Poisson likelihood case, this becomes, with $\lambda_b \equiv \boldsymbol{\theta}_b$,

$$p(D \,|\, \boldsymbol{\lambda}) = p(D \,|\, \lambda_1, \lambda_2, \ldots, \lambda_B) = \prod_{b=1}^B p(n_b \,|\, \lambda_b) = \prod_{b=1}^B \frac{e^{-\lambda_b} \lambda_b^{n_b}}{n_b!}. \tag{3.68}$$

In this our first strategy, we fix each Poisson parameter to the value of the model function in that bin,

$$\lambda_b = f(x_b, \boldsymbol{\beta}) \qquad b = 1, \ldots, B. \tag{3.69}$$

This equation represents a mapping from the K function parameters $\boldsymbol{\beta}$ to the B Poisson parameters,

$$\lambda_b = \lambda_b(\boldsymbol{\beta}) \qquad b = 1, \ldots, B. \tag{3.70}$$

and we can rewrite the likelihood as

$$p(D \,|\, \lambda_1, \lambda_2, \ldots, \lambda_B) = p(D \,|\, \lambda_1(\boldsymbol{\beta}), \ldots, \lambda_B(\boldsymbol{\beta})) = p(D \,|\, \boldsymbol{\beta}) \tag{3.71}$$

since the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\lambda}$ does not entail a Jacobian. This change allows us to apply Bayes' theorem directly to $\boldsymbol{\beta}$ rather than the intermediary $\boldsymbol{\lambda}$ parameters: the posterior for $\boldsymbol{\beta}$ is

$$\begin{aligned} p(\boldsymbol{\beta} \,|\, D, \mathcal{H}) = \frac{p(D \,|\, \boldsymbol{\beta}, \mathcal{H})\, p(\boldsymbol{\beta} \,|\, \mathcal{H})}{p(D \,|\, \mathcal{H})} &= \frac{p(\boldsymbol{\beta} \,|\, \mathcal{H})}{p(D \,|\, \mathcal{H})} \prod_b \frac{e^{-\lambda_b(\boldsymbol{\beta})} \lambda_b(\boldsymbol{\beta})^{n_b}}{n_b!} \\ &= \frac{p(\boldsymbol{\beta} \,|\, \mathcal{H})}{p(D \,|\, \mathcal{H})} \prod_b \frac{e^{-f_b(\boldsymbol{\beta})} f_b(\boldsymbol{\beta})^{n_b}}{n_b!} \end{aligned} \tag{3.72}$$

with

$$p(D \,|\, \mathcal{H}) = \int p(D \,|\, \boldsymbol{\beta}, \mathcal{H})\, p(\boldsymbol{\beta} \,|\, \mathcal{H})\, \mathrm{d}\boldsymbol{\beta} \tag{3.73}$$

which in most cases cannot be solved analytically. Since we are only interested in the functional dependence of the posterior on $\boldsymbol{\beta}$, we can take the logarithm and neglect all terms which are $\boldsymbol{\beta}$-independent,

$$\begin{aligned} L &\equiv \log p(\boldsymbol{\beta} \,|\, D, \mathcal{H}) && (3.74) \\ &= \log p(\boldsymbol{\beta} \,|\, \mathcal{H}) + \log p(D \,|\, \boldsymbol{\beta}, \mathcal{H}) + \text{constants } C \\ &= \log p(\boldsymbol{\beta} \,|\, \mathcal{H}) + \sum_b \big[- f_b(\boldsymbol{\beta}) + n_b \log f_b(\boldsymbol{\beta}) \big] && (3.75) \end{aligned}$$

We shall need expressions for the mode later. Since the logarithm is a monotonic function, the mode $\boldsymbol{\beta}^*$ of L and $p(\boldsymbol{\beta} \,|\, D, \mathcal{H})$ is the same, $\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} L(\boldsymbol{\beta})$, in other words, solving the system of equations

$$\left.\frac{\partial L}{\partial \beta_k}\right|_{\boldsymbol{\beta}^*} = \frac{1}{p(\boldsymbol{\beta} \,|\, \mathcal{H})} \frac{\partial p(\boldsymbol{\beta} \,|\, \mathcal{H})}{\partial \beta_k} + \sum_b \left(\frac{n_b}{f_b(\boldsymbol{\beta})} - 1 \right) \frac{\partial f_b(\boldsymbol{\beta})}{\partial \beta_k} = 0, \qquad k = 1, 2, \ldots, K. \tag{3.76}$$

3.2.4 Fit function parameter estimation: Negative Binomial case

In the second strategy, we model the case where λ_b is not fixed but can fluctuate according to some prior $p(\lambda_b \,|\, \text{hyperparameters})$ as already discussed. Of course the question is then how to connect λ_b with $f(x_b \,|\, \boldsymbol{\beta})$. We think it is reasonable to require that the *expectation value of* λ_b should be fixed to the function,

$$E(\lambda_b) = f(x_b, \boldsymbol{\beta}) \;=\; f_b(\boldsymbol{\beta}) \tag{3.77}$$

where we introduce shortened notation $f_b = f_b(\boldsymbol{\beta})$ for use below. We also choose a gamma prior with hyperparameters (r_b, c_b)

$$p(\lambda_b \,|\, c_b, r_b) = c_b^{r_b} \lambda_b^{r_b - 1} e^{-c_b \lambda_b} / \Gamma(r_b) \tag{3.78}$$

which is exactly the choice taken in Eq. (3.65). From this follows immediately that the joint likelihood becomes a product of Negative Binomials. Including explicitly the hyperparameters $\boldsymbol{c} = (c_1, \ldots, c_B)$ and $\boldsymbol{r} = (r_1, \ldots, r_B)$,

$$p(D \,|\, \boldsymbol{c}, \boldsymbol{r}) = \prod_{b=1}^{B} p(n_b \,|\, c_b, r_b) \;=\; \prod_b \frac{\Gamma(r_b + n_b)}{n_b!\,\Gamma(r_b)} \left(\frac{1}{c_b + 1}\right)^{n_b} \left(\frac{c_b}{c_b + 1}\right)^{r_b}. \tag{3.79}$$

Now the expectation value for λ_b within the Gamma prior is

$$f_b(\boldsymbol{\beta}) = E(\lambda_b) = \frac{r_b}{c_b} \tag{3.80}$$

Surprisingly, the same expectation value of λ_b in the Gamma distribution is also the expectation value of n_b in the Negative Binomial.

Furthermore, we know from Eq. (3.66) that the Negative Binomial parameter θ_b is related to the Gamma parameter c_b by $\theta_b = 1/(c_b + 1)$ and

$$c_b = \frac{1 - \theta_b}{\theta_b} = \frac{r_b}{f_b} \tag{3.81}$$

$$1 + c_b = \frac{1}{\theta_b} = \frac{r_b + f_b}{f_b} \tag{3.82}$$

so that we can eliminate c_b in favour of the expectation value f_b of Eq. (3.77), and so, with $\boldsymbol{f}(\boldsymbol{\beta}) = (f(x_1, \boldsymbol{\beta}), \ldots, f(x_B, \boldsymbol{\beta})) = (f_1, f_2, \ldots, f_B)$,

$$p(D \,|\, \boldsymbol{f}(\boldsymbol{\beta}), \boldsymbol{r}) = \prod_b \frac{\Gamma(r_b + n_b)}{n_b!\,\Gamma(r_b)} \left(\frac{f_b}{f_b + r_b}\right)^{n_b} \left(\frac{r_b}{f_b + r_b}\right)^{r_b}. \tag{3.83}$$

As discussed in Section 2.2.4, the parameters r_b govern the shape of the Gamma prior, while the c_b govern the inverse scale. We have already set the inverse scale parameters c_b to be determined by the model function f and r_b. There is no need to have B distinct shape hyperparameters $\boldsymbol{r}$, and so we set them all equal to a single shape parameter, $r_b = r$ for all b. The final form of the Negative Binomial likelihood is therefore

$$p(D \,|\, \boldsymbol{\beta}, r) = \prod_b \frac{\Gamma(r + n_b)}{n_b!\,\Gamma(r)} \left(\frac{f_b}{f_b + r}\right)^{n_b} \left(\frac{r}{f_b + r}\right)^{r}. \tag{3.84}$$

Bayes' theorem (3.10) yields

$$p(\boldsymbol{\beta}, r \,|\, D, \mathcal{H}) = \frac{p(D \,|\, \boldsymbol{\beta}, r, \mathcal{H}) p(\boldsymbol{\beta}, r \,|\, \mathcal{H})}{p(D \,|\, \mathcal{H})} \;\propto\; p(D \,|\, \boldsymbol{\beta}, r, \mathcal{H}) p(\boldsymbol{\beta}, r \,|\, \mathcal{H}). \tag{3.85}$$

As our primary interest is to estimate $\boldsymbol{\beta}$, we marginalise over nuisance parameter r to obtain the desired function parameter posterior,

$$p(\boldsymbol{\beta} \,|\, D, \mathcal{H}) = \int p(\boldsymbol{\beta}, r \,|\, D, \mathcal{H})\, \mathrm{d}r$$

Since the Negative Binomial likelihood has no conjugate priors, we assume $\boldsymbol{\beta}$ and r have a uniform prior; the only information is that both must be positive, $r > 0$ and $\boldsymbol{\beta} \geq 0$. Then the MAP is identical to ML, that is

$$p(\boldsymbol{\beta}, r \,|\, D, \mathcal{H}) \propto p(D \,|\, \boldsymbol{\beta}, r, \mathcal{H}).$$

Noting that for all $n_b \geq 1$ and $r > 0$ [19],

$$\begin{aligned} \log\left(\frac{\Gamma(r+n_b)}{\Gamma(r)}\right) &= \log\Big[r(r+1)\ldots(r+n_b-1)\Big] \\ &= \log\left[\prod_{j=0}^{n_b-1}(r+j)\right] \\ &= \sum_{j=0}^{n_b-1}\log(r+j), \end{aligned} \tag{3.86}$$

the log-posterior for the NB model is computed as

$$\begin{aligned} L &\equiv \log p(\boldsymbol{\beta}, r \,|\, D, \mathcal{H}) \equiv \log p(\boldsymbol{\beta}, r \,|\, \mathcal{H}) + \log p(D \,|\, \boldsymbol{\beta}, r, \mathcal{H}) + C \\ &= \log p(D \,|\, \boldsymbol{\beta}, r, \mathcal{H}) \\ &= \sum_{b=1}^{B}\left\{\sum_{j=0}^{n_b-1}\log(r+j) - \log(n_b!) + r\log r + n_b\log(f_b) - (n_b+r)\log(r+f_b)\right\}. \end{aligned}$$

The modes are obtained by the usual first derivatives,

$$\left.\frac{\partial L}{\partial \beta_k}\right|_{\boldsymbol{\beta}^*} = \sum_{b=1}^{B}\left\{\frac{n_b - f_b}{f_b(1+f_b/r)}\right\}\frac{\partial f_b}{\partial \beta_k} = 0, \qquad k = 1, 2, \ldots, K. \tag{3.87}$$

$$\left.\frac{\partial L}{\partial r}\right|_{r*} = \sum_{b=1}^{B}\left\{\sum_{j=0}^{n_b-1}\left(\frac{1}{r+j}\right) - \log(r+f_b) - \frac{r+n_b}{r+f_b} + \log r + 1\right\} = 0 \tag{3.88}$$

3.3 Model Comparison for discrete data spectra

3.3.1 Poisson versus Negative Binomial

Given data $D = \{n_b\}_{b=1}^{B}$, it is desired to test two hypotheses:

- $\mathcal{H}_0$ is a "point" or "null" hypothesis, where $r = \infty$, saying that in each bin the data was generated by a Poisson likelihood with parameter $\lambda_b(\boldsymbol{\beta})$.
- $\mathcal{H}_1$ claims that in each bin the data was generated by a Negative Binomial likelihood with unknown parameter r.

Therefore, we can say that model $\mathcal{H}_0$ is nested within model $\mathcal{H}_1$, and we are deciding how much complexity is demanded to interpret the variance of data D. Applying Eq. (3.52), we

obtain the Bayes Factor, with $\boldsymbol{\lambda} = \{\lambda_b\}_{b=1}^{B}$,

$$\begin{aligned}\mathcal{B}_{01} &= \frac{p(D\,|\,\mathcal{H}_0)}{p(D\,|\,\mathcal{H}_1)} \\ &= \frac{\displaystyle\int p(D\,|\,\boldsymbol{\lambda},\mathcal{H}_0)\,p(\boldsymbol{\lambda}\,|\,\mathcal{H}_0)\,\mathrm{d}\boldsymbol{\lambda}}{\displaystyle\int p(D\,|\,r,\boldsymbol{\beta},\mathcal{H}_1)\,p(\boldsymbol{\beta}\,|\,\mathcal{H}_1)p(r\,|\,\mathcal{H}_1)\,\mathrm{d}\boldsymbol{\beta}\mathrm{d}r} \\ &= \frac{\displaystyle\int\prod_{b=1}^{B}\frac{e^{-f_b}f_b^{n_b}}{n_b!}p(\boldsymbol{\beta}\,|\,\mathcal{H}_0)\,\mathrm{d}\boldsymbol{\beta}}{\displaystyle\int\prod_{b=1}^{B}\frac{\Gamma(r+n_b)}{n_b!\Gamma(r)}\left(\frac{f_b}{f_b+r}\right)^{n_b}\left(\frac{r}{f_b+r}\right)^{r}p(\boldsymbol{\beta},r\,|\,\mathcal{H}_1)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}r}\end{aligned} \tag{3.89}$$

or equivalently the logarithm

$$\log[\mathcal{B}_{01}] = \log(p(D\,|\,\mathcal{H}_0)) - \log(p(D\,|\,\mathcal{H}_1)) \tag{3.90}$$

We will make use of the guidelines of Jeffreys set out in Section 3.1.6 above.

3.3.2 Laplace Approximation of Bayes Factor for Poisson and NB likelihoods

Laplace's method can be applied to both numerator and denominator of the Bayes Factor. In general, the approximation works well for those likelihood functions which are not grossly skewed, with modest dimensionality K and sample size N. As usual, analytical results are preferred because of convergence and stability; but they fail when there is more than one mode. Also the form near the maxima may not be well-represented by a multidimensional Gaussian distribution in practice.

