University of Applied Sciences Munich
Department of Computer Science and Mathematics

Master Thesis

STOCHASTIC PROFILING OF SINGLE-CELL HETEROGENEITIES

A BAYESIAN EXTENTION

by Christoph Franz Kurz



Advisors: Prof. Dr. Martin Leitner
Dr. Christiane Fuchs
February 2015

Christoph Franz Kurz: *Stochastic profiling of single-cell heterogeneities*, a Bayesian extention, © February 2015

CONTENTS

1 INTRODUCTION 1

I BASICS 3

2 BAYESIAN INFERENCE 4
- 2.1 Bayes' Theorem . . . 5
- 2.2 The Bayesian Inference Model . . . 5
- 2.3 Prior Distributions . . . 7
  - 2.3.1 Informative Priors . . . 8
  - 2.3.2 Uninformative Priors . . . 8
  - 2.3.3 Weakly Informative Priors . . . 8
  - 2.3.4 Least Informative Priors . . . 9
  - 2.3.5 Proper and Improper Priors . . . 10
  - 2.3.6 Conjugate Priors . . . 11
- 2.4 Likelihood . . . 11
- 2.5 Bayesian Computation . . . 12
- 2.6 Markov Chain Monte Carlo Methods . . . 13
- 2.7 Model Validation . . . 15

3 THE LOGNORMAL DISTRIBUTION 17

4 MIXTURE MODELS 20
- 4.1 Finite Mixture Models . . . 20
- 4.2 Mixture Likelihoods and Posteriors . . . 20
- 4.3 MCMC for Mixtures . . . 22
- 4.4 The Label Switching Problem . . . 24

II THE SHOWCASE 26

5 MCMC FOR STOCHASTIC PROFILING 27
- 5.1 Methods . . . 27
- 5.2 Results . . . 31
  - 5.2.1 Required Sample Sizes . . . 31
  - 5.2.2 Estimating n . . . 32
  - 5.2.3 Finding the Best n . . . 33
  - 5.2.4 Fixed n . . . 36
  - 5.2.5 Estimation of Real Data . . . 37
- 5.3 Bivariate Lognormal Estimation . . . 39
- 5.4 Multivariate Lognormal Estimation . . . 48

6 CONCLUSION 50

A APPENDIX 51

BIBLIOGRAPHY 65

LIST OF FIGURES

Figure 2.1 A comparison of the Normal Distribution and the Uniform Distribution 9

Figure 2.2 Four independent Markov chains sampled with the Metropolis Algorithm. The target distribution is a binormal distribution, whereas the proposal distribution is a normal distribution centered at the previous sample. Starting points of each chain are marked by a solid diamond symbol. 14

Figure 4.1 A variety of distribution shapes from lognormal mixtures 21

Figure 5.1 The four parameter sets we used as gene expression data. All are mixture distributions of two lognormal distributions: $f(x) = p \cdot \mathcal{LN}(\mu_1, \sigma) + (1-p) \cdot \mathcal{LN}(\mu_2, \sigma)$ 29

Figure 5.2 Correlation matrices for samples from the GG and HARM samplers 30

Figure 5.3 A comparison of the GG and HARM algorithms on the same dataset. 31

Figure 5.4 MCMC estimation of the parameter n using the GG sampler. 33

Figure 5.5 MSE for the parameter set $p = 0.8$, $\mu_1 = 1$, $\mu_2 = 6$, and $\sigma = 0.3$ for both MCMC (*blue*) and MLE (*green*). 34

Figure 5.6 MSE for the parameter set $p = 0.225$, $\mu_1 = 0.5$, $\mu_2 = -2$, and $\sigma = 0.225$ for both MCMC (*blue*) and MLE (*green*). 34

Figure 5.7 MSE for the parameter set $p = 0.55$, $\mu_1 = 0.5$, $\mu_2 = -2$, and $\sigma = 0.3$ for both MCMC (*blue*) and MLE (*green*). 35

Figure 5.8 MSE for the parameter set $p = 0.2$, $\mu_1 = 0.5$, $\mu_2 = -2$, and $\sigma = 0.8$ for both MCMC (*blue*) and MLE (*green*). 35

Figure 5.9 Boxplots of MCMC and maximum likelihood of the parameter p for $n = 1, ..., 30$ of 100 runs each. 36

Figure 5.10 Intervals for MCMC estimation based on data with $n = 10$ but performed with different n. . 37

Figure 5.11 Line- and density plots of MCMC estimation for all parameters using the GG sampler for all parameters p, μ_{11}, μ_{21}, μ_{12}, μ_{22}, σ, and ρ. The orange line represents the MCMC estimate, the purple line the estimate of the MLE. The black dashed line marks the true value. It can be observerd that, for a small correlation between two genes ($\rho = 0.2$ here), MLE yields similar results to MCMC even though independence ($\rho = 0$) is assumed in MLE 42
Figure 5.12 Line- and density plots of MCMC estimation using the GG sampler for all parameters p, μ_{11}, μ_{21}, μ_{12}, μ_{22}, σ, and ρ. The orange line represents the MCMC estimate, the purple line the estimate of the MLE. The black dashed line marks the true value. It can observerd that, for a high correlation between two genes ($\rho = 0.8$ here), MCMC estimation, where ρ is estimated, yields better results than MLE where independence ($\rho = 0$) is assumed. 44
Figure 5.13 Correlation matrix heatmap for all 23 genes in a real data set, data set 1. 45
Figure 5.14 Lineplot of a multivariate estimation ($m = 23$) of parameter p of data set 1. 49
Figure A.1 Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 5$ 52
Figure A.2 Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 10$ 54
Figure A.3 Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 20$. 56
Figure A.4 Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 30$. 58
Figure A.5 Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 50$. 60

Figure A.6 Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 100$. 62

LIST OF TABLES

Table 5.1 The minimum sample size required for each n to achieve reasonable results for all four parameters p, μ_1, μ_2, σ (i.e., the true value being within the 95% interval). 32
Table 5.2 Comparison of the estimates for each parameter of MCMC and MLE on a real data set, the SOD2 data. 38
Table 5.3 Comparison of the estimates of MCMC and MLE on the two highest correlated genes in a real data set, data set 1. 46
Table 5.4 Comparison of the estimates of MCMC and MLE on the two highest correlated genes in a real data set, data set 2. 46
Table 5.5 Comparison of the estimates of MCMC and MLE on the two highest correlated genes in a real data set, data set 3. 47
Table 5.6 Required sample sizes and number of paramters not converging when estimating a mixture of m-dimensional multivariate lognormals. 49

LISTINGS

Listing 1 MCMC model definition for the univariate case 63
Listing 2 MCMC model definition for the bivariate case 64

1 INTRODUCTION

The human body contains a large number of different cells, with different functions. Nerve cells are vastly different from lung cells regarding their function, yet both contain the same DNA. The difference occurs because the genes differ in their expression: in different kinds of tissues, different genes are active. This is especially true for cancer cells. Such cells vary in their expression, yet they share the same phenotype, i.e., they look the same. So to identify cancerous cells, you would have to perform micro-array analysis to quantify the relative amount of messenger RNA in each cell. This is a very cost- and time-intense approach and measurement errors multiply. Janes et al. [2010] developed a method called *Stochastic Profiling*. By this method, a tissue sample of 10 cells is taken, and gene expression of the whole tissue is measured by micro-arrays. From this sample of several cells, the individual gene expression of the single cells is inferred by statistical methods and therefore less prone to measurement errors than single-cell measurements. Bajikar et al. [2014] improved this method using Maximum Likelihood Estimation (MLE) and showed that a 10-cell sample yields better results than measuring each cell individually, when the sample size is small. We will show that a Bayesian approach through Markov chain Monte Carlo (MCMC) methods works well for stochastic profiling and sometimes even outperforms the traditional maximum likelihood approach. Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations [Gelman et al., 2014].