Under those caveats, let us consider the Laplace approximation for the evidence, Eq. (3.49), first for hypothesis $\mathcal{H}_0$ of a Poisson likelihood

$$p(D\,|\,\mathcal{H}_0) \approx p(D\,|\,\boldsymbol{\beta}^*,\mathcal{H}_0)p(\boldsymbol{\beta}^*\,|\,\mathcal{H}_0)\sqrt{\frac{(2\pi)^K}{\det H_0(\boldsymbol{\beta}^*)}}$$

with a Hessian H_0 with components

$$\begin{aligned}(H_0)_{k,l} &= -\frac{\partial^2\log p(D\,|\,\mathcal{H}_0)}{\partial\beta_k\partial\beta_l} \\ &= -\left\{\frac{1}{p(\boldsymbol{\beta}\,|\,\mathcal{H}_0)^2}\frac{\partial p(\boldsymbol{\beta}\,|\,\mathcal{H}_0)}{\partial\beta_k}\frac{\partial p(\boldsymbol{\beta}\,|\,\mathcal{H}_0)}{\partial\beta_l} + \frac{1}{p(\boldsymbol{\beta}\,|\,\mathcal{H}_0)}\frac{\partial^2 p(\boldsymbol{\beta}\,|\,\mathcal{H}_0)}{\partial\beta_k\partial\beta_l}\right. \\ &\quad \left. + \sum_b\left[-\frac{n_b}{f_b^2}\frac{\partial f_b}{\partial\beta_k}\frac{\partial f_b}{\partial\beta_l} + \left(\frac{n_b}{f_b}-1\right)\frac{\partial^2 f_b}{\partial\beta_k\partial\beta_l}\right]\right\}\end{aligned} \tag{3.91}$$

When a uniform prior is used for $\boldsymbol{\beta}$, the Hessian reduces to

$$(H_0)_{k,l} = -\sum_b\left[-\frac{n_b}{f_b^2}\frac{\partial f_b}{\partial\beta_k}\frac{\partial f_b}{\partial\beta_l} + \left(\frac{n_b}{f_b}-1\right)\frac{\partial^2 f_b}{\partial\beta_k\partial\beta_l}\right] \tag{3.92}$$

Under hypothesis $\mathcal{H}_1$ of Negative Binomial likelihoods, the approximate evidence is

$$p(D\,|\,\mathcal{H}_1) \approx p(\boldsymbol{\beta}^*, r^*\,|\,\mathcal{H}_1)\, p(D\,|\,\boldsymbol{\beta}^*, r^*, \mathcal{H}_1) \sqrt{\frac{(2\pi)^K}{\det H_1(\boldsymbol{\beta}^*, r^*)}}$$

If uniform priors are used, $\log p(\boldsymbol{\beta}, r\,|\,\mathcal{H}_1) = 0$ and the log-posterior of the NB model is $L_1 = \log p(D\,|\,\boldsymbol{\beta}, r, \mathcal{H}_1)$, the matrix H_1 for r would consist of matrix elements

$$-\frac{\partial^2 L_1}{\partial r^2} = \sum_{b=1}^{B} \left\{ \sum_{j=0}^{n_b-1} \left(\frac{1}{r+j} \right)^2 - \frac{n_b r + f_b^2}{r(r+f_b)^2} \right\}; \tag{3.93}$$

the mixed terms are

$$-\frac{\partial^2 L_1}{\partial r \partial \beta_k} = -\sum_{b=1}^{B} \left\{ \frac{n_b - f_b}{(r+f_b)^2} \right\} \frac{\partial f_b}{\partial \beta_k}, \tag{3.94}$$

and matrix elements in terms of $\boldsymbol{\beta}$,

$$-\frac{\partial^2 L_1}{\partial \beta_k \partial \beta_l} = \sum_{b=1}^{B} \left\{ \left(\frac{n_b}{f_b^2} - \frac{r+n_b}{(r+f_b)^2} \right) \frac{\partial f_b}{\partial \beta_k} \frac{\partial f_b}{\partial \beta_l} - \frac{n_b - f_b}{f_b(1+f_b/r)} \frac{\partial^2 f_b}{\partial \beta_k \partial \beta_l} \right\}. \tag{3.95}$$

3.4 Solving the simplest case $f_b = \beta$

We start with the simplest possible model $f_b = \beta$ for all b, which is just fitting a fixed background signal.

Poisson likelihood: Assuming a uniform prior for β, the posterior distribution (3.72) is

$$p(\beta\,|\,D, \mathcal{H}_0) \propto p(D\,|\,\boldsymbol{\beta}, \mathcal{H}_0) = \prod_b \frac{e^{-\beta} \beta^{n_b}}{n_b!} \tag{3.96}$$

and from Eq. (3.76), the best parameter value is

$$\beta^* = \frac{1}{B} \sum_{b=1}^{B} n_b = \frac{N}{B} = \bar{n} \tag{3.97}$$

and since the second derivative is negative,

$$\frac{\partial^2 L}{\partial \beta^2} = -\sum_b \frac{n_b}{\beta^2} = -\frac{B}{\bar{n}} < 0 \tag{3.98}$$

the $\beta^* = \bar{n}$ is the desired MAP point. The asymptotic posterior distribution for β is the Gaussian,

$$p(\beta\,|\,D, \mathcal{H}_0) \approx \mathcal{N}\left(\bar{n}, \frac{\bar{n}}{B} \right). \tag{3.99}$$

Negative Binomial likelihood: First, we calculate the mode of parameters β^*, then obtain the Hessian matrix to evaluate the posterior distribution. From Eq. (3.87),

$$\beta^* = \frac{1}{B}\sum_{b=1}^{B} n_b = \bar{n}, \tag{3.100}$$

and from Eq. (3.95), the second derivative is again negative,

$$\frac{\partial^2 L}{\partial \beta^2} = -\sum_{b=1}^{B}\left\{\frac{n_b}{f_b^2} - \frac{r+n_b}{(r+f_b)^2}\right\} = -\frac{B}{\bar{n}(1+\bar{n}/r)} < 0, \tag{3.101}$$

while the off-diagonal term in the Hessian is

$$-\frac{\partial^2 L}{\partial r \partial \beta} = -\sum_{b=1}^{B}\left\{\frac{n_b - f_b}{(r+f_b)^2}\right\} = -\sum_{b=1}^{B}\left\{\frac{n_b - \bar{n}}{(r+\bar{n})^2}\right\} = 0. \tag{3.102}$$

We conclude that r and β are independent in this example [20].

Substituting β^* into Eq.(3.88) and calculating r^*,

$$\frac{\partial L}{\partial r} = \sum_{b=1}^{B}\left\{\sum_{j=0}^{n_b-1}\left(\frac{1}{r+j}\right) - \log(r+\bar{n}) + \log r\right\} = 0 \tag{3.103}$$

the equation cannot be solved analytically, so that a Newton-Raphson algorithm is applied to find the solution. The value of r is updated until $\frac{\partial L}{\partial r} = 0$.

$$r^{(t+1)} := r^{(t)} - \left(\frac{\partial L}{\partial r} \Big/ \frac{\partial^2 L}{\partial r^2}\right)\Big|_{r^{(t)}} \tag{3.104}$$

Alternatively, the value of $r^{(t+1)}$ is updated through

$$r^{(t+1)} := r^{(t)} + \alpha \nabla L(r)\big|_{r^{(t)}} \tag{3.105}$$

This technique is termed as steepest ascent, where the learning rate $\alpha > 0$, and the gradient $\nabla L(r)$ determines the direction of steepest ascent; r converges when L is maximized.

The posteriors for β and r in the NB model are hence

$$p(\beta \mid D, r, \mathcal{H}_1) \approx \mathcal{N}\left(\bar{n}, \frac{\bar{n}(1+\bar{n}/r^*)}{B}\right) \tag{3.106}$$

$$p(r \mid D, \beta, \mathcal{H}_1) \approx \mathcal{N}\Big(r^*, \mathrm{var}(r)\Big) \tag{3.107}$$

where $\mathrm{var}(r) = \left\{-\frac{\partial^2 L}{\partial r^2}\right\}^{-1}\Big|_{r^*}$ while the second derivative is

$$-\frac{\partial^2 L}{\partial r^2} = \sum_{b=1}^{B}\left\{\sum_{j=0}^{n_b-1}\left(\frac{1}{r+j}\right)^2\right\} - \frac{B\bar{n}}{r(r+\bar{n})}. \tag{3.108}$$

We see that even in the easiest paradigm, direct computation of parameters from the NB model impedes us from acquiring modes.

Chapter 4

Markov Chain Monte Carlo methods

While Section 3.2 has highlighted important analytical results for the construction of likelihoods for discrete data, it was also clear that numerical evaluation is necessary. Log posteriors and their Negative Binomial equivalents can be written down formally but the equations for the modes (maxima) cannot be solved analytically. Secondly, the number of parameters K may become large, and it is usually hard to visualise the form of the posteriors. Often it is also necessary to integrate out nuisance parameters which are not of interest. It is also clear from Section 3.3 that computation of the evidence involves integrals in parameter space which can rarely be handled analytically and that numeric methods are essential.

In this chapter, we introduce Monte Carlo (MC) methods and specifically the Markov Chain Monte Carlo (MCMC) family. We will use the parameters $\boldsymbol{\theta}$ as set out in Chapter 3 as our variables, but the MCMC methods are very general and not limited to the purpose of parameter space integration.

4.1 Theoretical basis

4.1.1 Underlying assumptions

MCMC evolved independently from Bayes and the logic used in MCMC literature is not Bayesian but frequentist. It relies on taking the number of elements in a sample to infinity. We will follow the traditional frequentist approach in this chapter.

The generic task solved by MCMC is as follows: Given a K-dimensional variable $\boldsymbol{\theta}$ and any probability of these variables $\pi(\boldsymbol{\theta})$, it is often easy to write down the unnormalised functional dependence of $\boldsymbol{\theta}$ but hard to explicitly integrate this in order to normalise $\pi(\boldsymbol{\theta})$. One then speaks of an unnormalised probability $\pi^*(\boldsymbol{\theta})$ which is related to the properly normalised one by[1]

$$\pi(\boldsymbol{\theta}) = \frac{\pi^*(\boldsymbol{\theta})}{Z}, \tag{4.1}$$

where $Z = \int \pi^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the unknown integral. A good example is the posterior distribution,

$$p(\boldsymbol{\theta} \mid D, \mathcal{H}) = \frac{p(D \mid \boldsymbol{\theta}, \mathcal{H})\, p(\boldsymbol{\theta} \mid \mathcal{H})}{p(D \mid \mathcal{H})} \tag{4.2}$$

[1] It should be clear from the context when π refers to the mathematical constant.

in which $\pi^*(\boldsymbol{\theta})$ refers to the product of likelihood and prior, $p(D\,|\,\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}\,|\,\mathcal{H})$, while Z is the evidence. The difference between Bayesian and frequentist solutions often lies in the fact that $\pi^*(\boldsymbol{\theta})$ is identified only with the likelihood in the frequentist case, while Bayesians include the prior and the likelihood. For the sake of developing the theory of MCMC, it does not matter which of these two definitions is used.

In terms of the generic notation, the tasks of MCMC would include finding marginal distributions such as

$$\pi(\theta_1) = \iint \pi(\theta_1,\theta_2,\theta_3)\,\mathrm{d}\theta_2\,\mathrm{d}\theta_3 \tag{4.3}$$

or integrating over all parameters

$$Z = \int \pi^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{4.4}$$

or calculating the expectation value of some function of the parameters $f(\boldsymbol{\theta})$,

$$E(f(\boldsymbol{\theta})) = \int f(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \;=\; \frac{\int f(\boldsymbol{\theta})\,\pi^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\int \pi^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \tag{4.5}$$

Monte Carlo always creates a *sample of numbers* distributed according to some $\pi^*(\boldsymbol{\theta})$. It must be emphasised that the sample referred to is not a data sample but a sample of parameter values or of functions of those values. The basic sampling procedure may be characterised as follows:

1. Start the simulation process by generating the first sample vector $\boldsymbol{\theta}^{(1)}$ according to the joint distribution $p(\theta_1,\theta_2,\cdots\theta_K)$, and compute $f^{(1)} = f(\boldsymbol{\theta}^{(1)})$.

2. Generate the second independent sample vector $\boldsymbol{\theta}^{(2)}$ to obtain $f^{(2)} = f(\boldsymbol{\theta}^{(2)})$. (Regarding independence, see Section 4.2.1.) This procedure is repeated multiple times. After constructing S number of i.i.d. samples, $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{S}$, the distribution of random variable $f^{(j)} = f(\boldsymbol{\theta}^{(j)})$ is i.i.d. as well. According to the Law of Large Numbers, the sample average $\langle f\rangle$ converges to its population mean as in Section 2.3.1.

$$\lim_{S\to\infty}\frac{1}{S}\sum_{j=1}^{S} f^{(j)} = \langle f\rangle \approx E(f). \tag{4.6}$$

3. Calculate the mean and variance of estimator $\langle f\rangle$. The Law of Large Numbers and the Central Limit Theorem (Section 2.3.2) guarantee that, when the variables are i.i.d., the distribution of the standardised variable

$$f^* = \frac{\langle f\rangle - E(\langle f\rangle)}{\sqrt{\mathrm{var}(\langle f\rangle)}} \tag{4.7}$$

with

$$\mathrm{var}(\langle f\rangle) = \mathrm{var}\left[\frac{1}{S}\sum_{j=1}^{S} f^{(j)}\right] \;=\; \frac{\mathrm{var}(f)}{S} \tag{4.8}$$

converges to a standardised Gaussian distribution,

$$p(f^*) \to \text{Gaussian}(0,1) \tag{4.9}$$

Note that the variance of $\langle f \rangle$ shrinks on a scale of $1/S$ proportional to its target variance $\text{var}(f)$. In practice, the "unbiased estimator" of the population variance

$$\sigma^2_{\text{est}} = \frac{1}{S-1}\sum_{j=1}^{S}(f^{(j)} - \langle f \rangle)^2 \tag{4.10}$$

is used for uncertainty estimation, but the difference is not important.

4.1.2 Stochastic processes

Before concentrating on the Metropolis-Hastings case, we first sketch the basic concepts and issues of stochastic processes in general, roughly following the text of [21].

Markov property: The state space $\mathcal{A} = \mathcal{A}(\boldsymbol{\theta})$ can be defined as either a set of discrete values which for the sake of simplicity we just number by an index i, or by real values whose numbers can be categorised into discrete bins which we also call i for the moment [1]. When the random variable $\boldsymbol{\theta}_t$ takes discrete values, $\boldsymbol{\theta}_t = i$, for $i \in \mathcal{A}$, we say that the process is in a state i at time t. The next variable $\boldsymbol{\theta}_{t+1}$ is determined by a conditional probability which in principle depends on the entire history of states,

$$p(\boldsymbol{\theta}_{t+1}{=}j \,|\, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \ldots, \boldsymbol{\theta}_0). \tag{4.11}$$

The index t can be interpreted as sequence index of data; it is not physical time. Such a stochastic process is called a Markov Chain if, for all states $\boldsymbol{\theta}_t, t \geq 0$, the probability of reaching state $\boldsymbol{\theta}_{t+1} = j$ depends only on the previous state $\boldsymbol{\theta}_t = i$,

$$p(\boldsymbol{\theta}_{t+1}{=}j \,|\, \boldsymbol{\theta}_t{=}i, \boldsymbol{\theta}_{t-1}, \ldots, \boldsymbol{\theta}_0) = p(\boldsymbol{\theta}_{t+1}{=}j \,|\, \boldsymbol{\theta}_t{=}i) = A_{ij} \tag{4.12}$$

using A_{ij} as brief notation for the conditional probability, which is also called the transition probability. Because the same function is used for the conditional probability $p(\boldsymbol{\theta}_{t+1} \,|\, \boldsymbol{\theta}_t)$ for any t, it is *time stationary* or *time homogeneous*. The set of transitions $\{A_{ij} \,|\, i, j \in \mathcal{A}\}$ can be represented as a square matrix, where each row sums to one due to the normalisation of the conditional probability, $\sum_{j\in\mathcal{A}} A_{ij} = 1$.