There has been constant growth in the development and application of Bayesian inference in statistics. One reason for this is the availability of computational algorithms to compute the range of integrals that are necessary in a Bayesian posterior analysis. Owing to the speed of modern computers, it is now possible to use the Bayesian paradigm to fit very complex models that cannot be fit by alternative frequentist models [Albert, 2009].

In the next part of this thesis we will give a short introduction to Bayesian inference and MCMC algorithms. We will then formally introduce the lognormal distribution as the distribution that gene expression is usually modeled by. The final chapter of the first part will be about mixture models which play an important role for us.

In the following part, we will present the results of our MCMC approach on stochastic profiling. We will answer the question of how

large the sample size should be for various numbers of cells n, how well this n can be estimated, and try to find the best n for estimation. We fix n and perform estimation with various assumptions for it. Then, we estimate real data and extend our model to the bivariate case, where correlation is also assessed. Finally, we generalize to the multivariate case and show what problems we encountered there.

All computations were done using the `R` programming language, especially the packages `LaplacesDemon` for MCMC and `stochprofML` for MLE. Additional computations and all visualizations were performed in `mathematica 10`.

Part I

BASICS

2 BAYESIAN INFERENCE

Outline
In this chapter we introduce the Bayesian inference concepts this work is built upon. First, we briefly describe the difference of Bayesian inference and frequentist inference. Then, we discuss Bayes' theorem and the concept of prior distributions. Finally, we turn to likelihood and give a brief introduction to computational methods and Markov chain Monte Carlo algorithms. We conclude with a section about model validation.

One of the first things an undergraduate student hears about statistics is that there are two different approaches: Bayesian inference and frequentist inference, often called frequentism and Bayesianism.

Fundamentally, Bayesian and frequentist approaches concern the difference in the nature of probability. Frequentists define probability as the limit of an event's relative frequency in a large number of trials. In the limit of a large number of measurements, the *frequency* of a given value indicates the probability of measuring that value, so probabilities are related to frequencies of events. Mathematically, the probability of an event x can be expressed as the limit of the quotient of the frequency of this event n_x and the number of all observed events n as n goes to infinity:

$$P(x) = \lim_{n\to\infty} \frac{n_x}{n}.$$

Bayesian inference, on the other hand, extends the concept of probability to cover degrees of certainty about statements: it is able to assign probabilities to any statement, even when a random process is not involved. Within Bayesian inference, there are also different interpretations of probability, and even different approaches based on these interpretations. On one side, we have the *objective* Bayesian inference [Berger, 2006], on the other, the *subjective* Bayesian inference [Anscombe and Aumann, 1963; Goldstein, 2006]. Objective Bayesian inference, or the traditional Bayesian inference, is usually associated with Bayes and Price [1763], Laplace [1814], and Jeffreys [1961], who put Bayes' theorem and Laplace's formulation on an axiomatic basis. Subjective Bayesianism is based on Ramsey [1926], De Finetti [1931], and Savage [1954]. De Finetti is also responsible for the rebirth of Bayesian inference in the 1930s. We will not investigate the differences in these interpretations further, though we prefer the *subjective* interpretation throughout this work.

In this introduction, we cannot and will not cover all of Bayesian theory. For an easily accessible introduction, we recommend Kruschke [2010]. For a more formal introduction, consider Robert [2007].

In the following chapters, a basic knowledge of probability theory and mathematical notation is assumed.



Thomas Bayes (1701–1761) it is dubious whether this picture is really him, but no verified portrait of him survived. Image taken from Wikipedia [2015a]

2.1 BAYES' THEOREM

Bayes' theorem[1] is named after Reverend Thomas Bayes (1701–1761) and is also referred to as Bayes' law or Bayes' rule. Bayes' theorem expresses the conditional probability (or *posterior probability*) of an event A after B is observed in terms of the *prior probability* of A, prior probability of B, and the conditional probability of B given A:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

It is derived from the rules of conditional probability and shows the relation between two conditionals that are the reverse of each other.

A more general definition is as follows:

Theorem 2.1 (Bayes' Theorem)**.** *Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then for $i = 1, 2, \ldots$,*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}. \tag{1}$$

2.2 THE BAYESIAN INFERENCE MODEL

As the name suggests, Bayesian inference is based on Bayes' theorem. Let's look at (1) again, now replacing B with observations $\mathbf{X} = x_1, x_2, \ldots$, A with parameter set θ, and probabilities $\mathbb{P}(\cdot)$ with densities p, and we get the following:

in some literature, you will find the notation π or f for the densities

$$p(\theta|\mathbf{X}) = \frac{p(\theta, \mathbf{X})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} \tag{2}$$

where $p(\mathbf{X}) = \sum_\theta p(\theta)p(p\mathbf{X}|\theta)$, the sum of all possible values of θ in the discrete case, or $p(\mathbf{X}) = \int p(\theta)p(p\mathbf{X}|\theta)d\theta$ if θ is continuous. Note that, because p is a distribution, it is subject to its distribution parameters, usually called *hyperparameters*. So the formal definition of (2) is

Definition 2.1. *The model for Bayesian inference is defined as follows:*

$$p(\theta|\mathbf{X}, \alpha) = \frac{p(\mathbf{X}|\theta)p(\theta|\alpha)}{p(\mathbf{X}|\alpha)} \tag{3}$$

where

1 Stigler [1983] claims that Bayes' theorem has originally been discovered earlier by Nicholas Saunderson (1682–1739), who became the Lucasian Professor of Mathematics at Cambridge at age 29, in succession to Isaac Newton.

- $\mathbf{X} = x_1, x_2, ..., x_n$ *is a set of* n *observed data points*
- $\theta = \theta_1, \theta_2, ...$ *is the parameter of the distribution of data point* $x \sim p(\theta|\alpha)$
- α *is the hyperparameter of the parameter,* $\alpha \sim p(\theta|\alpha)$

In this model, $p(\theta|\alpha)$ is called the *prior distribution*, $p(\mathbf{X}|\theta)$ is the *likelihood* of $\mathbf{X}$ under a model (sometimes called the *sampling distribution*), and $p(\theta|\mathbf{X}, \alpha)$ is the *posterior distribution*, also called the *joint posterior distribution*, of parameter set θ that expresses uncertainty about θ after taking both prior and data into account.

The denominator

$$p(\mathbf{X}|\alpha) = \int_\theta p(\mathbf{X}|\theta)p(\theta|\alpha)d\theta$$

is the distribution of the observed data marginalized over the parameters and therefore called the *marginal likelihood* of $\mathbf{X}$, or the *prior predictive distribution*[2] of $\mathbf{X}$. Because it does not depend on θ, it may be set to an unknown constant C. The prior predictive distribution indicates what $\mathbf{X}$ should look like, given the model, before $\mathbf{X}$ has been observed. It is unconditional on previous observations and therefore *prior*. It is *predictive* as well, because it is the distribution for a quantity that is observable [Gelman et al., 2014]. The presence of the marginal likelihood normalizes the joint posterior distribution, $p(\theta|\mathbf{X}, \alpha)$, makes sure it is a proper distribution, and integrates to 1.

In Bayesian statistics, the marginal likelihood *is sometimes called the* evidence *or* model evidence.

If we simplify $p(\mathbf{X}|\alpha)$ to a constant C, (3) becomes:

$$p(\theta|\mathbf{X}, \alpha) = \frac{p(\mathbf{X}|\theta)p(\theta|\alpha)}{C} \propto p(\mathbf{X}|\theta)p(\theta|\alpha).$$

In words:

$$p(\theta|\mathbf{X}, \alpha) \propto \text{Likelihood} \times \text{Prior}.$$

a *is* proportional *to* b, *if* $a = m \cdot b$, *with a constant* m. *The relation is denoted* $a \propto b$.

We can therefore define:

Definition 2.2. *A Bayesian statistical model is made of a parametric statistical model,* $p(\mathbf{X}|\theta)$, *called likelihood, and a prior distribution on the parameters,* $p(\theta|\alpha)$ *[Robert, 2007].*

The goal in Bayesian inference is to quantify and manipulate your degrees of beliefs. In other words, Bayesian inference is the analysis of beliefs. That means, we are usually not interested in summarizing the unnormalized joint posterior distribution, but the marginal distributions of the parameters.

Note that we skipped the hyperparameter(s) α *for readability.*

2 This term goes back to Jeffreys [1961]

STRUCTURE OF BAYESIAN INFERENCE

(a) $p(\theta)$ is the *prior distribution* for parameter(s) θ. It is the distribution of θ before any data is observed. It takes *prior* knowledge into account.

(b) $p(\mathbf{X}|\theta)$ is the *likelihood*, the probability distribution of $\mathbf{X}$ given the model parameters θ.

(c) $p(\theta|\mathbf{X})$ is the *(joint) posterior distribution*. It is the distribution of the parameter(s) after taking both the prior and the observed data into account.

2.3 PRIOR DISTRIBUTIONS

In Bayesian inference, the *prior distribution* represents the information about an uncertain parameter θ before data are taken into consideration.

The *prior distribution* affects the *posterior distribution*, because the prior is multiplied by the *likelihood* and then normalized to yield the posterior. However, as the sample size increases, the prior will have less impact on the posterior. For small sample sizes, well-defined and reasonable choices of prior distributions are important.

Traditionally, there have been two categories of priors: *uninformative priors* and *informative priors.* However, in recent years, more distinction of priors has been made, and we stick to the definition of Gelman et al. [2014] where three categories of priors are distinguished:

(a) Prior distributions giving numerical information that is crucial to estimation of the model. This would be a traditional *informative prior*, which might come from a literature review or explicitly from an earlier data analysis.

(b) Prior distributions that are uniform, or nearly so, and basically allow the information from the likelihood to be interpreted probabilistically. These are *noninformative* or *uninformative priors*, or maybe, in some cases, *weakly informative*.

(c) Prior distributions that are not supplying any controversial information but are strong enough to pull the data away from inappropriate inferences that are consistent with the likelihood. This might be called a *weakly informative prior*.

In section 2.3.4, we will present a fourth prior, the *least informative prior*. This prior is favored by objective Bayesians.

2.3.1 *Informative Priors*

When prior information about θ is available, you should include that in your model for the distribution of θ. For example, say you want to model the mean weight of male adults in a country. The mean weight follows a normal distribution and you even know that weights above, say 400kg, are very unlikely, if not impossible.

The Guinness world record for the heaviest human is 394kg [Records, 2012].

In many situtations, it is difficult to estimate all possible values of θ, but a prior distribution should always include all plausible values. In the case of a normal distribution, for example, it does not necessarily have to be centered around the true value, because the data often overwhelm any reasonable prior specification. To ensure that the data will not outweigh the prior, Ibrahim and Chen [2000] introduced the *power prior*. This is especially important if your current model is based on a previous model (which is common practice), so data from the previous model also influences the new model. This cumulative effect may have a tremendous impact and should be avoided. If both data points are similar, this may increase the precision, but if they differ, you may be trapped in the *tails* of the prior distribution for θ, i.e., the prior distribution contributes less density to the tails.

2.3.2 *Uninformative Priors*

Technically, an *uninformative* or *noninformative prior* is a prior that has no influence on the posterior. Therefore, the posterior is identical to the likelihood. In practice, this often does not work out so well. For example, the famous *Jeffreys's Prior* [Jeffreys, 1961], which is defined as

$$p(\theta) \propto \sqrt{I(\theta)},$$

Jeffreys's Prior is named after Sir Harold Jeffreys, *so while "Jeffreys' Prior" or "Jeffreys Prior" are viable spellings,* Jeffreys *himself preferred "Jeffreys's Prior".*

where $I(\theta)$ is the Fisher-Information of θ, is an uninformative prior in the traditional sense. Another example is the use of a flat[3] prior, but this prior often does not retain its flatness under transformations and can be quite influentual. We will see in the next sections that there exists no such thing as a truly *uninformative prior*.

2.3.3 *Weakly Informative Priors*

The concept of *weakly informative priors* has been introduced by Gelman et al. [2014]. A prior is characterized as *weakly informative* if it is proper (we'll come to this later) but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available. Many problems have some natural constraints that allow a weak informative prior. For example, the great-

3 The Uniform Distribution is usually referred to as *the* flat prior. Its PDF is a flat line; this is the reason to call the prior *flat*. In section 2.3.4, we will see that this prior is often not that uninformative.

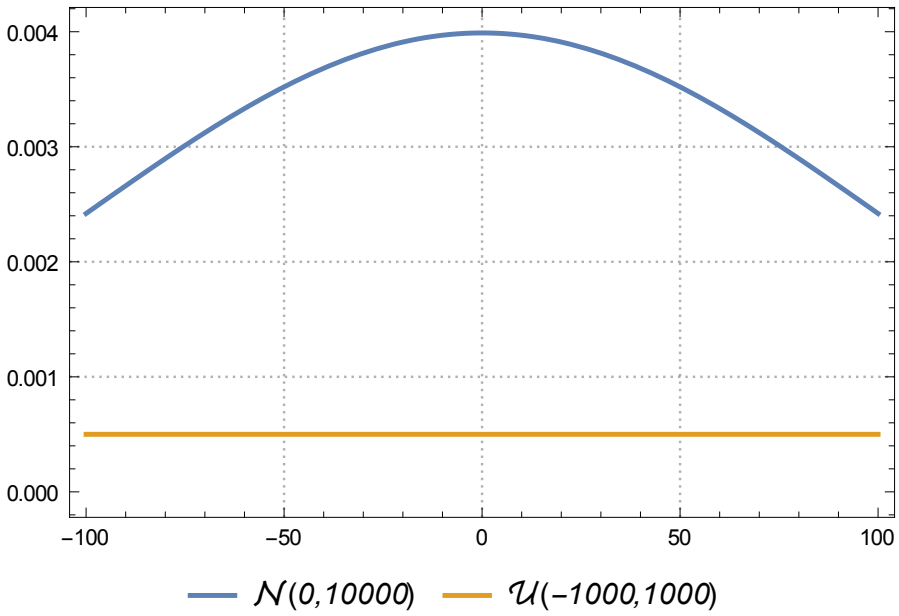


Figure 2.1: A comparison of the Normal Distribution and the Uniform Distribution

circle distance between two cities cannot be larger than the perimeter of the earth ($\approx$ 43000 km).