Vector-matrix formulation: After initialising the state's distribution with probability $\pi_0(\boldsymbol{\theta})$, the processes's distribution changes on the next step to

$$\pi_1(j) = \sum_{i\in\mathcal{A}} \pi_0(i)\, A_{ij} \tag{4.13}$$

This can be written in a neat vector-matrix form with $\boldsymbol{\pi}(\boldsymbol{\theta})$ representing row vectors of outcomes $\pi(\boldsymbol{\theta}{=}i)$

$$\boldsymbol{\pi}_1(\boldsymbol{\theta}_1) = \boldsymbol{\pi}_0(\boldsymbol{\theta}_0)A. \tag{4.14}$$

Furthermore, we would like to define a n-step transition, from state i to destination state e after n transitions.

$$A_{ie}^{(n)} = p(\boldsymbol{\theta}_{t+n}{=}e \,|\, \boldsymbol{\theta}_t{=}i). \tag{4.15}$$

The Chapman-Kolmogorov equations provide a way to compute $(n+m)$-step transitions as follows,

$$\begin{aligned} A_{ie}^{(n+m)} = p(\boldsymbol{\theta}_{n+m}{=}e \,|\, \boldsymbol{\theta}_0) \;&=\; \sum_{k\in\mathcal{A}} p(\boldsymbol{\theta}_{n+m}{=}e, \boldsymbol{\theta}_n{=}k \,|\, \boldsymbol{\theta}_0{=}i) \\ &= \sum_{k\in\mathcal{A}} p(\boldsymbol{\theta}_n{=}k \,|\, \boldsymbol{\theta}_0{=}i)\, p(\boldsymbol{\theta}_{n+m}{=}e \,|\, \boldsymbol{\theta}_n{=}k, \boldsymbol{\theta}_0{=}i) \\ &= \sum_{k\in\mathcal{A}} p(\boldsymbol{\theta}_n{=}k \,|\, \boldsymbol{\theta}_0{=}i)\, p(\boldsymbol{\theta}_{n+m}{=}e \,|\, \boldsymbol{\theta}_n{=}k) \\ &= \sum_{k\in\mathcal{A}} A_{ik}^{(n)} A_{ke}^{(m)} \end{aligned} \tag{4.16}$$

We may interpret this as the process that starts at state i and goes to an intermediate state k after n transitons. Summing over all the possible k bridge states yields the final state e after m additional transitions. Alternatively, we can use matrix notation to express these $(n+m)$ transitions [21]

$$A^{(n+m)} = A^{(n)} A^{(m)} \tag{4.17}$$

Hence we derive the following property: the n-step transition is just the n^{th} multiplicative power on matrix A.

$$A^{(n)} = A \cdot A^{n-1} = A \cdots A = (A)^n \tag{4.18}$$

Stationarity: When $n \to \infty$, we will end up with the stationary or "equilibrium" distribution row vector $\boldsymbol{\pi}(\boldsymbol{\theta})$,

$$\boldsymbol{\pi}(\boldsymbol{\theta}) = \boldsymbol{\pi}_0(\boldsymbol{\theta}_0) \cdot \lim_{n\to\infty} A^n, \qquad \text{for all } \; \boldsymbol{\pi}_0(\boldsymbol{\theta}_0). \tag{4.19}$$

Preconditions: The preconditions for this strong statement include *ergodicity* and so-called *recurrence*. A process is called *ergodic* if the stationary state is reached for every possible initial state. In practice, we want to construct an ergodic Markov Chain which obeys the following properties [4] [22]:

1. Irreducible: all states j must be reachable from all starting points i or mathematically $A_{i,j}^m > 0 \; \forall i, j$ and $m > 0$, so that all the states will be visited eventually. By contrast, a reducible process could consist of two or more subsets of states where states from one subset cannot reach those of the other subsets.

2. Recurrent: no matter in which initial state it starts, all of the states in the chain's stationary outcome space can be revisited infinitely many times. This is in contrast to transient states which are visited only in the initial *burn-in* phase of the process; see Section 4.2.1.

3. Aperiodic: the chain is not periodic in the sense that any state i is not revisited regularly every m timesteps, with m smaller than the total number of states in the

process. To prevent periodicity, it is necessary to enforce so-called *detailed balance*, the condition that

$$\pi(i)A_{ij} = \pi(j)A_{ji} \qquad \forall i, j \tag{4.20}$$

which is necessary to reach true stationarity rather than limited cycles in which the system converges not to a single state $\boldsymbol{\pi}(\boldsymbol{\theta})$ but to a cycle of states which repeat after passing through C states, $\boldsymbol{\pi}_{t+C} = \boldsymbol{\pi}_t$.

Detailed balance can be interpreted as time reversibility: the rate of transition process $\pi(i)A_{ij}$ hopping from any state $\boldsymbol{\theta}_t = i$ to any other state $\boldsymbol{\theta}_{t+1} = j$ stays the same when the sequence of time states is reversed, $\pi(j)A_{ji}$. In other words, $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ are exchangeable variables. The consequence of imposing the Markov condition on them means that, in addition to the usual $p(\boldsymbol{\theta}_{t+1} \,|\, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \ldots, \boldsymbol{\theta}_0) = p(\boldsymbol{\theta}_{t+1} \,|\, \boldsymbol{\theta}_t)$, we must also have $p(\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{t+2}, \ldots, \boldsymbol{\theta}_s) = p(\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}_{t+1})$.

An example to illustrate the Markov Chain: We now consider an example of a Markov Chain which has only four states, $\mathcal{A} = \{\text{states } 1, 2, 3, 4\}$ and a transition matrix visualised in terms of the graph shown in Figure 4.1.

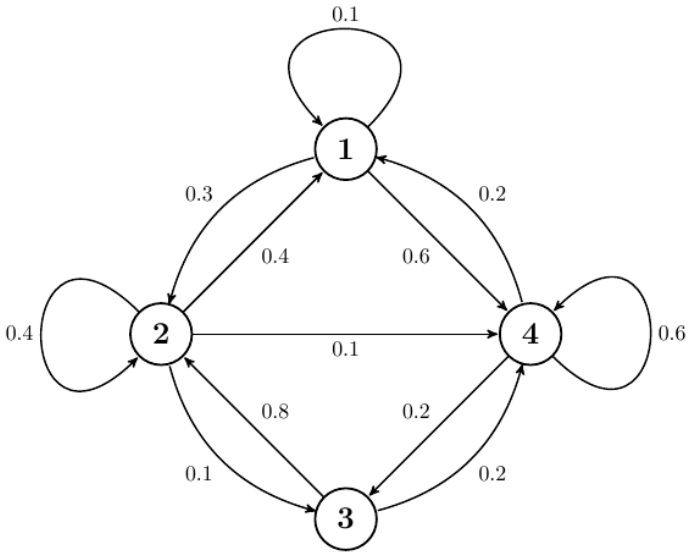


Figure 4.1: Transition diagram for toy Markov Chain model with just four states, shown as circles and transition probabilities as arrows. The latex code of this figure is taken from [23].

The transition matrix $A(\theta_{t+1} \,|\, \theta_t)$ is, by construction,

$$A(\theta_{t+1} \,|\, \theta_t) = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.3 & 0.0 & 0.6 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.0 & 0.8 & 0.0 & 0.2 \\ 0.2 & 0.0 & 0.2 & 0.6 \end{bmatrix} \tag{4.21}$$

Starting with an initial state of a row vector

$$\pi_0(\boldsymbol{\theta}_0) = (0.1, 0.32, 0.52, 0.06) \tag{4.22}$$

then the next state is given by a one step transition

$$\pi_1(\boldsymbol{\theta}_1) = \pi_0(\boldsymbol{\theta}_0)A = (0.15, 0.574, 0.044, 0.232) \tag{4.23}$$

and after the n^{th} iteration

$$\pi_n(\boldsymbol{\theta}_n) = \pi_0(\boldsymbol{\theta}_0)A^n. \tag{4.24}$$

No matter which initial $\pi_0(\boldsymbol{\theta})$ we use, the large-n product converges to

$$\pi(\boldsymbol{\theta}) = (0.207, 0.252, 0.111, 0.430) \tag{4.25}$$

and we conclude $\pi(\boldsymbol{\theta})$ is the stationary distribution. In other words, the chain starts from an arbitrary initial state, and once the process has run for many iterations, the initial state is "forgotten" and the chain reaches stationarity. Once $\pi(\boldsymbol{\theta})$ has reached this equilibrium distribution, it has the property that, with probability 1, the average of any function $f(\theta)$ of i.i.d. random variable $\boldsymbol{\theta}$ converges to its expected value,

$$\lim_{S\to\infty} \frac{1}{S} \sum_{l=1}^{S} f(\boldsymbol{\theta}_l) = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, \mathrm{d}\theta = E[f(\boldsymbol{\theta})] \tag{4.26}$$

where the $\boldsymbol{\theta}_l$ are samples from the ergodic chain and are hence regarded as originating from the target distribution.

4.1.3 Metropolis-Hastings Algorithm

Among different possible MCMC methods, the Metropolis-Hastings algorithm is the oldest of the method using an evolutionary framework which is the best suited to working in high-dimensional spaces. The scheme of MCMC methods is to formulate a Markov Chain on parameter space $\mathcal{A}(\boldsymbol{\theta})$ whose unnormalised stationary distribution $\pi^*(\boldsymbol{\theta})$ is proportional to the target distribution of interest (e.g. the posterior distribution $p(\boldsymbol{\theta} \mid D, \mathcal{H})$). By carefully choosing a proposal distribution or "kernel", non-independent samples are drawn from a succession or "chain" of states. Once the algorithm has reached equilibrium, the states generated in such chains visit regions of parameter space with a frequency proportional to $\pi^*(\boldsymbol{\theta})$.

The Metropolis-Hastings algorithm relies on a two-step stochastic process of first proposing a new state and then accepting or rejecting this proposal. First, one must invent a transition probability or *proposal distribution* $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$ which is the probability of the system proposing to move to a particular new state $\boldsymbol{\theta}'$ when it is in state $\boldsymbol{\theta}$. Given the current state $\boldsymbol{\theta}$ and the proposed new state $\boldsymbol{\theta}'$, an acceptance probability determines whether $\boldsymbol{\theta}'$ does become the next state in the chain.

The algorithm itself is as follows [22]:

1. Choose a proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$, which can be symmetric or asymmetric in the sense that for any two states $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$, $q(\boldsymbol{\theta}_a \mid \boldsymbol{\theta}_b)$ is equal to $q(\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_a)$ or not.

2. Start from an arbitrary initial value $\boldsymbol{\theta}_0$ generated from $\pi^*(\boldsymbol{\theta})$.

3. At every time t starting with $t = 0$, generate a tentative new state $\boldsymbol{\theta}'$ with probability $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t)$.

4. Define the acceptance ratio

$$\alpha = \frac{\pi(\boldsymbol{\theta}')\, q(\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}_t)} \tag{4.27}$$

and calculate the acceptance probability

$$r = \min(1, \alpha) \tag{4.28}$$

which is a ratio and can therefore also be calculated in terms of unnormalised probabilities,

$$r = \min\left(1, \frac{\pi^*(\boldsymbol{\theta}')\, q(\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}')}{\pi^*(\boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}_t)}\right), \tag{4.29}$$

which in Bayesian applications conveniently removes the need to calculate the evidence.

5. The min functions in Eq. (4.28) and Eq. (4.29) mean that states with higher probability are always accepted. States with lower probability are not rejected outright; Metropolis occasionally allows lower possible states. This is accomplished by generating a random number u from a uniform distribution $U(0, 1)$ and then deciding on acceptance according to the rule

   If $\alpha \geq 1$ or $\alpha > u$ then "accept": set the new state $\boldsymbol{\theta}_{t+1}$ to the proposed state $\boldsymbol{\theta}'$
   else "reject": set the new state $\boldsymbol{\theta}_{t+1}$ to the current state $\boldsymbol{\theta}_t$

6. Steps 3 to 5 are repeated many times.

The output of this algorithm is a "raw" sequence of states $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \ldots)$ which must be further processed as set out in Section 4.2 below.

4.1.4 Probabilistic interpretation

With the above explanations, we now consider the Metropolis-Hastings algorithm from the point of view of probability theory. We introduce a new variable c_t with the interpretation that an action or choice is made at time t [9] [21]. At time t, the chain is in state $\boldsymbol{\theta}$, and we propose a new state $\boldsymbol{\theta}'$ for time $t+1$. Using the product rule, the adapted one-step transition probability $p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)$ is then

$$p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t) = p(c_{t+1} \,|\, \boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)\, p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)$$

since states in a Markov Chain are reversible, we may confidently assert

$$p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, c_{t+1} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t) = p(c_{t+1} \,|\, \boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)\, p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t).$$

Rewriting these two equations as

$$p(c_{t+1} \,|\, \boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t) = \frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)} \tag{4.30}$$

$$p(c_{t+1} \,|\, \boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t) = \frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, c_{t+1} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)} \tag{4.31}$$

and dividing equation (4.30) by equation (4.31), we obtain the acceptance ratio α,

$$\begin{aligned}\alpha &= \frac{p(c_{t+1} \mid \boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}{p(c_{t+1} \mid \boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)} \\ &= \frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)} \frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}\end{aligned} \tag{4.32}$$

furthermore, detailed balance of MH algorithm offers the following

$$p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)\, p(\boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t) = p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)\, p(\boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t) \tag{4.33}$$

from which we obtain

$$\frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)} = \frac{p(\boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)}{p(\boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}.$$

Substituting the above equation into equation(4.32), it follows

$$\begin{aligned}\alpha &= \frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}', c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}, c_{t+1} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)} \frac{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}', c_t)}{p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}, c_t)} \\ &= \frac{p(\boldsymbol{\theta}_t{=}\boldsymbol{\theta}')p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}')}{p(\boldsymbol{\theta}_t{=}\boldsymbol{\theta})p(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta})} \\ &= \frac{\pi(\boldsymbol{\theta}_t{=}\boldsymbol{\theta}')q(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta} \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}_t{=}\boldsymbol{\theta})q(\boldsymbol{\theta}_{t+1}{=}\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t{=}\boldsymbol{\theta})} \\ &= \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}\end{aligned} \tag{4.34}$$

This is the fourth step of the Metropolis-Hastings algorithm. Here c_t is the acceptance probability at time t and certainly, $c_t = 1$.

We then calculate the acceptance probability, $r = \min\left(1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}\right)$. The relative acceptance probability of two adjacent states [24] is

$$\begin{aligned}\frac{r(\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1})}{r(\boldsymbol{\theta}_{t+1} \to \boldsymbol{\theta}_t)} &= \frac{\min\left(1, \dfrac{\pi(\boldsymbol{\theta}_{t+1})\, q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}{\pi(t)q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)}\right)}{\min\left(1, \dfrac{\pi(\boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)}{\pi(\boldsymbol{\theta}_{t+1})q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}\right)} \\ &= \begin{cases} 1 \Big/ \left[\dfrac{\pi(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)}{\pi(\boldsymbol{\theta}_{t+1})q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}\right] = \dfrac{\pi(\boldsymbol{\theta}_{t+1})q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}{\pi(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)}, & \text{if} \quad \dfrac{\pi(\boldsymbol{\theta}_{t+1})q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}{\pi(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)} > 1 \\ \dfrac{\pi(\boldsymbol{\theta}_{t+1})q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}{\pi(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)}, & \text{otherwise} \end{cases} \\ &= \frac{\pi(\boldsymbol{\theta}_{t+1})}{\pi(\boldsymbol{\theta}_t)} \frac{q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})}{q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)}.\end{aligned} \tag{4.35}$$

If a symmetric proposal distribution is used,

$$q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') \tag{4.36}$$

for example a Gaussian distribution centered on the current state, then the relative acceptance simplifies even more,

$$\frac{r(\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1})}{r(\boldsymbol{\theta}_{t+1} \to \boldsymbol{\theta}_t)} = \frac{\pi(\boldsymbol{\theta}_{t+1})}{\pi(\boldsymbol{\theta}_t)}, \tag{4.37}$$

i.e. the sequential states will be accepted proportionally to their corresponding distribution $\pi(\boldsymbol{\theta})$ if we run the iterations long enough.

4.2 On Implementing MCMC

No method can claim universal success or applicability. A few problems encountered with MCMC are disussed here.

4.2.1 Immediate consequences of stochastic-process sampling

Burn-In: Burn-in is the colloquial name for the period where the Markov Chain has not yet converged. Output from the burn-in period is not stored as it does not represent the target stationary distribution. The length of the burn-in time will be examined in Section 4.2.2.3.

Proposal Distribution: The proposal distribution plays a crucial role in the efficiency of MCMC methods, as we want to assure the sampling space covers the parameter space as much as possible and not missing important parts [25]. One important aspect is the variance of proposal distribution q. On the one hand, a proposal function with a small variance (such as a Gaussian with small width of σ) accepts most of the proposals as acceptance probabilities are large; however, since the new states $\boldsymbol{\theta}_{t+1}$ are close to the old one $\boldsymbol{\theta}_t$, this chain will converge very slowly and a finite sample may not be representative. States from a proposal function with large variance, on the other hand, will be rejected most of the time because probabilities are low, and so the chain stays in the same state for long periods and taking a big leap once in a while. In both cases, the parameter space is not fully searched which has the risk of missing modes in multi-modal target distribution.

A rule of thumb criterion for a good proposal function is that the acceptance rate should be between 25% for high-dimensional models and about 50% for models of dimension 1 or 2 [26].

Adaptive MCMC is an algorithm used to increase accuracy and efficiency by adapting parameters of the proposal distribution while the simulation is running [27]. Information provided by earlier samples is used to update parameters in order both to obtain convergence and ensure an acceptable acceptance rate.