As this example shows, constructing a weakly informative prior can be tricky: virtually all statistical models are weakly informative, e.g., they are restricted by the size of the earth (or the universe) or the weight of an atom. A model should have enough prior information to learn from the data, but the prior information must be weak enough to actually be able to learn from it, i.e., the prior should not dominate the model.

Gelman [2008] suggests $\theta \sim \mathcal{N}(0, 10000)$ as a weakly informative prior for a centered and scaled predictor, where θ is normally distributed with $\mu = 0$ and $\sigma^2 = 10000$. In Figure 2.1, you can see that the distribution is very flat, even compared with a uniform distribution. Although it is not perfectly flat, and this yields good properties for numerical approximation algorithms. Regions with flat densities are a particular problem in frequentist inference but can be problematic in Bayesian inference as well (we will see that when we come to MCMC algorithms).

2.3.4 *Least Informative Priors*

The difference between *weakly informative* and *least informative* is hard to define. A *least informative prior* is usually a prior that is determined only by the model and the observed data. As we have seen in the examples in section 2.3.3, this is not necessarily the case for *weakly informative priors*.

Flat Priors

We have already briefly mentioned the flat prior as an *uninformative prior*. However, if θ is uniformly distributed from $-\infty$ to ∞, $\theta \sim \mathcal{U}(-\infty, \infty)$, then this prior is not proper, i.e., does not integrate to 1, because the integral of $p(\theta)$ is infinity. Even though this allows us "to let the data speak for themselves'", such a prior should be avoided; it may cause the posterior to become improper as well and thus invalid.

Both BAYES and LAPLACE used flat priors in their examples; LAPLACE's use of this prior distribution was later referred to as the *principle of indifference* or *principle of insufficient reason*, and is now called the *flat prior*. A *flat prior* seems reasonably uninformative because all values are considered equiprobable. But let's look at a binomial distribution $X \sim \text{Binom}(n, \theta)$, where we want to put a prior on θ. $\theta \in [0, 1]$, therefore the flat prior on θ is the uniform distribution $p(\theta) = 1$. A valid new parametrization to map θ to the full scale of reals is the logit transformation $\rho = \log \frac{\theta}{1-\theta}$. But under this transformation, the prior $p(\theta)$ is not flat any more. This shows that a prior is uninformative in one parametrization, but becomes informative in another, and this is especially problematic in higher dimensions. LAPLACE was aware that this prior was not truly uninformative and used it as a *least informative prior* [Gelman et al., 2014].

Jeffreys's Prior

We have already looked at *Jeffreys's Prior* as an *uninformative prior* in section 2.3.2, and it is quite reasonable in one dimension. Jeffreys's Prior is based on the principle of invariance: one should be able to apply this prior to certain situations, apply a change of variable, and still get the same answer [Jeffreys, 1961]. In multi-parameter situations, Jeffreys's prior may encounter inappropriate aspects and may lose its uninformative status, so we classify it as a *least informative prior* as well.

2.3.5 *Proper and Improper Priors*

We have already seen that it is important for the prior distribution to be proper.

Definition 2.3 (Proper and Improper Prior Distributions)**.** *A prior distribution is* improper*, when*

$$\int p(\theta) d\theta = \infty.$$

On the other hand, we call a prior density $p(\theta)$ proper *if it does not depend on data and integrates to 1. (If* $p(\theta)$ *integrates to any positive finite value,*

it is called an unnormalized density *and can be renormalized – multiplied by a constant – to integrate to 1 [Gelman et al., 2014].*

For example, defining the prior density for σ^2 as $p(\sigma^2) \propto (\sigma^2)^{-1}$ is an improper prior, having an infinite integral over the range $(0, \infty)$.

As we have already noted, the unbounded uniform prior distribution is improper, because $p(\theta) \propto 1$, for $-\infty < \theta < \infty$.

Improper priors should be avoided, and posterior distributions from improper priors must be interpreted carefully; the posterior is non-integrable and thus it should only be seen as an approximation in situations where the likelihood dominates the prior density.

2.3.6 *Conjugate Priors*

When the posterior distribution $p(\theta|\mathbf{X})$ is in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood [Raïffa and Schlaifer, 1961].

Definition 2.4 (Conjugate Prior Distributions)**.** *If $\mathcal{F}$ is a family of sampling distributions $p(\mathbf{X}|\theta)$, and $\mathcal{P}$ is a family of prior distributions for θ, then the family $\mathcal{P}$ is* conjugate *for $\mathcal{F}$ if*

$$p(\theta|\mathbf{X}) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}.$$

[Gelman et al., 2014]

For example, the family of normal distributions is self-conjugate, i.e., a conjugate prior for a normal likelihood function. If the likelihood is normal, then choosing a normal prior will result in a posterior distribution that is also normal. All members of the exponential family have conjugate priors; see Robert [2007] and Gelman et al. [2014] for a catalog. Using a conjugate prior leads to a closed-form expression of the posterior and can thus be computed analytically rather than numerically.

2.4 LIKELIHOOD

Although the concept *likelihood* was used by both Bayes and Laplace, it was formally introduced by Fisher [1921]. As we have already defined in Definition 2.2, in a model for Bayesian inference, the prior and the likelihood must be specified, analytically or approximatevily. The likelihood $p(\mathbf{X})$ contains the the information provided by the sample data. It is defined as:

The likelihood function is sometimes denoted as $L(\theta|\mathbf{X}) = p(\mathbf{X}|\theta)$.

$$p(\mathbf{X}|\theta) = \prod_{i=1}^{n} p(x_i|\theta), \quad \text{where } \mathbf{X} = x_1, x_2, ..., x_n.$$

Because data $\mathbf{X}$ only affect the posterior through the likelihood function, Bayesian inference obeys the so-called *Likelihood Principle:*

Definition 2.5 (Likelihood Principle)**.** *The information brought by an observation* $\mathbf{X}$ *about* θ *is entirely contained in the likelihood function* $p(\mathbf{X}|\theta)$*. Moreover, if* x_1 *and* x_2 *are two observations depending on the same parameter* θ*, such that there exists a constant* C *satisfying*

$$p(x_1|\theta) = C \cdot p(x_2|\theta)$$

for every θ*, they then bring the same information about* θ *and must lead to identical inferences [Robert, 2007].*

The *likelihood principle* states that two likelihood functions are equivalent if one is a scalar multiple of the other. The *likelihood* gives a measure of how *likely* a certain value of θ is. This must not be confused with probability[4] in a frequentist manner, although the likelihood principle is the true connection between Frequentists and Bayesians, in a sense that Bayesian inference obeys it and frequentist inference does not:

> [Frequentist] procedures typically violate the likelihood principle, as long-run behavior under hypothetical repetitions depends on the entire distribution $p(x|\theta)$, $x \in \mathbf{X}$ and not only on the likelihood [Bernardo and Smith, 2009].