Effective sample size and the autocorrelation function The raw states $\{\boldsymbol{\theta}_t\}_{t=1}^{S}$ produced by MCMC are clearly not independent, which is a big but unavoidable disadvantage of the method, therefore, the accuracy of this estimation is worth scrutiny. The theoretical

variance of MCMC estimator $\langle f \rangle$ of any function $f(\boldsymbol{\theta})$ [22] is

$$\begin{aligned}
\text{var}_{MCMC}(\langle f \rangle) &= E[(\langle f \rangle - \mu)(\langle f \rangle - \mu)] = E\Big[\frac{1}{S}\sum_{t=1}^{S}(f(\boldsymbol{\theta}_t) - \mu)\frac{1}{S}\sum_{k=1}^{S}(f(\boldsymbol{\theta}_k) - \mu)\Big] \\
&= E\Big[\frac{1}{S^2}\sum_{t,k}^{S}(f(\boldsymbol{\theta}_t) - \mu)(f(\boldsymbol{\theta}_k) - \mu)\Big] \\
&= \frac{1}{S^2}\sum_{t=1}^{S} E\Big[(f(\boldsymbol{\theta}_t) - \mu)^2\Big] + \frac{1}{S^2}\sum_{t\neq k} E\Big[(f(\boldsymbol{\theta}_t) - \mu)(f(\boldsymbol{\theta}_k) - \mu)\Big] \\
&= \frac{1}{S}E[(f(\boldsymbol{\theta}) - \mu)^2] + \frac{2}{S^2}\sum_{t=1}^{S-1}\sum_{k=t+1}^{S}\text{Cov}\Big[f(\boldsymbol{\theta}_t), f(\boldsymbol{\theta}_k)\Big]
\end{aligned}$$

and after changing variables to $\Delta t = k - t$,

$$\begin{aligned}
\text{var}_{MCMC}(\langle f \rangle) &= \frac{1}{S}\text{var}(f) + \frac{2}{S^2}\sum_{t=1}^{S-1}\sum_{\Delta_t=1}^{S-t}\text{Cov}\Big[f(\boldsymbol{\theta}_t, f(\boldsymbol{\theta}_{t+\Delta_t})\Big] \\
&= \text{var}_I(\langle f \rangle) + \frac{2}{S^2}\sum_{\Delta_t=1}^{S-1}(S - \Delta_t)\text{Cov}\Big[f(\boldsymbol{\theta}_t), f(\boldsymbol{\theta}_{t+\Delta_t})\Big]
\end{aligned}$$

where the first term $\text{var}_I(\langle f \rangle)$ is the theoretical variance when the chain variables are independent as assumed in Monte Carlo methods, the second term lies in the fact that $\text{Cov}\Big[f(\boldsymbol{\theta}_t, f(\boldsymbol{\theta}_{t+\Delta_t})\Big]$ does not depend on t by stationarity [28]. For large S, this tends asymptotically to

$$\text{var}_{MCMC}(\langle f \rangle) = \text{var}_I(\langle f \rangle) + \frac{2}{S}\sum_{\Delta_t=1}^{S}\text{Cov}\Big[f(\boldsymbol{\theta}_t), f(\boldsymbol{\theta}_{t+\Delta_t})\Big] \tag{4.38}$$

Looking at the actual MCMC theoretical variance sample average $\langle f \rangle$, we see that it can therefore be expressed formally in terms of a theoretical variance for independent samples $\text{var}_I(\langle f \rangle)$ plus a sum over theoretical covariances which estimate the correlations or non-independence. The *Effective Sample Size* (ESS) is correspondingly defined as

$$\text{ESS} = \frac{\text{var}_I(\langle f \rangle)}{\text{var}_{MCMC}(\langle f \rangle)} \tag{4.39}$$

While the above calculations were for theoretical expectations, variances and covariances, these can be replaced (under a frequentist argument) by the corresponding finite-S sample averages. The above motivates why in practice one measures the so-called *autocorrelation function* [29]

$$\rho(\Delta_t) = \frac{\frac{1}{S-\Delta t}\sum_{t=1}^{S-\Delta_t}(f(\boldsymbol{\theta}_t) - \langle f \rangle)(f(\boldsymbol{\theta}_{t+\Delta_t}) - \langle f \rangle)}{\frac{1}{S-1}\sum_{t=1}^{S}(f(\boldsymbol{\theta}_t) - \langle f \rangle)^2}$$

which is clearly an approximation to the theoretical ratio

$$\rho(\Delta_t) \simeq \frac{\mathrm{Cov}(f(\boldsymbol{\theta}_t), f(\boldsymbol{\theta}_{t+\Delta_t}))}{\mathrm{var}_I(f(\boldsymbol{\theta}_t))}. \tag{4.40}$$

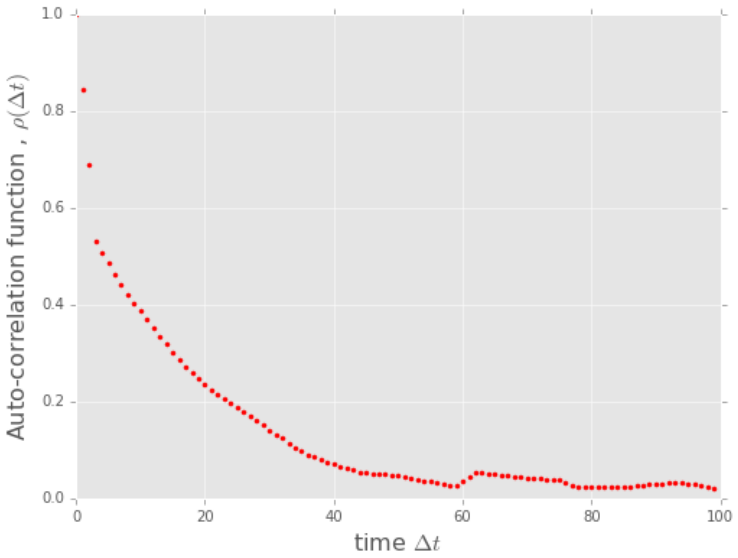


Figure 4.2: Autocorrelation of samples of $\boldsymbol{\theta}_t$ vs time difference Δt. We can see that it displays a decreasing exponential curve and the correlation decreases to zero after about 60 iterations.

Thinning: While the fact that sample points $\boldsymbol{\theta}_t$ are correlated is clearly a big disadvantage, the advantages of the MCMC method are so great that one prefers to live with them and try to reduce their influence by throwing more computing power at the problem. *Thinning* is one of the common methods used to reduce auto-correlation: We only keep the every τ_{eq}^{th} sample and throw away the intermediate states, thereby creating a "thinned" sample with fewer states but with a quasi-independence which the original chain did not have. We then use the thinned sample to estimate the desired function.

4.2.2 Convergence Diagnostics of MCMC

In addition to the above issues, *convergence* of the Markov Chain is the decisive criterion for MCMC performance evaluation; we will discuss convergence in the next section.

Whether or not the target posterior distribution can be represented by a stationary distribution of MCMC samples is the key to make valid inferences. In practice, one can never run infinite iterations which would guarantee that the sample average equals the theoretical expectation value. Therefore, it is necessary to investigate the minimum number of iterations to ensure a credible approximation to the interested distribution. The assessment of convergence is known as *convergence diagnostics*. The main idea of convergence diagnostics is to analyse statistical properties of samples drawn from the chain.

4.2.2.1 Simple Methods

The most straightforward way to check convergence is the *trace plot* where one plots the value of the random variable (the parameters in our case) against time. Excluding the burn-in phase, the trace plot indicates convergence if the parameter values stay in a narrow band with no dramatic fluctuations, one can also investigate every set of m iterations[2], noting that the system cannot be said to be converged until all of the variables satisfy the convergence criterion [30].

Alternatively, convergence can be inferred by running multiple chains, rather than a single long chain and observing whether or not they all arrive in the same band [31]. This is one signature of Markov chains called *ergodicity*, where all states in the chain is aperiodic, recurrent and non-null [22]. A typical trace plot is shown in Figure 4.3.

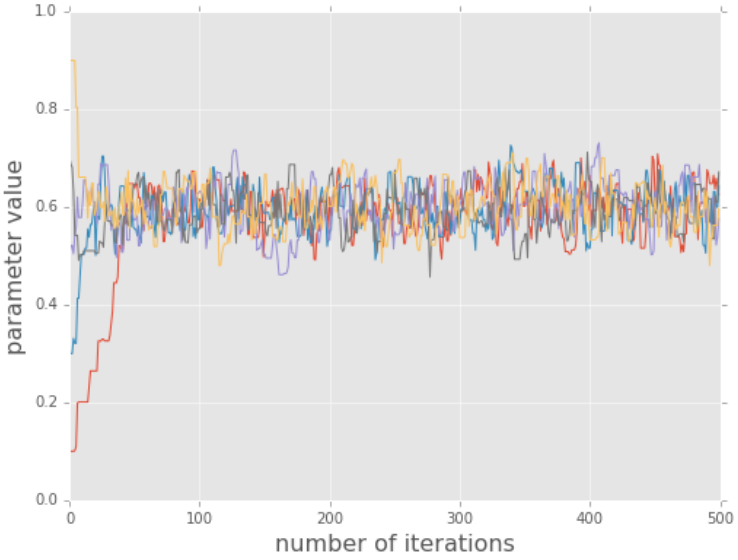


Figure 4.3: MCMC trace plots for a single variable : chains start with various dispersed initial values. After a few steps, fewer than 100 in the present example, they all stay within the same band around mean value 0.6.

One persistent problem with chain evaluation can occur: the chain may linger within some local minimum for an extended period of time even while providing excellent results with the above methods. It may appear to an observer that the chain has reached a state of stability when in fact it has not. This phenomenon of metastability reveals that further research is required when examining convergence. Below we consider a few prominent methods from the MCMC literature.

4.2.2.2 Geweke

Geweke [32] proposed a method to assess diagnostic convergence usually used for burn-in estimation. It divides the samples into three different time periods in a single chain, the first S_A samples, the last S_B samples and the remainder. We start by defining the total number

[2]The value of m is user defined.

of samples as S, $(S_A + S_B \leq S)$, where the parameter of interest is $\boldsymbol{\theta}$. Geweke suggests $S_A = 0.1S$ and $S_B = 0.5S$. He defines a quantity comparing the mean and variance of the two segments,

$$Z = \frac{\bar{\boldsymbol{\theta}}_A - \bar{\boldsymbol{\theta}}_B}{\sqrt{\text{var}(\boldsymbol{\theta}_A) + \text{var}(\boldsymbol{\theta}_B)}}. \tag{4.41}$$

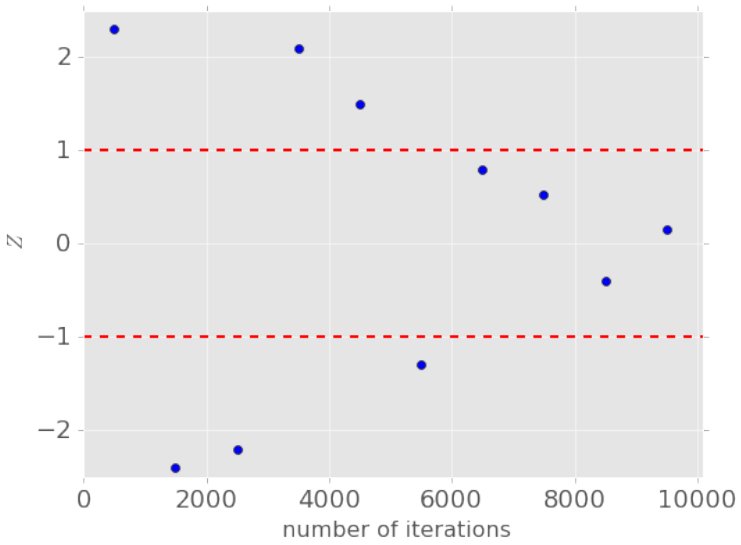


Figure 4.4: Z values for various portions of the chain, comparing with the last 50% [33]. In this example, we have 20000 iterations in total. Segments from the first half are chosen as $\bar{\boldsymbol{\theta}}_A$, while $\bar{\boldsymbol{\theta}}_B$ is the bottom half. Initial Z scores represent portions that are close to the start; previous segments are then gradually excluded to check burn-in. The fluctuations of Z which lie within the interval [-1,1] indicate the point in time at which convergence has been reached, in the present case after 6500 iterations.

If the distribution of Z is approximately Gaussian $\mathcal{N}(0, 1)$, the majority of points should not deviate more than two standard deviations from the mean. When the two fractions of the samples can be considered similar, the chain is deemed to have converged. In practice, the initial time (from where we want to observe) is user defined.

An initial time that is too late in the process may result in part of the converged samples being discarded as burn-in. The antithetical case where the selected time is too early may result in the parameters not exploring the entire parameter space.

4.2.2.3 Raftery and Lewis

Raftery and Lewis [34] provided a method to answer three frequently asked questions:

- *How much time* is needed for a sufficient simulation run?
- *How much space* is required between every retained k^{th} sample to avoid "stickiness" or correlations?
- *When should the initial* burn-in iterations be discarded?

We note that it is very unlikely that a Markov chain can start at an arbitrary point and originate from a stationary distribution of this chain. Furthermore, this method can be used to determine the desired length of iterations for a desired precision, and the output from this method can diagnose slow convergence or lack of convergence.

4.2.2.4 Gelman and Rubin

Gelman and Rubin [35] presented a monitoring convergence method by running m parallel chains, each starting with dispersed points over target distribution, and comparing the variance within a chain (W), and between chains (B) to check whether convergence has been achieved . This method contains two parts:

1. Generating multiple over-dispersed initial values

- Locate the modes of the target distribution (possibly multivariate) using optimization methods (MAP, Laplace approximation) or EM algorithm [36], so that important regions will not be neglected.
- Generating N samples from a mixture of t-distribution where these modes are centered at, then use SIR (Sampling Importance Resampling) to obtain m samples, where the histogram of m importance-resampled draws represent the target distribution.

2. Using these initial values as the start to run M independent multiple chains. Each chains run for $2S$ iterations, of which the first S samples are discarded as burn-in [37]. For each scalar quantity of interest, calculate the variance B between M chains,

$$B = \frac{S}{M-1}\sum_{j=1}^{M}(\bar{\boldsymbol{\theta}}_{.j} - \bar{\boldsymbol{\theta}}_{..}) \tag{4.42}$$

where $\bar{\boldsymbol{\theta}}_{.j} = \frac{1}{S}\sum_{i=1}^{S}\boldsymbol{\theta}_{ij}$, $\bar{\boldsymbol{\theta}}_{..} = \frac{1}{M}\sum_{j=1}^{M}\bar{\boldsymbol{\theta}}_{.j}$
and the intra-chain variance W is

$$W = \frac{1}{M}\sum_{j=1}^{M}s_j^2 \tag{4.43}$$

where sample variance $s_j^2 = \frac{1}{S-1}\sum_{i}^{S}(\theta_{ij} - \bar{\theta}_{.j})$. The estimate posterior variance is defined as

$$V = \frac{S-1}{S}W + \frac{M+1}{SM}B \tag{4.44}$$

The potential scale reduction factor (PSRF) is the square root of V/W, refined by Brooks and Gelman(1997) [38],

$$R = \sqrt{\frac{d+3}{d+1}\cdot\frac{V}{W}} = \sqrt{\frac{d+3}{d+1}\Big(\frac{S-1}{S} + \frac{M+1}{SM}\frac{B}{W}\Big)} \tag{4.45}$$

where $d = \frac{2V^2}{\mathrm{var}(V)}$. If all the M sequences have converged, the PSRF should close to 1, less than 1.1 in practice.

4.3 Calculating Evidence

The Laplace approximation to evidence of Eq. (3.49) is convenient but may not be accurate, depending on the specific form of the likelihood and prior. Monte Carlo integration of Eq. (3.48) is then a necessity, especially when the number of parameters K is large.