2.5 BAYESIAN COMPUTATION

The recipe for evaluating a problem in Bayesian inference is usually quite simple: it boils down to the integration of a ratio of integrals. For example, if we are interested in the posterior mean of the function $g(x)$, (3) comes down to computing

$$\mathbb{E}[g(\theta)|\mathbf{X}] = \frac{\int g(\theta)p(\theta)p(\mathbf{X}|\theta)d\theta}{\int p(\theta)p(\mathbf{X}|\theta)d\theta}.$$

To evaluate these integrals, a number of numerical quadrature methods exist. However, these quadrature methods have limited use for Bayesian integration problems, because the choice of these methods depends on the location and shape of the posterior, and the number of evaluations of the posterior density grows exponentially as a function of the number of components of θ [Albert, 2009].

Numercial approximation algorithms in Bayesian inference include, among others: Approximate Bayesian Computation (ABC) [Rubin, 1984], Importance Sampling [Hastings, 1970], Laplace Approximation [Azevedo-Filho and Shachter, 1994], Markov Chain Monte Carlo (MCMC) [Metropolis et al., 1953], and Variational Bayes [Bishop, 2006]. We will only cover MCMC and not go into detail about the other algorithms.

4 *Probability* is used to quantify unknown outcomes based on known parameters, whereas *likelihood* is the other way round, it allows us to estimate unknown parameters based on known outcomes.

2.6 MARKOV CHAIN MONTE CARLO METHODS



Andrey Andreyevich Markov (1856–1922). Image from Wikipedia [2015b]

MCMC algorithms have their origins in statistical physics and became more and more popular in Bayesian inference as a result of advances in computational power. Markov chain simulations draw parameters θ from approximate distributions, eventually leading to the target distribution $p(\theta|\mathbf{X})$ as its *equilibrium distribution*. The quality of the sample improves at each step and each new draw θ_{t+1} is dependent on the previous θ_t but not on $(\theta_0, ..., \theta_{t-1})$, hence it is generating a Markov chain. Following a sufficient *burn-in period* (with k steps), the chain approaches its stationary distribution, which we will see in an example shortly. The key to the method's success, however, is not the Markov property, but rather that the approximate distributions are improved at each step [Gelman et al., 2014]. The following definitions are based on Robert and Casella [2005]; for a more thorough introduction to Markov chains see Norris [1998].

Definition 2.6 (Markov chain). *A sequence $X_0, X_1, ..., X_n, ...$ of random variables is a* Markov chain, *denoted by (X_n), if, for any t, the conditional distribution of X_t given $x_{t-1}, x_{t-2}, ..., x_0$ is the same as the distribution of X_t given x_t; that is*

$$P(X_{k+1} \in A|x_0, x_1, x_2, ..., x_k) = P(X_{k+1} \in A|x_k). \tag{4}$$

The Markov chain is time homogeneous, *or simply* homogeneous, *if the distribution of $(X_{t_1}, ..., X_{t_k})$ given x_{t_0} is the same as the distribution of $(X_{t_1-t_0}, X_{t_2-t_0}, ..., X_{t_k-t_0})$ given x_0 for every k and every $(k+1)$-uplet $t_0 \leqslant t_1 \leqslant ... \leqslant t_k$.*

Monte Carlo methods are a class of stochastics algorithms that rely on repeated pseudo-random sampling for computation. The combination of Markov chains and Monte Carlo methods is called MCMC. The most prominent MCMC samplers are the random-walk Metropolis-Hastings[5] and Gibbs sampling.

Algorithm 2.1. *The Metropolis-Hastings Algorithm*

1: Draw a starting point $\theta^{(0)}$, for which $p(\theta^{(0)}|\mathbf{X}) > 0$, from a *starting distribution* $p_0(\theta)$
2: **for** $t = 1, 2, ...$ **do**
3: Sample a proposal θ^* from a *proposal distribution* $J_t(\theta^*|\theta^{(t-1)})$.
4: Calculate the ratio of the densities

$$r = \frac{p(\theta^*|\mathbf{X})/J_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|\mathbf{X})/J_t(\theta^{(t-1)}|\theta^*)}.$$

5 The Metropolis algorithm is among the ten algorithms that have had the greatest influence on the development and practice of science and engineering in the 20th century [Beichl and Sullivan, 2000].

5: Set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(t-1)} & \text{otherwise.} \end{cases}$$

6: **end for**

Figure 2.2 shows an example of the first steps in a Markov chain simulation. Here $\theta \sim \mathcal{N}\left(\binom{1}{2}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$ is a vector with two components, $\theta = (\theta_1, \theta_2)$, and the colored lines represent the paths of four independent random walks, each starting in a corner of the figure space. The target distribution is indicated by the contour lines, where the innermost circle represents the maximum of the distribution function.

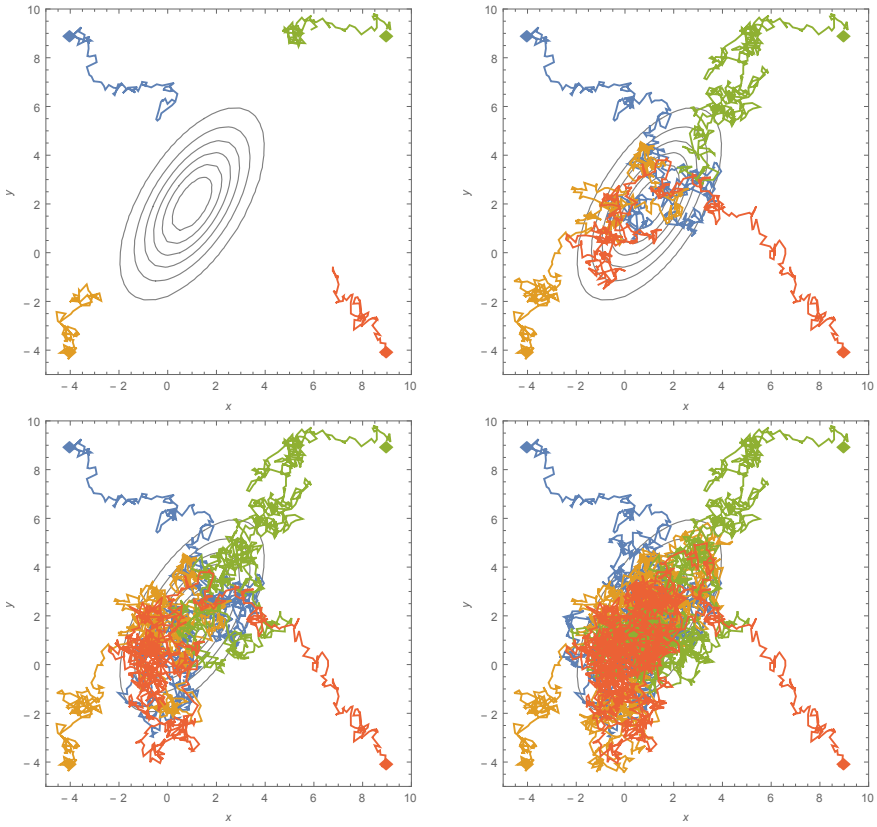


Figure 2.2: Four independent Markov chains sampled with the Metropolis Algorithm. The target distribution is a binormal distribution, whereas the proposal distribution is a normal distribution centered at the previous sample and with variance equal to 1. Starting points of each chain are marked by a solid diamond symbol. (*Top left*) After 100 iterations. (*Top right*) After 500 iterations. Still far from convergence. (*Bottom left*) After 1000 iterations. (*Bottom right*) After 2000 iterations, points cluster around the modus of the target distribution.

Gibbs sampling, or a *Gibbs-sampler*, is a special case of the Metropolis-Hastings algorithm, developed by Geman and Geman [1984], wherein

the random value (i.e., the ratio r) is always accepted. Suppose θ has been divided into d subvectors $\theta = (\theta_1, ..., \theta_d)$, then the Gibbs sampler cycles through the subvectors of θ, drawing each subset conditional on the value of all the others. At each iteration t, each $\theta_j^{(t)}$ is sampled from the conditional distribution $p(\theta_j|(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_d^{(t)}), \mathbf{X})$, so each parameter subvector θ_j from the distribution of that parameter is conditioned on all the other parameter subvectors, making use of the latest values and updating the variable with its new values [Gelman et al., 2014].

2.7 MODEL VALIDATION

We have now accomplished the three steps of Bayesian analysis: modeling our prior knowledge, constructing the likelihood, and computing the posterior distribution of all estimands. We now come to the relatively easy step of assessing our model. Bayesian inference has *Bayes' factors* [Jeffreys, 1961] for model validation, which serve as a kind of hypothesis testing for model fitting. Yet the common way to assess the adequacy of the fit of the model to the data and to our substantial knowledge is the use of a genereralization of the *Akaike Information Criterion (AIC)*, called the *Deviance Information Criterion (DIC)*:

$$\text{DIC} = -2 \log p(\mathbf{X}|\theta).$$

So deviance is defined as -2 times the log-likelihood. The difference from the AIC is that it replaces the maximum likelihood estimator by θ and therefore does not require the calculation of the maximum of the likelihood over θ, which is not readily available in MCMC simulations. Note that, like AIC, the DIC is only valid when the posterior distribution is approximately multivariate normal.

Deviance does not take model complexity into account; in our case, this is the effective number of parameters in the model. There are two approaches to measure model complexity, one by Spiegelhalter et al. [2002] and the other by Gelman et al. [2014]. Ando [2011] provides a discussion of several Bayes' model selection criteria and developed a method that slightly improves both methods already mentioned.

The *Watanabe-Akaike or Widely Applicable Information Criterion (WAIC)* [Watanabe, 2009] can be viewed as an improvment over the DIC for Bayesian models, because it is fully Bayesian[6]. WAIC is computationally more expensive because it requires the log pointwise posterior over numerous samples but, when available, it is preferred over AIC and DIC. The WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate, which is especially relevant in a predictive context.

6 Plummer [2008] and van der Linde [2005] showed that the DIC is not fully Bayesian because it is based on a point estimate.

Still, formulas such as the AIC, DIC, and WAIC fail in various examples: AIC does not work in settings with strong prior information, DIC gives nonsensical results when the posterior distribution is not well summarized by its mean, and WAIC relies on a data partition that would cause difficulties with structured models such as for spatial or network data [Gelman et al., 2014]. In these cases, the use of the already mentioned *Bayes' factors* can be appropriate, but generally they are not recommended because the marginal likelihood is highly sensitive to arbitrarily assigned aspects of the model and often biased. See Robert et al. [2011] for discussion and Carlin and Louis [2011] for an introduction to Bayes' factors and examples.

3

THE LOGNORMAL DISTRIBUTION

OUTLINE

In this chapter we address the problem of Bayesian estimation of the parameters of the univariate *Lognormal Distribution*. The lognormal distribution is commonly used as a model in the analysis of gene expression data, but is used in many applications in the fields of physical sciences, social sciences and engineering as well [Aitchison and Brown, 1957]. In this section we will define the univariate and bivariate lognormal distribution and briefly explore the distribution of sums of lognormally distributed random variables.

A random variable Y is *lognormally distributed*, $Y \sim \mathcal{LN}(\mu, \sigma^2)$, with parameters $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$, if it has the probability density function:

$$f(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma y} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) & y > 0 \\ 0 & y \leqslant 0. \end{cases} \tag{5}$$

Therefore, a lognormally distributed random variable is a random variable whose logarithm is normally distributed. The lognormal distribution is skewed with a longer tail to the right. Besides the classic reference on the lognormal distribution by Aitchison and Brown [1957], the compendium edited by Crow and Shimizu [1987] and the books by Patel and Read [1996] and Balakrishnan and Chen [1999] provide more recent updates.

The aritmethic mean and variance of the lognormal distribution are:

$$\begin{aligned} E[Y] &= e^{\mu + \frac{1}{2}\sigma^2}, \\ \text{Var}[Y] &= e^{2\mu+\sigma^2}\left(e^{\sigma^2-1}\right) = (E[Y])^2\left(e^{\sigma^2-1}\right). \end{aligned}$$

If the mean and variance are known, the location (μ) and scale (σ) parameters can be obtained:

$$\begin{aligned} \mu &= \ln(E[Y]) - \frac{1}{2}\ln\left(1 + \frac{\text{Var}[Y]}{(E[Y])^2}\right) = \ln(E[Y]) - \frac{1}{2}\sigma^2. \\ \sigma^2 &= \ln\left(1 + \frac{\text{Var}[Y]}{(E[Y])^2}\right) \end{aligned}$$

The *likelihood* of the lognormal distribution is:

$$L(y; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma y_i} \exp\left(-\frac{(\ln y_i - \mu)^2}{2\sigma^2}\right), \quad y > 0. \tag{6}$$

The Bivariate Lognormal Distribution

Let the random variable $\mathbf{X} = (X_1, X_2)$ be bivariate *normally* distributed, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \end{pmatrix}$ is the mean vector and, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{x_1}^2 & \rho_N \sigma_{x_1} \sigma_{x_2} \\ \rho_N \sigma_{x_1} \sigma_{x_2} & \sigma_{x_2}^2 \end{pmatrix}$ is the covariance matrix.

$$\rho_N = \frac{\operatorname{Cov}(X_1, X_2)}{\sigma_{x_1} \sigma_{x_2}}$$

denotes the correlation between X_1 and X_2. The bivariate *lognormal* random variable $\mathbf{Y} = (Y_1, Y_2)$ is derived from $\mathbf{X}$ by exponentiation of the X_1 and X_2 components of $\mathbf{X}$. It has the same parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\mathbf{X}$.