4.3.1 Direct Monte Carlo integration

The easiest way of computing the evidence would be to draw samples $\boldsymbol{\theta}^{(j)}$ from the prior distribution $p(\boldsymbol{\theta} \,|\, \mathcal{H})$ and use them in an average over likelihoods,

$$p(D \,|\, \mathcal{H}) \approx \frac{1}{S} \sum_{j=1}^{S} p(D \,|\, \boldsymbol{\theta}^{(j)}), \qquad \boldsymbol{\theta}^{(j)} \sim p(\boldsymbol{\theta} \,|\, \mathcal{H}). \tag{4.46}$$

However, the variance of this estimator is large because it is sensitive to the choice of prior. If the posterior is concentrated relative to the prior in a smaller subspace of the parameter space, most of the samples drawn will have small likelihood values, and as a result the estimator is dominated by only those few values $\boldsymbol{\theta}^{(j)}$ which happen to have large likelihood values.

4.3.2 Harmonic Mean Approximation

Let $Z = p(D \,|\, \mathcal{H})$ be the evidence. Bayes' Theorem can be written as

$$Z \cdot p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) = p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) \cdot p(\boldsymbol{\theta} \,|\, \mathcal{H}).$$

Dividing both sides by the likelihood $p(D \,|\, \boldsymbol{\theta}, \mathcal{H})$ and integrating over parameter space Ω, the integral over the prior is 1 due to normalisation, so

$$Z \cdot \int_{\Omega} \frac{p(\boldsymbol{\theta} \,|\, D, \mathcal{H})}{p(D \,|\, \boldsymbol{\theta}, \mathcal{H})} \,\mathrm{d}\boldsymbol{\theta} = \int_{\Omega} p(\boldsymbol{\theta} \,|\, \mathcal{H}) \,\mathrm{d}\boldsymbol{\theta} = 1,$$

then the inverse of the evidence equals the expectation value of the inverse likelihood weighted by the posterior,

$$\frac{1}{Z} = \int_{\Omega} \frac{p(\boldsymbol{\theta} \,|\, D, \mathcal{H})}{p(D \,|\, \boldsymbol{\theta}, \mathcal{H})} \,\mathrm{d}\boldsymbol{\theta} = E\left[\frac{1}{p(D \,|\, \boldsymbol{\theta}, \mathcal{H})}\right]_{p(\boldsymbol{\theta} \,|\, D, \mathcal{H})}.$$

Sampling from the posterior distribution $p(\boldsymbol{\theta} \,|\, D, \mathcal{H})$ yields the Harmonic Mean approximation of evidence,

$$Z = p(D \,|\, \mathcal{H}) = \left[\frac{1}{S} \sum_{j=1}^{S} \frac{1}{p(D \,|\, \boldsymbol{\theta}^{(j)})}\right]^{-1}, \quad \boldsymbol{\theta}^{(j)} \sim p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) \tag{4.47}$$

Unfortunately, this method is also rather unstable because outliers with small likelihoods will dominate the average [39].

4.3.3 Nested Sampling

Skilling [40] invented a method to calculate evidence by means of variable transformation, in which a multi-dimensional integral over parameter space is transformed into a one-dimensional integral over likelihood space. Define $\mathrm{d}X = p(\boldsymbol{\theta} \,|\, \mathcal{H})\,\mathrm{d}\boldsymbol{\theta}$, and write the likelihood function as $L(\boldsymbol{\theta}) = p(D \,|\, \boldsymbol{\theta}, \mathcal{H})$. The variable $X(\lambda)$ is the cumulative prior probability over that subspace of the parameter space whose likelihood exceeds a given threshold $\lambda > 0$,

$$X(\lambda) = \int_{L(\boldsymbol{\theta})>\lambda} p(\boldsymbol{\theta} \,|\, \mathcal{H})\,\mathrm{d}\boldsymbol{\theta} \tag{4.48}$$

As λ increases, the function $X(\lambda)$ decreases from 1 to 0, and likewise the inverse function $\lambda(X)$ is a monotonically decreasing function of X [41]; see Figure 4.5. Therefore, the evidence can be written as a one-dimensional integral over X,

$$Z = \int p(D \,|\, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \,|\, \mathcal{H})\,\mathrm{d}\boldsymbol{\theta} \;=\; \int_0^1 \lambda(X)\,\mathrm{d}X. \tag{4.49}$$

The step-by-step nested sampling algorithm works as follows:

1. Draw a set of N points $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ from the prior without any criteria value λ.
2. Calculate the likelihood value $p(D \,|\, \boldsymbol{\theta}^{(i)})$ of each point $\boldsymbol{\theta}^{(i)}$ and order them by their likelihood values.
3. Eliminate the point $\boldsymbol{\theta}'$ from the active set which has the lowest likelihood $L(\boldsymbol{\theta}')_{\text{lowest}}$, and store $L_j = L_{\text{lowest}}$.
4. Generate a new point $\boldsymbol{\theta}_{\text{new}}$ from the prior and accept it if $L(\boldsymbol{\theta}_{\text{new}}) > L_j$. Including the accepted point in the active set and compute the new mass X_{j+1} from the latest active set.
5. Repeat steps 2 to 4 until termination.

Thus the algorithm explores the nested shells of likelihood contours shown by example in Figure 4.6 as the prior volume decreases. The evidence is then estimated by some one-dimensional numerical integration technique,

$$Z = \sum_{j=1}^{S} L_j \Delta_j \tag{4.50}$$

where $\Delta_j = X_{j-1} - X_j$ or $\frac{1}{2}(X_{j-1} - X_{j+1})$, and $X_0 = 1$. Δ_j is the width of successive prior mass points. L_j is the lowest likelihood value at each iteration.

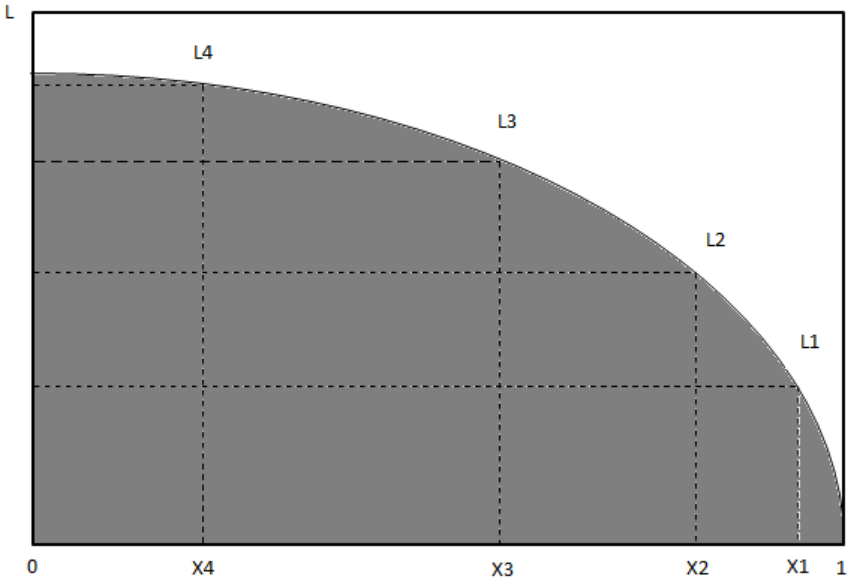


Figure 4.5: Relation between the variable λ and the cumulative prior mass X. The shaded area under the curve is desired evidence.

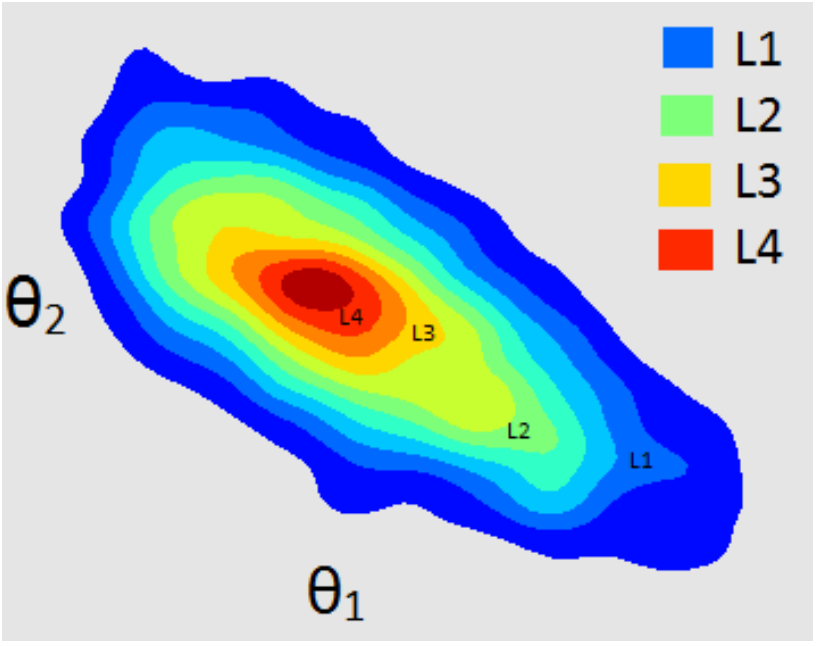


Figure 4.6: Likelihood contour plot over 2-D parameter space, prior volumes are outlined by iso-likelihood in corresponding to points $L1, \ldots, L4$ of Figure 4.5. The "hotter" the volume, the higher the likelihood value.

We know that the prior mass X_j is uniformly distributed over the interval $[0, 1]$, $X_0 = 1$ and constrained by $X_j < X_{j-1}$. The probability of drawing a maximum value X from a sample of size N is given by the binomial distribution (B.6) in Appendix B

$$p(X) = NX^{N-1}, \tag{4.51}$$

then at the j-th iteration,

$$X_j = \prod_{t=1}^{j} p(X_t) \cdot 1.$$

Since $E(\log X) = -1/N$ and $\sigma(\log X) = 1/N$ [40], the variable is distributed approximately as

$$\log X_j \approx -(j \pm \sqrt{j})/N$$

and the sequence of X_j can be evaluated through

$$X_j = \exp(-j/N). \tag{4.52}$$

The posterior can be represented by points which have the lowest likelihood value at each iteration, the posterior weight for each point is

$$p(\boldsymbol{\theta})_j = \frac{L_j \Delta_j}{Z} \tag{4.53}$$

where Z is the final converged value calculated through Eq. (4.50). The posterior mean and variance for the parameters are computed as

$$E(\boldsymbol{\theta}) = \frac{1}{Z} \sum_{j=1}^{S} L_j \Delta_j \boldsymbol{\theta}_j \tag{4.54}$$

$$\mathrm{var}(\boldsymbol{\theta}) = \left(\frac{1}{Z} \sum_{j=1}^{S} L_j \Delta_j \boldsymbol{\theta}_j^2 \right) - E(\boldsymbol{\theta})^2 \tag{4.55}$$

However, one disadvantage of the algorithm is the absence of a rigorous approach to the convergence criterion.

How to generate X_j samples The biggest challenge in Nested Sampling is to draw points from the prior parameter distribution under the constraint that the likelihood value exceeds the lowest likelihood, $L > L_j$ or $X < X_j$ at each iteration. The Metropolis-Hastings algorithm can be used to meet the requirement, but to avoid blindly drawing from the prior, ellipsoid sampling was introduced in Ref. [42]. Defining the covariance matrix of the current set of active points as $\boldsymbol{C}$, the current points enclosed in a hyper-ellipsoid can be expressed as

$$\boldsymbol{\theta}^T (\boldsymbol{C})^{-1} \boldsymbol{\theta} \leq k, \tag{4.56}$$

where $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_K$, matrix element $C_{i,j} = \frac{1}{N-1} \sum_{n=1}^{N} (\theta_i^{(n)} - \mu_i)^T (\theta_j^{(n)} - \mu_j)$, $i, j = 1, 2, \ldots, K$, covariance matrix $\boldsymbol{C}$ has eigenvector A which gives the orientation of the ellipse and eigenvalue λ, and k is a user defined enlargement factor, $k = \max[\boldsymbol{\theta}^T (\boldsymbol{C})^{-1} \boldsymbol{\theta}]$. Many ellipsoids will be used to bound the points if a single ellipsoid is not a well approximation to cover the active set.

New points are drawn which exceed this iso-likelihood bound ellipsoidal. Instead of drawing points $\boldsymbol{\theta}$ from a K-ellipsoid, one can also draw points $\boldsymbol{X}$ from a unit K-sphere and then map them to the surface of this ellipsoid. In the case of multi-modal, the single ellipsoid is splitted into multi-ellipsoids, in which actives points form separate regions with relatively high likelihoods.[43]

Chapter 5

Numerical experiments with simple models and conclusions

In this chapter, we apply some of the insights gained to a few simple numerical examples. Due to the lack of time and computer power to fully utilise Bayesian estimation, some simulation experiments did not achieve the desired accuracy and some investigations had to be postponed.

The generic situation we consider here is that of a few Gaussian peaks with a flat background of counts, mimicking a situation where one would know from theory that peaks exist at known locations. The simplest task would be to infer the peak's amplitudes in various situations where these amplitudes are large or small compared to the level of background noise. In addition, it may happen that a given data set looks like it has a peak at a location which in reality is not a peak at all but "spurious", a mere fluctuation. In this case, one would want the tools to apply a model which can accommodate the location of the spurious peak.

The simulations set out below follow the usual two-step process. In the first step, we create some synthetic data based on a choice for the number of real peaks and their amplitudes; standard deviations and locations are considered constants throughout. In the second step, the created data is analysed with a variety of hypotheses, including the one which was used to actually simulate the data, but also others.

We have designed the following four model hypotheses to analyse each simulated data set:

1. $\mathcal{H}_1$: the data are Poisson distributed with rate parameter $\lambda_b = f(x_b, \boldsymbol{\beta})$, $n_b \sim \text{Poisson}(f_b(\boldsymbol{\beta}))$. The model for the rate parameter is the two-peak function,

$$f(x_b, \boldsymbol{\beta}) = \beta_1 \exp\left\{ -\frac{(x_b - \mu_1)^2}{2\sigma^2} \right\} + \beta_2 \exp\left\{ -\frac{(x_b - \mu_2)^2}{2\sigma^2} \right\} + \beta_c, \tag{5.1}$$

   where β_1, β_2 are the signal amplitudes and β_c represents the background noise, and μ_1, μ_2 and σ are known constants.

2. $\mathcal{H}_2$: $n_b \sim \text{Poisson}(f_b(\boldsymbol{\beta}))$. Compared to hypothesis $\mathcal{H}_1$, two peaks are accompanied by a spurious peak with amplitude β_3 and μ_3, the width of the third peak is assumed to be the same as the other two peaks for simplicity, and μ_3 is hard-coded in view of the

data in this thesis, it could be an interested parameter in practice.

$$f(x_b, \boldsymbol{\beta}) = \beta_1 \exp\left\{-\frac{(x_b-\mu_1)^2}{2\sigma^2}\right\} + \beta_2 \exp\left\{-\frac{(x_b-\mu_2)^2}{2\sigma^2}\right\} + \beta_3 \exp\left\{-\frac{(x_b-\mu_3)^2}{2\sigma^2}\right\} + \beta_c \tag{5.2}$$

3. $\mathcal{H}_3$: the data was generated according to Negative Binomial distributions in each bin, $n_b \sim NBD(f_b(\boldsymbol{\beta}), r)$. The signal has two peaks and the model function is the same as Eq. (5.1).

4. $\mathcal{H}_4$: $n_b \sim NBD(f_b(\boldsymbol{\beta}), r)$. The spectrum has two peaks and a spurious peak, and we use the same model function as in $\mathcal{H}_2$, Eq. (5.2).

As was already discussed in Chapter 3, the likelihood functions and posterior of the Poisson models $\mathcal{H}_1$ and $\mathcal{H}_2$ are given by Eq. (3.68) and Eq. (3.72), while the likelihood functions and posterior of NB models $\mathcal{H}_3$ and $\mathcal{H}_4$ are described by Eq. (3.84) and Eq. (3.85).

In the following simulations, the one-dimensional variable x ranges over the interval $[0, 20]$, with varying bin numbers B and corresponding bin widths $\varepsilon = 20/B$.

We used Python package `nestle` [43] to calculate model evidence on the $\log_e$ scale, nestle uses nested sampling with user defined method "MCMC" or "ellipsoid" as introduced in Chap 4, and package `corner` [44] to contour plot posterior parameter distributions.