Law and Kelton [2000] show that, if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ follows a bivariate normal distribution, then $\mathbf{Y} = \exp(\mathbf{X})$ has a bivariate lognormal distribution with mean

$$\mathrm{E}[\mathbf{Y}]_i = e^{\mu_i + \frac{1}{2}\Sigma_{ii}}$$

and covariance matrix

$$\operatorname{Var}[\mathbf{Y}]_{ij} = e^{\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj})}(e^{\Sigma_{ij}} - 1), \quad i, j \in \{1, 2\}.$$

Let ρ the correlation coefficient of the transformed variable,

$$\rho = \frac{e^{\rho_N \sigma_{y_1} \sigma_{y_2}} - 1}{\sqrt{(e^{\sigma_{y_1}^2} - 1)(e^{\sigma_{y_2}^2} - 1)}},$$

then, for $y_1, y_2 > 0$, the joint probability density function is

$$f(y_1, y_2) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\ln(y_1)-\mu_{y_1}}{\sigma_{y_1}}\right)^2 - 2\rho\frac{\ln(y_1)-\mu_{y_1}}{\sigma_{y_1}}\frac{\ln(y_2)-\mu_{y_2}}{\sigma_{y_2}} + \left(\frac{\ln(y_2)-\mu_{y_2}}{\sigma_{y_2}}\right)^2\right]\right)}{2\pi y_1 y_2 \sigma_{y_1} \sigma_{y_2}\sqrt{1-\rho^2}}. \tag{7}$$

The distributions of sums of lognormally distributed random variables

Let $X_j \sim \mathcal{LN}(\mu_j, \sigma_j^2)$ be independent lognormally distributed variables with possibly varying μ and σ parameters, and let $Y = \sum_{j=1}^{n} X_j$, the sum of n lognormally distributed parameters. No closed-form expression of the distribution of Y can be given, but it can be reasonably approximated by another lognormal distribution Z at the right tail. Its probability density function at the neighborhood of 0 does not resemble any lognormal distribution. A commonly used approximation due to Fenton [1960] (but previously stated by Wilkinson [1942]

and mathematically justified by Marlow [1967]) can be obtained by matching the mean and variance of another lognormal distribution:

$$\sigma_Z^2 = \ln\left[\frac{\sum_{j=1}^n e^{2\mu_j+\sigma_j^2}(e^{\sigma_j^2}-1)}{(\sum_{j=1}^n e^{\mu_j+\frac{1}{2}\sigma_j^2})^2}+1\right],$$
$$\mu_Z = \ln\left[\sum_{j=1}^n e^{\mu_j+\frac{1}{2}\sigma_j^2}\right]-\frac{\sigma_Z^2}{2}.$$

In the case that all X_j have the same variance, $\sigma_j = \sigma$, these formulas simplify to

This is usually assumed when modeling gene expression.

$$\sigma_Z^2 = \ln\left[(e^{\sigma^2}-1)\frac{\sum_{j=1}^n e^{2\mu_j}}{(\sum_{j=1}^n e^{\mu_j})^2}+1\right],$$
$$\mu_Z = \ln\left[\sum_{j=1}^n e^{\mu_j}\right]+\frac{\sigma^2}{2}-\frac{\sigma_Z^2}{2}.$$

4 MIXTURE MODELS

Outline

In this chapter we cover finite mixtures of distributions, especially the case of lognormal distributions. In mixture models simple distributions are made more complex and less informative. Bayesian inference on these models has been intensified since the introduction of MCMC algorithms, because the posterior distribution may be complicated and thus cannot be computed analytically. We will show why this is the case and present the MCMC approach to mixture distribution. In most parts of this chapter we follow Marin and Robert [2014], where additional examples can be found.

4.1 FINITE MIXTURE MODELS

Definition 4.1. *A* mixture distribution *is a combination of* k *known distributions* $f_j(x)$ *leading to the mixture model*

$$\sum_{j=1}^{k} p_j f_j(x), \qquad \sum_{j=1}^{k} p_j = 1, \quad p_j \geqslant 0,\ k > 1.$$

If all f_j*s are from the same parametric family (e.g., the lognormal distribution), with unknown parameters* θ_j*, this leads to*

$$\sum_{j=1}^{k} p_j f(x|\theta_j), \qquad \sum_{j=1}^{k} p_j = 1, \quad p_j \geqslant 0,\ k > 1. \tag{8}$$

with parameters including both the weights p_j *and the component parameters* θ_j $(j = 1, ..., k)$.

Note that, for example, the sum of two random variables of two different normal distributions, each with a different mean, will have two peaks and will not resemble a normal distribution. On the other hand, a mixture created from two normal distributions with the same mean will again be normally distributed and only have one peak. Figure 4.1 shows examples of different mixtures of lognormal distributions.

A mixture of two lognormal *distributions is not lognormally distributed even if the log-means are the same*

4.2 MIXTURE LIKELIHOODS AND POSTERIORS

Let $\mathbf{X} = (x_1, ..., x_n)$ be iid samples from model 8, $\mathbf{p} = (p_1, ..., p_k)$ and $\theta = (\theta_1, ..., \theta_k)$. Then the likelihood is:

$$L(\theta, \mathbf{p}|\mathbf{X}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i|\theta_j) \tag{9}$$

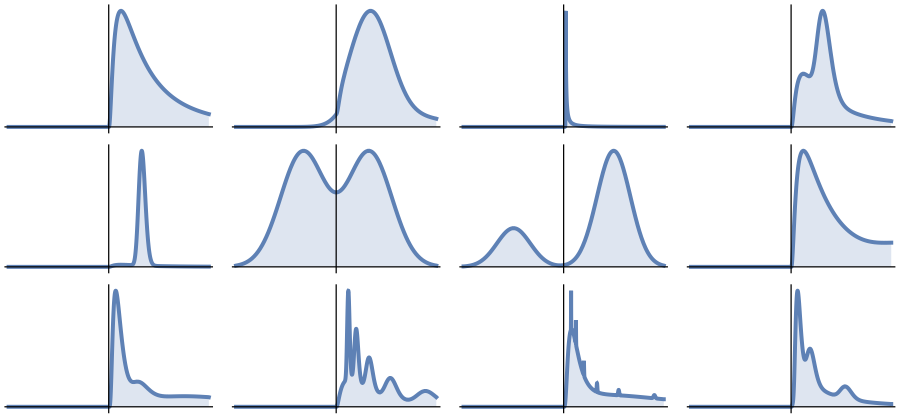


Figure 4.1: A variety of distribution shapes from lognormal mixtures

The likelihood consists of k^n terms and is computationally feasible in $\mathcal{O}(nk)$ operations. However, if the inner sums are expanded, the model is difficult to compute and thus precludes analytic solutions for maximum likelihood estimation.

For any prior $p(\theta, \mathbf{p})$, the posterior distribution of $(\theta, \mathbf{p})$ is proportional to:

$$p(\theta, \mathbf{p}|\mathbf{X}) \propto \left[\prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i|\theta_j)\right] \times p(\theta, \mathbf{p}). \tag{10}$$

We now investigate why the derivation of posterior expectations is only possible in an exponential time of order $\mathcal{O}(k^n)$. Marin and Robert [2014] suggest considering the missing-variable representation[1] of the mixture model where each x_i is associated with a variable z_i that indicates its component, i.e., the index z_i of the distribution from which it was generated. The resulting hierarchical model is:

$$z_i|\mathbf{p} \sim \mathcal{M}_k(p_1, ..., p_k),$$

where $\mathcal{M}_k$ denotes the multinomial distribution, and

$$x_i|z_i, \theta \sim f(\cdot|\theta_{z_i}).$$

The likelihood (9) modifies to:

$$L(\theta, \mathbf{p}|\mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})$$

and for the posterior we get

$$p(\theta, \mathbf{p}|\mathbf{X}, \mathbf{Z}) \propto \left[\prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})\right] \times p(\theta, \mathbf{p}),$$

1 A *missing-variable model* is a model in which the complexity originates from the fact that some piece of information about an original and more standard (simpler) model is *missing* [Marin and Robert, 2014]. Mixture models are the archetype of a missing-variable model because they have a rather simple representation but need complex processing.

where $\mathbf{Z} = (z_1, ..., z_2)$. Set $\mathcal{Z} = \{1, ..., k\}^n$ the set of the k^n possible values of the vector $\mathbf{Z}$, then $\mathbf{Z}$ can be composed into a partition of r subsets

$$\mathcal{Z} = \bigcup_{j=1}^{r} \mathcal{Z}_j,$$

with

$$\mathcal{Z}_j = \left\{ \mathbf{Z} : \sum_{i=1}^{n} \mathbb{1}_{z_i=1}, ..., \sum_{i=1}^{n} \mathbb{1}_{z_i=k} \right\},$$

which consists of all allocations with the given allocation size $(n_1, ..., n_k)$. If we use lexigraphical ordering on the $(n_1, ..., n_k)$'s this means that $j = 1$ corresponds to $(n_1, ..., n_k) = (n, 0, ..., 0)$, $j = 2$ to $(n_1, ..., n_k) = (n-1, 1, ..., 0)$, $j = 3$ to $(n_1, ..., n_k) = (n-1, 0, 1, ..., 0)$, and so on.