Test problem 1

The first data set is a "low signal" scenario. Data counts are generated from a Poisson distribution $\mathcal{H}_1$ with parameters $\beta_1 = 1.1, \beta_2 = 1.1, \beta_3 = 0$ and $\beta_c = 1.0$, so that the peaks barely exceed the background. The following table summarises the simulation results. B is the number of bins we set in the data simulation. Note that the total number of counts $N = \sum_b n_b$ is not set by hand but can fluctuate from run to run as the n_b do. For the purposes of the analysis, it is of course fixed. We used a uniform prior for the dispersion parameter r with minimum zero and two upper bounds $r_{\max}$ as specified below.

DATA SET 1 simulated from Poisson($D \mid \beta_1 = 1.1, \beta_2 = 1.1, \beta_3 = 0., \beta_c = 1.$) N=22, B=20				
estimates	true values	$\mathcal{H}_1$	$\mathcal{H}_3$	$2\log\mathcal{B}_{13}$
In $\mathcal{H}_3$, prior bound $r_{\max}$ =1				
$\hat{\beta}_1$:	1.1	1.41±0.92	1.57±0.95	
$\hat{\beta}_2$:	1.1	1.34±0.90	1.57±0.95	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	1.	1.11±0.24	1.29±0.45	
$\hat{r}$:	—	—	0.83±0.14	
$\log\hat{Z}$:	—	-28.021±0.025	-32.122±0.007	$> 6, \mathcal{H}_1$
In $\mathcal{H}_3$, prior bound $r_{\max}$ =100				
$\hat{\beta}_1$:	1.1	1.41±0.92	1.42±0.92	
$\hat{\beta}_2$:	1.1	1.34±0.90	1.36±0.91	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	1.	1.11±0.24	1.12±0.24	
$\hat{r}$:	—	—	52.54±27.72	
$\log\hat{Z}$:	—	-28.021±0.025	-28.108±0.007	< 2

- We use Eq. (3.90) and apply the judging rule in Section 3.1.6 to select the model. In the first $\log\mathcal{B}_{13}$ block, $\log\mathcal{B}_{13} = \log\hat{Z}(\mathcal{H}_1) - \log\hat{Z}(\mathcal{H}_3) \approx 4.1$ and hypothesis $\mathcal{H}_1$ is thus strongly favored, while in the second block, the calculated $2\log\mathcal{B}_{13} < 2$ and it is indeterminate to say which model is better.

- The estimated values for $\beta_1, \beta_2, \beta_c$ and r are calculated from the `nestle` package using Eqs. (4.54); they are compatible with the true values for both $\mathcal{H}_1$ (Poisson model) and $\mathcal{H}_3$ (Negative Binomial model) but the variance is large. This is to be expected, given the very low total number of counts N and the even lower number which actually make up the "peaks", therefore hypotheses $\mathcal{H}_2$ and $\mathcal{H}_4$ are ignored.

  In addition, this simulation used an inappropriate binning and bin number B: the bin width ϵ is greater than the Gaussian peak width σ. As a result, the peak cannot be resolved with this coarse binning or equivalently too few events occurred, and a Gaussian shape model definitely disagrees with simulated the data. In the following test problems, we have corrected this and increase B to ensure that $\epsilon < \sigma$.

- We can see that the evidence is sensitive to the range of the prior of r. Setting a narrow prior with $r_{\max} = 1$ implies that r is constrained to small values; we have effectively

forced the model to remain significantly non-Poissonian. For the wider prior with $r_{\max} = 100$, this constraint is removed and the algorithm therefore increases r to a large value, so that the Negative Binomial model has evolved into a near-Poissonian one: r has become irrelevant and the variance on r is correspondingly large.

As shown in Figure 5.1, the forced-NB model has a lower evidence than the pure Poisson model of $\mathcal{H}_1$, while the near-Poissonian NB model has about the same evidence as $\mathcal{H}_1$.

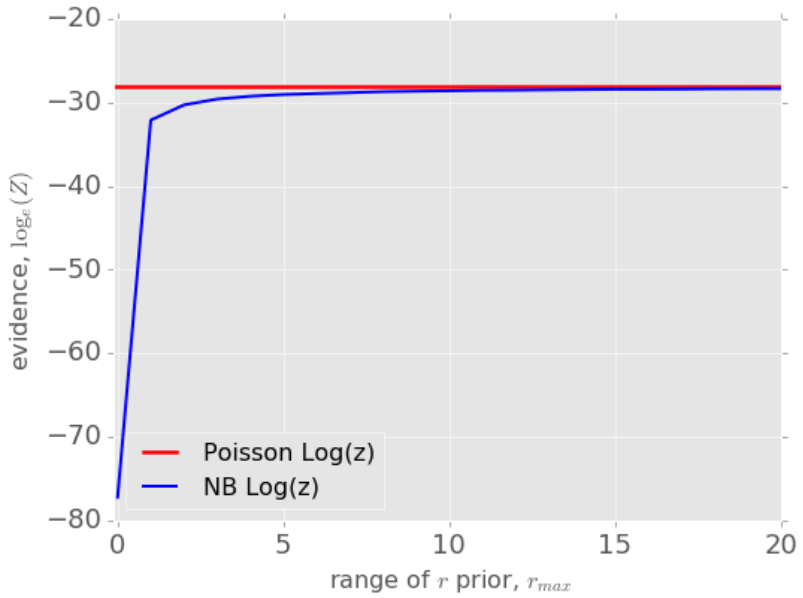


Figure 5.1: Relationship between log evidence and the range of the r prior. The evidence for the NB model grows quickly with $r_{\max}$ for small values of $r_{\max}$ and then converges with the evidence value of the Poisson model.

In asymptotic analysis Eq.(3.48), the evidence is inverse proportional to the range of the prior parameter if a uniform distribution is used. However, the above figure reveals that no matter how large $r_{\max}$ is, the evidence of the NB model approaches a fixed value. The evidence calculated from the `nestle` package seems to violate the theory; in fact, the nested sampling algorithm starts by drawing prior samples from a hypercube if uniform priors are given. A wide prior results in a large hypercube volume, so that for the same number of points S in the sample, the parameter space is sampled very roughly in that case. Points with low likelihood values are filtered and are not counted in the evidence summation.

Test problem 2

The second data set is also generated from a Poisson distribution. Now we have a stronger signal and more counts, resulting in better fitting. The same pattern regarding the evolution of the NB model emerges as in Test problem 1 and the evolution of the NB model is therefore independent of the number of counts.

DATA SET 2 simulated from Poisson$(D \mid \beta_1 = 4., \beta_2 = 10., \beta_3 = 0., \beta_c = 4.)$ N=564, B=100				
estimates	true values	$\mathcal{H}_1$	$\mathcal{H}_3$	$2\log\mathcal{B}$
In NB model, $r_{\max} = 1.$				
$\hat{\beta}_1$:	4.	2.53±0.90	3.91±2.75	
$\hat{\beta}_2$:	10.	9.69 ±1.23	12.00±5.20	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	4.	4.16 ± 0.26	4.19±0.60	
$\hat{r}$:	—	—	0.98 ± 0.02	
$\log\hat{Z}$:	—	-229.545±0.017	-286.281±0.017	$> 10, \mathcal{H}_1$
In NB model, $r_{\max} = 100.$				
$\hat{\beta}_1$:	4.	2.53 ± 0.90	2.54±0.95	
$\hat{\beta}_2$:	10.	9.69 ± 1.23	9.75±1.35	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	4.	4.16 ± 0.26	4.15±0.27	
$\hat{r}$:	—	—	66.72 ± 20.57	
$\log\hat{Z}$:	—	-229.531±0.017	-230.530±0.017	< 2

Same as test problem 1: the evidence in NB model grows as the range of r prior increases.

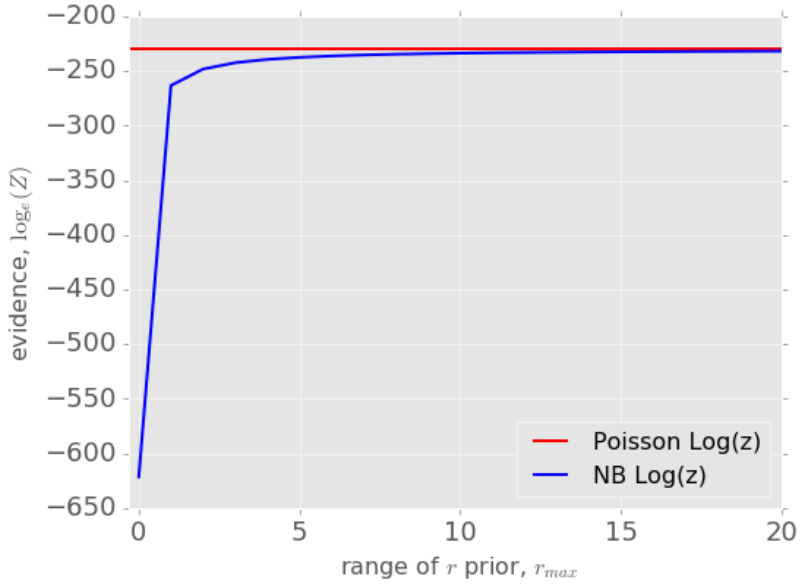


Figure 5.2: Relationship between $\log_e(Z)$ and $r_{\max}$.

Test problem 3

In this case, a much larger number of bins was used while both the peaks and background counts were kept reasonably large. Correspondingly the total number of counts is much larger than in the previous cases. However, this is not a "strong signal" scenario yet, because the average number of counts per bin is still relatively small, $N/B \simeq 6.8$.

DATA SET 3 simulated from Poisson$(D \mid \beta_1 = 6., \beta_2 = 9., \beta_3 = 0., \beta_c = 5.)$ N=6838, B=1000				
estimates	true values	$\mathcal{H}_1$	$\mathcal{H}_3$	$2 \log \mathcal{B}$
In $\mathcal{H}_3$, r prior bound $r_{\max} = 1.$				
$\hat{\beta}_1$:	6.	5.97±0.36	6.09±1.16	
$\hat{\beta}_2$:	9.	8.73±0.39	8.88±1.37	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	5.	5.00±0.09	5.02±0.23	
$\hat{r}$:	—	—	1.00±0.00	
$\log \hat{Z}$:	—	-2294.970±0.034	-2957.728±0.037	$> 10, \mathcal{H}_1$
In $\mathcal{H}_3$, r prior bound $r_{\max} = 1000.$				
$\hat{\beta}_1$:	6.	5.97±0.36	5.97±0.36	
$\hat{\beta}_2$:	9.	8.73±0.39	8.72±0.40	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	5.	5.00±0.09	5.00±0.09	
$\hat{r}$:	—	—	667.59±217.49	
$\log \hat{Z}$:	—	-2294.970±0.034	-2295.946±0.035	< 2

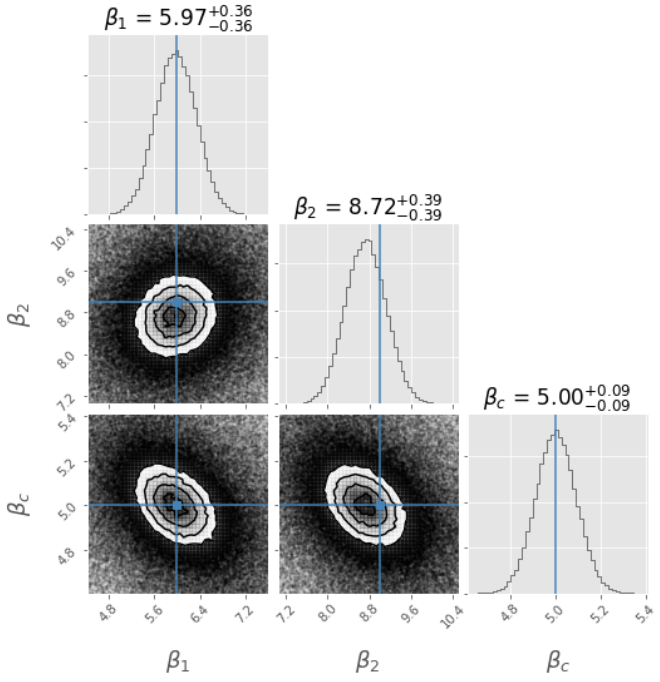


Figure 5.3: Posterior contour plots of model $\mathcal{H}_1$ parameter pairs; data set 3. Blue lines indicate true values. All the sample distributions converged.

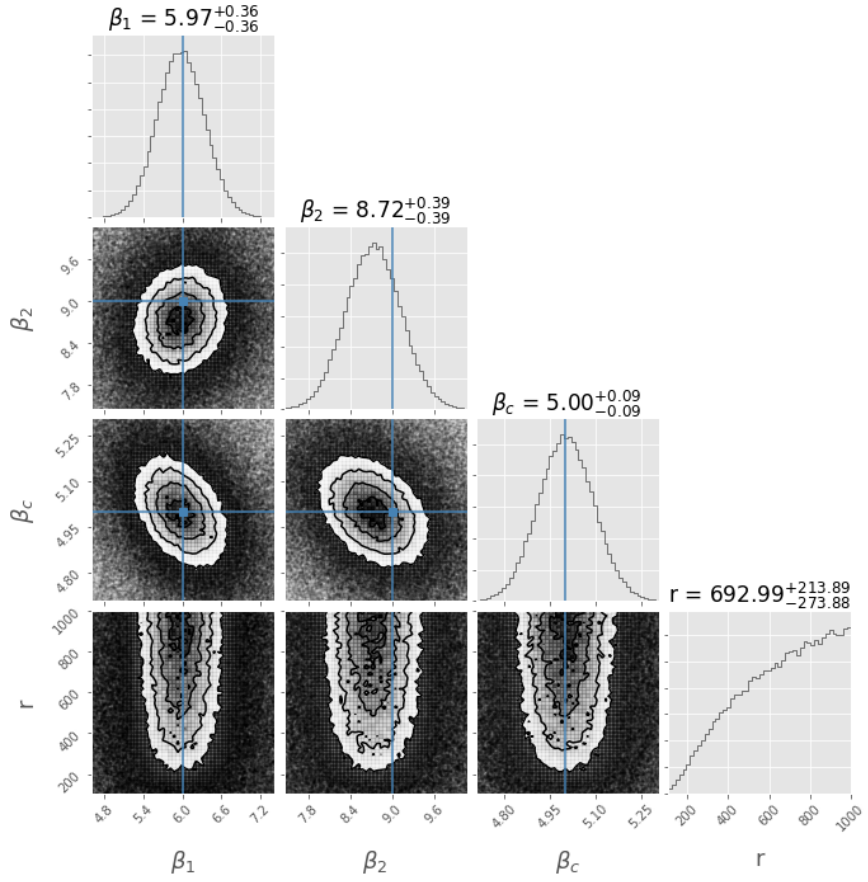


Figure 5.4: Posterior contour plot of model $\mathcal{H}_3$ parameter pairs. The posterior of r has a large variance.

Nevertheless, the larger total N means that meaningful contour plots can be made which give more information on the parameter posterior, its marginals and two-dimensional cross sections. In Figures 5.3 and 5.4, contour plots of $\mathcal{H}_1$ and $\mathcal{H}_3$ parameter posteriors are shown along with the true parameter values indicated by blue lines. Also shown are the projections (marginal distributions) of each parameter.

- The best-fit parameters have smaller variances than in the very-low count Problem 1 and Problem 2 cases, as they should. All MAP values are compatible with the true ones. Also the resulting estimated parameters values are less dependent on r prior as more data are collected.

- However, $\hat{\beta}_1$ and $\hat{\beta}_2$ are underestimated by both models while β_c is estimated reasonably accurately. Nontrivial parameter-parameter correlations are also visible in the form of tilted ellipses in the various cross sections (peak parameters β_1 and β_2 are both negative correlated with background parameter β_c). Frequentist confidence interval plots are clearly proxies for posterior contour plots (the contour outlines are sigma levels).

Test problem 4

In this test, we consider data generated from Negative Binomial distributions in each bin. For small values of its parameter r, the Negative Binomial has large fluctuations and "fat tails". Figure 5.5 shows a toy example where a number of Poisson posterior predictions are shown for $\lambda = 5$. The red dots represent the mean of the prediction and the red bars the standard deviations of the Poisson data. By contrast, the blue Negative Binomial points are clearly far more dispersed than their Poisson equivalents.

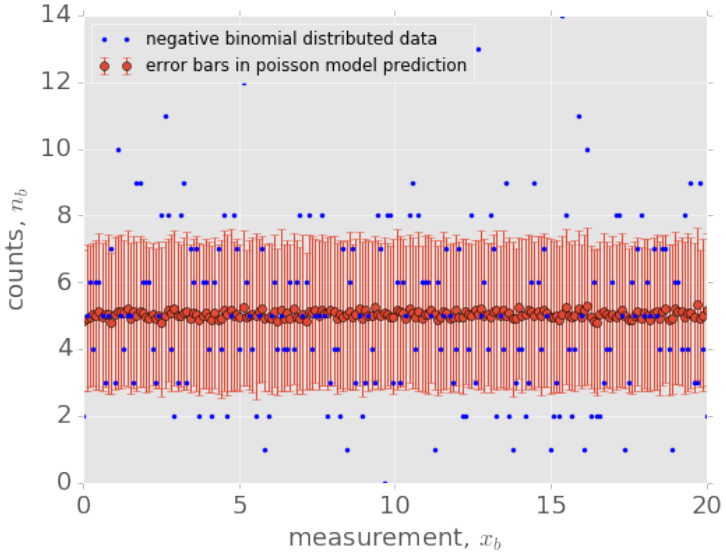


Figure 5.5: Red error bars come from poisson model prediction, while blue dots are simulated Negative Binomial data NB($\beta = 5, r = 50$). We can see that the data structure is not fully captured by a Poisson model.

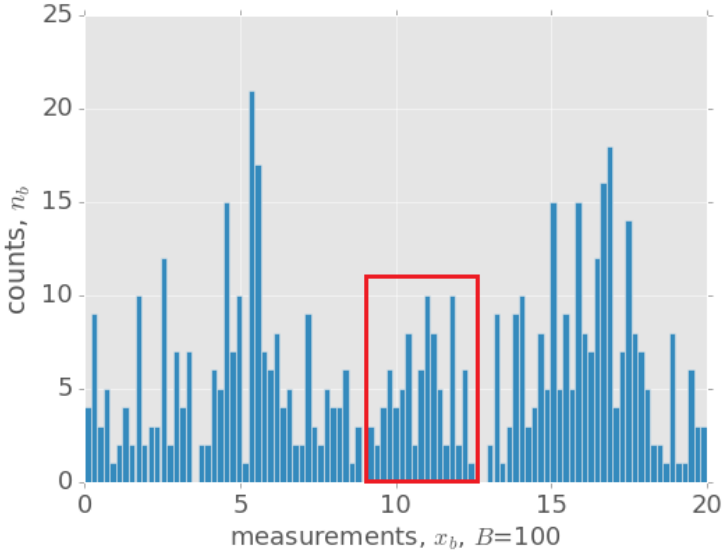


Figure 5.6: Example of a spectrum generated from NB distributions.

Given "real" Negative Binomial data, the chance of peak misidentification is larger than in the Poisson-generated case. In Figure 5.6, we show a typical spectrum which has, of course, the obvious peaks at $x = 5$ and 16 but also spurious peaks at various other locations.

The fourth data set was generated from NB binwise distributions with function parameters $\beta_1 = 4, \beta_2 = 10, \beta_3 = 0, \beta_c = 4$ and $r = 4$. In this test problem, we will include a hypothetical third peak at $x_b = 11$ (in the red rectangle) in order to see whether the scheme identifies it as spurious or real. The spurious-peak (three-peak) scenarios are captured in hypothesis $\mathcal{H}_2$ (Poisson likelihood) and $\mathcal{H}_4$ (NB likelihood) respectively.

DATA SET 4 simulated from NB$(D \mid \beta_1 = 4., \beta_2 = 10., \beta_3 = 0., \beta_c = 4., r = 4.)$ N=559, B=100						
esti-mati-ons	true val-ues	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$2 \log \mathcal{B}$
In $\mathcal{H}_3$ and $\mathcal{H}_4$, $r_{\max} = 50$.						
$\hat{\beta}_1$:	4.	5.44±1.01	6.08±1.04	5.46±1.71	6.13±1.73	
$\hat{\beta}_2$:	10.	8.12±1.16	8.85±1.19	8.70±2.24	9.54±2.27	
$\hat{\beta}_3$:	0.	—	2.21±0.84	—	2.33±1.19	
$\hat{\beta}_c$:	4.	3.94±0.25	3.50±0.28	3.97±0.36	3.53±0.38	
$\hat{r}$:	4.	—	—	4.58±1.38	4.85±1.52	
$\log \hat{Z}$:	—	-287.687±0.017	-287.145±0.020	-266.990±0.018	-267.939±0.020	$< 2, \mathcal{H}_3$

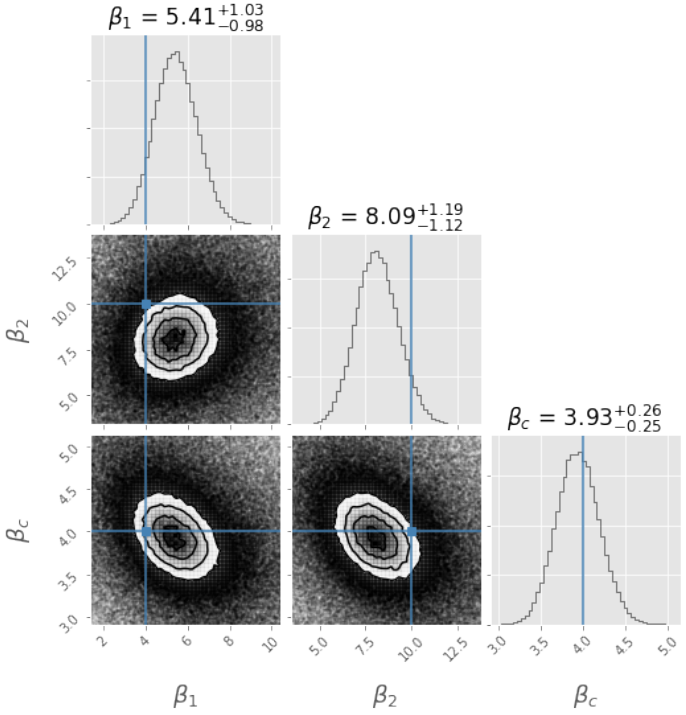


Figure 5.7: Posterior contour plot of hypothesis $\mathcal{H}_1$ parameters. Data set 4.

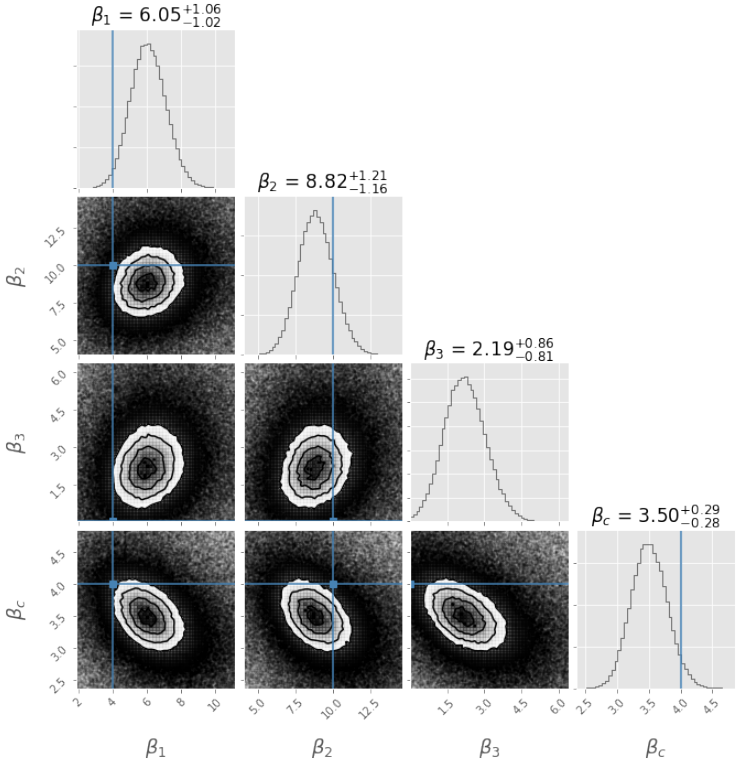


Figure 5.8: Posterior contour plot of hypothesis $\mathcal{H}_2$ parameters. Data set 4.

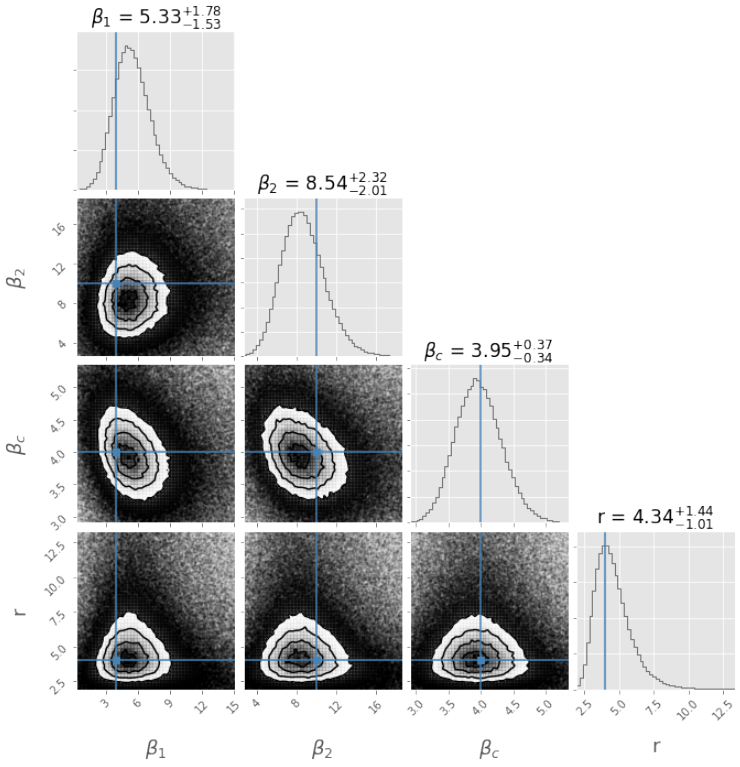


Figure 5.9: Posterior contour plot of hypothesis $\mathcal{H}_3$ parameters. Data set 4.

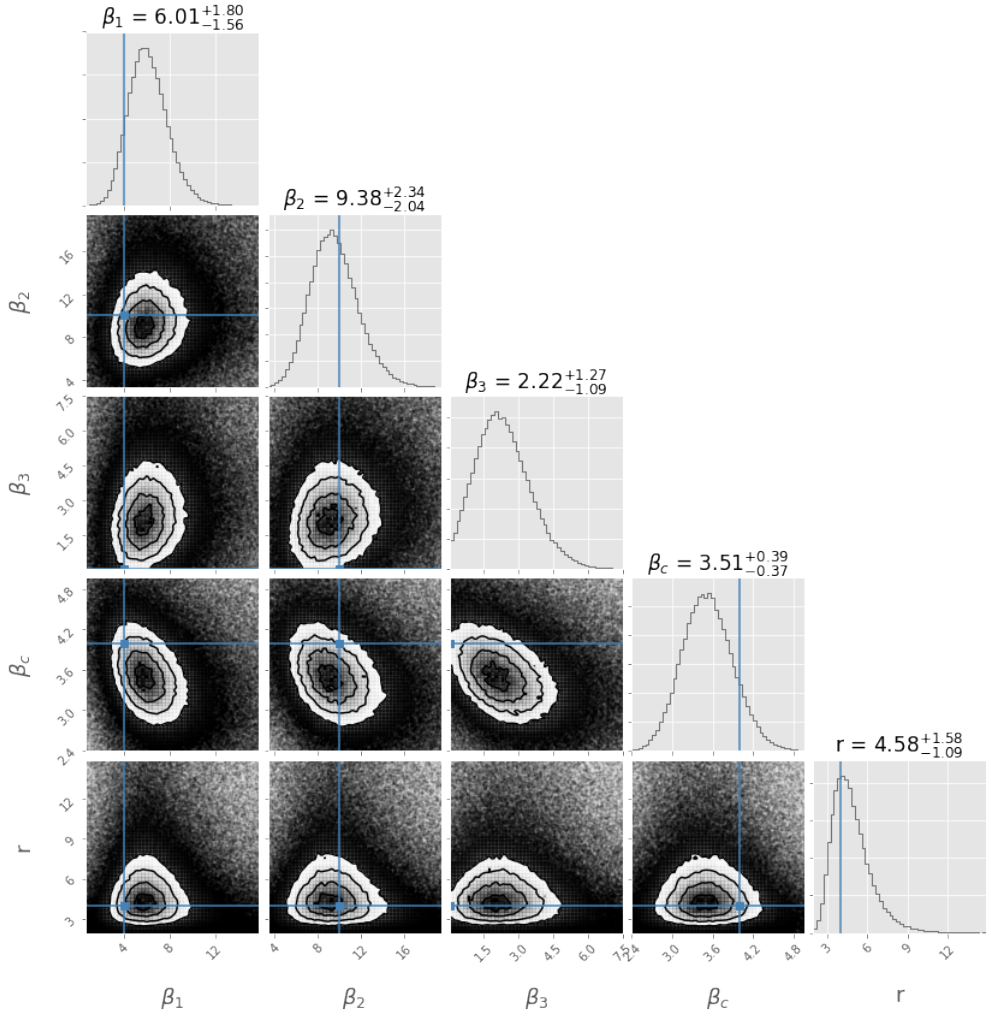


Figure 5.10: Posterior contour plot of hypothesis $\mathcal{H}_4$ parameters. Data set 4.

- From the four evidences, we firstly learn that the two Poisson model scenarios (2-peak or 3-peak) are both disfavoured compared to their NB equivalents. The evidence for the 2-peak NB is slightly larger (less negative) than the 3-peak one (remember that these are log scales). A look at the evidence value for the spurious peak's model $\mathcal{H}_4$ compared to $\mathcal{H}_3$ equivalent also favours the conclusion that the posterior correctly identifies the peak at 11 as spurious. This comes, however, at the price of larger variances for the other two real peak amplitudes.

- A quick look at Data Set 5 below with the same parameters but a larger total count merely confirms that the Negative Binomial model is strongly favoured compared to the Poisson competition, as it should be. Also the best-fit parameter values are more realistic for the NB case.

DATA SET 5 simulated from NB($D \mid \beta_1 = 4., \beta_2 = 10., \beta_3 = 0., \beta_c = 4., r = 4.$) N=5714, B=1000				
estimates	true values	$\mathcal{H}_1$	$\mathcal{H}_3$	$2 \log \mathcal{B}$
In $\mathcal{H}_3$, $r_{\max} = 50$.				
$\hat{\beta}_1$:	4.	3.66±0.31	3.72±0.49	
$\hat{\beta}_2$:	10.	9.19±0.38	9.18±0.68	
$\hat{\beta}_3$:	0.	—	—	
$\hat{\beta}_c$:	4.	4.11±0.08	4.11±0.12	
$\hat{r}$:	4.	—	5.12±0.75	
$\log \hat{Z}$:	—	-2789.601±0.035	-2573.054±0.022	$> 10, \mathcal{H}_3$

Conclusions

The toy models show by example the Bayesian way to estimate parameters and compare competing models in discrete spectra. We have learnt by example that evidence is directly related to the prior settings and that care must therefore be taken to ensure that the prior either reflects actual information or else is as unbiased as possible. In the absence of information, we must navigate between two competing needs. On the one hand, a wider prior prevents unjustified bias and allows the system to evolve to larger evidences. However, a wider prior comes at the cost of much larger parameter space volumes which must be searched by the Monte Carlo algorithms, requiring both much longer run times and more cases where the MC algorithm fails to converge to a stable result.

Using a Poisson model to fit Negative Binomial data is clearly very bad, while using a Negative Binomial model to fit Poisson data is fine, except of course that in the above simulated test examples we know the parameter r to be redundant. For the purposes of inference, however, the evidence does whatever it does; it either favours the simpler Poisson model by means of the Occam's Razor automatically built into the Bayesian method, or the data is inconclusive regarding the competing models.

We have seen that the posterior calculations automatically eliminated the small r prior in NB model (the extra complexity does not exist) for Poisson-distributed data even for small count numbers. This is satisfying but does not answer the question of model comparison when the number of parameters in different model hypotheses is not the same. That remains an active research topic which is beyond the scope of this thesis.

Appendix A

Raftery and Lewis

In the context of Bayesian confidence intervals, suppose we want to estimate a quantile q for posterior parameter θ, from the quantile Eq.(2.15), we have $\hat{q} = p(\theta \leq \theta_q \,|\, D)$, where θ_q is the quantile value. The mathematical expression for the estimated $\hat{q}$ within an accuracy of r with probability s is

$$p(|\hat{q} - q| \leq r) = s$$

To achieve this goal, computing θ_t for each iteration t, the value of Z_t is given by

$$Z_t = \begin{cases} 1, & \text{if } \theta_t \leq \theta_q \\ 0, & \text{otherwise} \end{cases}$$

Z_t is a binary process born from a Markov chain θ_t, it is not necessarily a Markov chain, but the new process $Z_t^{(k)}$, where $Z_t^{(k)} = Z_{1+(t-1)k}$, consisting of only those of the k^{th} simulation from the original chain, can be regarded as a two-state Markov chain if k is sufficiently large.

$$A = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

where α, β are transition probabilities.
The eigenvectors of matrix A is obtained through stationary equation $\boldsymbol{\pi} = \boldsymbol{\pi}A$, $\boldsymbol{\pi} = [\pi(0), \pi(1)] = \frac{1}{\alpha+\beta}[\beta, \alpha]$. Suppose after m iterations, this two-state chain has stabilised, in other words, we have [34].

$$A^m = \begin{bmatrix} \pi(0) & \pi(1) \\ \pi(0) & \pi(1) \end{bmatrix} + \frac{\lambda^m}{\alpha+\beta} \begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}$$

Alternatively, one can write A^m as

$$A^m = \begin{bmatrix} \pi(0) + \lambda^m \pi(1) & \pi(1) - \lambda^m \pi(1) \\ \pi(0) - \lambda^m \pi(0) & \pi(1) + \lambda^m \pi(0) \end{bmatrix}$$

where $\lambda = (1-\alpha-\beta), \quad (0 < \lambda < 1)$. From the above equation, we can also get the two-state distribution of Z_t, it follows from a binary distribution.

Therefore the requirement becomes

$$|\, A_{ij}^m - \pi(j) \,| \leq \epsilon \quad \text{for } i, j = 0, 1$$

by simplifying it,

$$\lambda^m \leq \frac{(\alpha+\beta)\epsilon}{\max(\alpha,\beta)}$$

this equation holds if $m = \log\left\{\frac{(\alpha+\beta)\epsilon}{\max(\alpha,\beta)}\right\} / \log\lambda$, thus we obtain the number iterations M to be abandoned for Markov chain θ

$$M = mk$$

Next we turn to estimate the total number of iterations N.

The mean and the variance of $\bar{Z_t}^{(k)}$ is given by

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\begin{aligned}\mathbf{Var}(\bar{Z_t}^{(k)}) &= \frac{1}{n}\mathbf{Var}(Z_t^{(k)}) + \frac{2}{n^2}\sum_{\Delta_t=1}^{n-1}(n-\Delta_t)\mathbf{Cov}\big(Z_t^{(k)}, Z_{t+\Delta_t}^{(k)}\big) \\ &= \frac{1}{n}\Big\{\mathbf{Var}(Z_t^{(k)}) + 2\sum_{\Delta_t=1}^{\infty}\mathbf{Cov}\big(Z_t^{(k)}, Z_{t+\Delta_t}^{(k)}\big)\Big\} \qquad \text{(A.1)}\end{aligned}$$

we compute the covariance for this two-state Markov process first [45],

$$\begin{aligned}\mathbf{Cov}\big(Z_t^{(k)}, Z_{t+\Delta_t}^{(k)}\big) &= \mathbf{E}[(Z_t^{(k)}=1)\cdot(Z_{t+\Delta_t}^{(k)}=1)] - \mathbf{E}[(Z_t^{(k)}=1)]\mathbf{E}[(Z_{t+\Delta_t}^{(k)}=1)] \\ &= p(Z_{t+\Delta_t}^{(k)}=1, Z_t^{(k)}=1) - \pi^2(1) \\ &= p(Z_t^{(k)}=1)p(Z_{t+\Delta_t}^{(k)}=1 \,|\, Z_t^{(k)}=1) - \pi^2(1) \\ &= \pi(1)A_{11}^{\Delta_t} - \pi^2(1) \\ &= \pi(1)[\pi(1) + \lambda^{\Delta_t}\pi(0)] - \pi^2(1) \\ &= \pi(0)\pi(1)\lambda^{\Delta_t} \qquad \text{(A.2)}\end{aligned}$$

Since it is easy to have $\mathbf{Var}(Z_t^{(k)}) = \pi(0)\pi(1)$, then by plugging Eq.$(A.2)$ into Eq.$(A.1)$, and eventually, we obtain the variance of $\mathbf{Var}(\bar{Z_t}^{(k)})$

$$\begin{aligned}\mathbf{Var}(\bar{Z_t}^{(k)}) &= \frac{1}{n}\Big\{\pi(0)\pi(1) + 2\sum_{\Delta_t=1}^{\infty}\pi(0)\pi(1)\lambda^{\Delta_t}\Big\} \\ &= \frac{\pi(0)\pi(1)}{n}\Big[1 + 2\sum_{\Delta_t=1}^{\infty}\lambda^{\Delta_t}\Big] \\ &= \frac{\pi(0)\pi(1)}{n}(1 + \frac{2\lambda}{1-\lambda}) \\ &= \frac{\pi(0)\pi(1)}{n}\frac{1+\lambda}{1-\lambda} \\ &= \frac{1}{n}\frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3}\end{aligned}$$

In the limit of $n \to \infty$, the binomial distribution of Z_t^k is approximately a Gaussian distribution. Note that $\hat{q} = \bar{Z_t}^{(k)}$, the requirement for $p\big(|\bar{Z_t}^{(k)} - q| \leq r\big) = s$ is accomplished by

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{q-r}^{q+r} \mathrm{d}\bar{Z_t}^{(k)} \exp\Big\{ - \frac{(\bar{Z_t}^{(k)} - q)^2}{2\sigma^2} \Big\} = \mathrm{erf}\Big(\frac{r}{\sqrt{2}\sigma}\Big) = s$$

Since the error function is connected with the cumulative distribution ϕ, which is the integral of standard normal distribution, in particular, $\phi(x) = \frac{1}{2}[1 + \mathrm{erf}(\frac{x}{\sqrt{2}})]$, now we can have the following equation,

$$\phi\Big(\frac{r}{\sigma}\Big) = \frac{1}{2}(s+1)$$

which is the same as $\frac{r}{\sqrt{\mathbf{Var}(\bar{Z_t}^{(k)})}} = \phi^{-1}\big(\frac{1}{2}(s+1)\big)$, thus we obtain the sample size of $\bar{Z_t}^{(k)}$,

$$n = \frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3} \Big\{ \frac{\phi^{-1}\big(\frac{1}{2}(s+1)\big)}{r} \Big\}^2$$

It implies that the dense[1] chain has sample size,

$$N = nk$$

In practice, information like the minimum number of iteration to run is quite useful, then we compute $N_{\min}$ by supposing successive samples are independent, in other words $M = 0, k = 1$

$$N_{\min} = \pi(0)\pi(1) \Big\{ \frac{\phi^{-1}\big(\frac{1}{2}(s+1)\big)}{r} \Big\}^2$$

[1]The chained has not been thinned

Appendix B

Order statistics

Denote a sample of size N, with $\{x_i\}_{i=1}^{N}$, independently and identically distributed with probability denstity function $p(X)$ and cumulative density function $U = F(x)$. After sorting the sample in increasing order, let $P(x_n)$ be the probability that obtaining the n-th element $x_n, n \in N$, from such a sample is given by joint probability of three propositions [46]

1. $n-1$ number of elements smaller or equal to X,
2. one element within the interval $x_n \in (X, X + \mathrm{d}X]$,
3. $N - n$ items are greater $X + \mathrm{d}X$.

Since the number of ways for such an observation is given by

$$\frac{N!}{(n-1)!1!(N-n)!} = \frac{1}{B(n, N-n+1)}, \tag{B.1}$$

and each proposition has probability

$$F^{n-1}(X)[F(X+\mathrm{d}X) - F(X)][1 - F(X+\mathrm{d}X)]^{N-n}, \tag{B.2}$$

the desired probability that getting x_n from the interval $(X, X + \mathrm{d}X]$ is the product of elements(B.1,B.2)

$$\begin{aligned} P(X < x_n \leq X + \mathrm{d}X) &= \frac{1}{B(n, N-n+1)} F^{n-1}(X)[F(X+\mathrm{d}X) - F(X)][1 - F(X+\mathrm{d}X)]^{N-n} \\ &= \frac{F^{n-1}(X)[1 - F(X+\mathrm{d}X)]^{N-n}}{B(n, N-n+1)} p(X)dX. \end{aligned} \tag{B.3}$$

By setting the limit $\mathrm{d}X \approx 0$, $P(x_n)$ is given by

$$\begin{aligned} P(x_n) &\approx \frac{P(X < x_n \leq X + \mathrm{d}X)}{dX} \approx \frac{F^{n-1}(x_n)[1 - F(x_n)]^{N-n}}{B(n, N-n+1)} p(x_n) \\ &= Beta(n, N-n+1)p(x_n) \end{aligned} \tag{B.4}$$

Using Jacobian transformation, the probability that obtaining U_n from an ordered sample of size N is

$$P(U_n) = P(x_n)\left|\frac{\mathrm{d}X}{\mathrm{d}U}\right| = Beta(n, N-n+1) \tag{B.5}$$

In the special case that $n = N$, we obtain the probability of drawing the maximum value from interval [0,1]

$$P(U_N) = N(U_N)^{N-1} \tag{B.6}$$

where $U_N = F(x_N)$.

Bibliography

[1] Wolfgang von der Linden, Volker Dose and Udo von Toussaint. *Bayesian Probability theory Applications in the Physical Sciences*. Cambridge University Press, 2014.

[2] Eric D. Feigelson, G.Jogesh Babu. *Statistical Challenges in Modern Astronomy*. Springer Verlag, 1993.

[3] Hans C. Eggers. *Physics757: Entropy and Information Lecture Notes*. 2016.

[4] Phil Gregory. *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press, 2005.

[5] Allen Caldwell. *Model-Based Data Analysis Parameter Inference and Model Testing Lecture Note*. 2016.

[6] Eric Weisstein. Gamma Distribution – from Wolfram MathWorld.

[7] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, Fourth Edition, 2010.

[8] William Feller. *An Introduction to Probability Theory and Its Application*, John Wiley & Sons, Inc. Third Edition 1967.

[9] Hans C. Eggers. *Physics344: Monte Carlo Methods in Physics Theory Lecture Notes*. 2016.

[10] Christian P. Robert, George Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

[11] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press. 2005.

[12] E.T. Jaynes. *Probability Theory The Logic of Science*. Cambridge University Press. 2003.

[13] Cameron Davidson Pilon. *Bayesian methods for hackers*. Addison-Wesley 2016.

[14] Jose M.Bernardo. Adrian F.M.Smith. *Bayesian Theory*. 1994.

[15] E.T.Jaynes. *Papers on Probability, Statistics and Statistical Physics*. N0. 1963. Kluwer Academic, 1989.

[16] G.Larry Bretthorst. *Maximum Entropy and Bayesian Methods Fundamental Theories of Physics*. Springer Science Business Media,B.V, 1993.

[17] Harold Jeffreys. *The Theory of Probability.* Oxford University Press, Third edition, 1961.

[18] Robert E. Kass and Adrian E. Raftery *Bayes Factor.* Journal of the American Association.90(430) 1995.

[19] Joseph M. Hilbe. *Modelling Count Data.* Cambridge Univ Press, 2014.

[20] John Hinde. *Overdispersion: Models and Estimation A Short Course for SINAPE 1998.* Statistics, 2007.

[21] Sheldon M. Ross. *Introduction to Probability Models.* Academic Press, Sixth edition, 1997.

[22] Kevin P. Murphy. *Machine Learning A Probabilistic Perspective*, The MIT Press 2012.

[23] Unknown. graphics - Drawing Graph of Markov Chain with 'Patches' using Tikz - TeX - LaTeX Stack Exchange.

[24] John K. Kruschke. *Doing Bayesian Data Analysis.* Academic Press, First edition, 2010.

[25] Antti Solonen. *Monte Carlo Methods in Parameter Estimation of Nonlinear Models.* Lappeenranta University of Technology, Master's Thesis, 2006.

[26] Gareth O. Roberts and Jeffrey S. Rosenthal. *Optimal scaling for various Metropolis Hastings algorithms*, 2001.

[27] Christophe Andrieu and J Thoms. *A tutorial on adaptive MCMC.* Statistics and Computing, 18(4):343–373, 2008.

[28] CJ. Geyer. *Markov chain Monte Carlo lecture notes.* Course notes, Spring Quarter, 1998.

[29] M.E.J. Newman and G.T. Barkema. *Monte Carlo Methods in Statistical Physics.* Oxford University Press, 2001.

[30] S Sinharay. *Assessing convergence of the Markov Chain Monte Carlo algorithms: A review.* ETS Research Report Series, March, 2003.

[31] Mary Kathryn Cowles and Bradley P. Carlin. *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review.* 2014.

[32] John Geweke. *Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments.* Bayesian Statistics 4, pages 169–193, 1992.

[33] Chris Fonnesbeck, Anand Patil, David Huard, John Salvatier. *https:// pymc-devs.github.io/pymc/index.html.*

[34] A.E. Raftery and S.M. Lewis. *The number of iterations, convergence diagnostics and generic Metropolis algorithms.* Practical Markov Chain Monte Carlo, 7(98):763–773, 1995.

[35] Andrew Gelman and Donald B. Rubin. *A Single Series from the Gibbs Sampler Provides a False Sense of Security.* Bayesian Statistics, 4(July):625–631, 1992.

[36] A.P. Dempster, N.M. Laird and Donald B Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society Series B Methodological, 39(1):1–38, 1977.

[37] A. Gelman and D.B. Rubin. *Inference from iterative simulation using multiple sequences.* Statistical science, 7(4):457–472, 1992.

[38] Stephen P. Brooks and Andrew Gelman. *General methods for monitoring convergence of iterative simulations.* Journal of computational and graphical statistics, 7(4):434–455, 1998.

[39] Martin D. Weinberg *Computing the Bayes Factor from a Markov chain Monte Carlo Simulation of the Posterior Distribution* . Bayesian Anal. Volume 7, Number 3, 737-770. 2010.

[40] John Skilling. *Nested sampling for general Bayesian computation.* Bayesian Analysis, 1(4):833–860, 2006.

[41] F. Feroz, M.P. Hobson and M. Bridges. *MultiNest: An efficient and robust Bayesian inference tool for cosmology and particle physics.* Monthly Notices of the Royal Astronomical Society, 398(4):1601–1614, 2009.

[42] J. R. Shaw, M. Bridges and M. P. Hobson. *Efficient Bayesian inference for multimodal problems in cosmology.* April, 2017.

[43] *GitHub - kbarbary_ nestle Pure Python, MIT-licensed implementation of nested sampling algorithms for evaluating Bayesian evidence.*

[44] Foreman-Mackey Daniel and Contributors. *GitHub corner.* The Journal of Open Source Software, 2016.

[45] M.N. Islam and C.D. O'shaughnessy. *On the Markov Chain Binomial Model.* Applied Mathematics, 4(December):1726–1730, 2013.

[46] H.A. David and H.N. Nagaraja. *Order Statistics*, John Wiley & Sons, Inc. Third Edition. 2003.