The posterior of $(\theta, \mathbf{p})$ can then be written in closed form:

$$p(\theta, \mathbf{p}|\mathbf{X}) = \sum_{\mathbf{Z}\in\mathcal{Z}} p(\theta, \mathbf{p}|\mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{r} \sum_{\mathbf{Z}\in\mathcal{Z}_i} p(\mathbf{Z}) p(\theta, \mathbf{p}|\mathbf{X}, \mathbf{Z}), \qquad (11)$$

where $p(\mathbf{Z})$ denotes the marginal posterior probability of the allocation $\mathbf{Z}$ conditional on the observations $\mathbf{X}$, derived by integrating out the parameters θ and $\mathbf{p}$.

In this posterior (11), each possible partition $\mathbf{Z}$ of the dataset is allocated a posterior probability $p(\mathbf{Z})$, and then a posterior distribution is constructed for the parameters conditional on this allocation. There are k^n terms in this sum so unfortunately this is very hard to compute. However, there are MCMC approaches that take advantage of the missing-variable structure and do not require the exploration of the k^n possible values of $\mathbf{Z}$ by only looking at the most likely ones.

4.3 MCMC FOR MIXTURES

We will now stick with the missing variable model and derive the Gibbs sampler for this model according to Marin and Robert [2014]. For the joint posterior distribution (10), the full conditional distribution of $\mathbf{Z}$ given data $\mathbf{X}$ and the parameters is proportional to

$$p(\mathbf{Z}|\mathbf{X}, \theta, \mathbf{p}) \propto \prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i}),$$

and can be computed at the cost of $\mathcal{O}(n)$. For distributions $f(\cdot|\theta)$, where a conjugate prior exists, the Gibbs sampler can be derived easily: If $\mathbf{p}$ and θ are independent, then, given $\mathbf{Z}$, the vectors $\mathbf{p}$ and $\mathbf{X}$ are independent:

$$p(\mathbf{p}|\mathbf{Z}, \mathbf{X}) \propto p(\mathbf{p}) f(\mathbf{Z}|\mathbf{p}) f(\mathbf{X}|\mathbf{Z}) \propto p(\mathbf{p}) f(\mathbf{Z}|\mathbf{p}) \propto p(\mathbf{p}|\mathbf{Z})$$

It can be seen that θ is also independent a posteriori from $\mathbf{p}$ given $\mathbf{Z}$ and $\mathbf{X}$ with density $p(\theta|\mathbf{Z}, \mathbf{X})$. The Gibbs sampler for this problem involves the successive simulation of $\mathbf{Z}$ and $(\mathbf{p}, \theta)$:

Algorithm 4.1. *The Gibbs Sampler for Mixture Distributions*

1: Draw starting vectors $\mathbf{p}^{(0)}$ and $\theta^{(0)}$
2: **for** $t = 1, 2, \ldots$ **do**
3: **for** $i = 1, \ldots, n$ **do**
4: Generate $z_i^{(t)}$ such that

$$P(z_i = j|\theta, \mathbf{p}) \propto p_j^{(t-1)} f(x_i|\theta_j^{(t-1)}).$$

5: **end for**
6: Generate $\mathbf{p}^{(t)}$ according to $p(\mathbf{p}|\mathbf{Z}^{(t)})$.
7: Generate $\theta^{(t)}$ according to $p(\theta|\mathbf{Z}^{(t)}, \mathbf{X})$.
8: **end for**

The marginal sampling distribution of the z_is is a multinomial distribution $\mathcal{M}_k(p_1, \ldots, p_k)$ for which a conjugate prior distribution on $\mathbf{p}$ exists, namely the *Dirichlet distribution*, $\mathbf{p} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$.

The Dirichlet distribution of order $k \geqslant 2$ with parameters $\alpha, \ldots, \alpha_k$ has a probability density function with respect to the Lebesgue measure on the Euclidean space $\mathbb{R}^{k-1}$

$$f(p_1, \ldots, p_{k-1}; \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} p_j^{\alpha_j - 1}$$

on the open $(k-1)$-dimensional simplex defined by

$$(p_1, \ldots, p_k) \in [0, 1]^k, \sum_{j=1}^{k} p_j = 1$$

and zero elsewhere.

If we denote $n_j = \sum_{l=1}^{n} \mathbb{1}_{j_l = j} (1 \leqslant j \leqslant k)$ the allocation sizes, the posterior distribution of $\mathbf{p}$ given $\mathbf{Z}$ is

$$\mathbf{p}|\mathbf{Z} \sim \mathcal{D}(n_1 + \alpha_1, \ldots, n_k + \alpha_k).$$

The fundamental feature of Gibbs sampling, and its derivation from conditional convergence, implies that it is often restricted in the width of its moves and that, in some situations, this restriction may jeopardize convergence [Marin and Robert, 2014]. When dealing with mixtures of distributions, conditioning on $\mathbf{Z}$ implies that the proposals for $(\theta, \mathbf{p})$ are quite concentrated and do not allow drastic changes in the allocations at the next step. Sometimes, e.g., when initialized in a saddlepoint, the Gibbs sampler needs a considerable number of iterations to obtain a significant modification of $\mathbf{Z}$ to leave this stable position. In mixtures, where two or more modes exist, this is often a problem, and there is little one can do about it as its components are given by the joint distribution. The best approach is to mix the Gibbs sampler with another MCMC algorithm. The Griddy-Gibbs algorithm (see chapter 5.1) we use later in our computations avoids this pitfall by using a grid structure to evaluate the model at a different set of points.

4.4 THE LABEL SWITCHING PROBLEM

The label switching problem is caused by the likelihood of a Bayesian mixture model being invariant under permutations of the indices of the components. For example, the normal mixtures $0.2 \cdot \mathcal{N}(1,2) + 0.8 \cdot \mathcal{N}(3,4)$ and $0.8 \cdot \mathcal{N}(3,4) + 0.2 \cdot \mathcal{N}(1,2)$ are exactly the same. There is no *first* component in this mixture. The permutation can change multiple times between MCMC iterations making it difficult to infer component-specific parameters of the model: in our example, the pairs (θ_1, p) and $(\theta_2, 1-p)$ are exchangeable.

In a k-component mixture, the number of modes of the likelihood is of order $\mathcal{O}(k!)$ as, if $((\theta_1, ..., \theta_k), (p_1, ..., p_k))$ is a local maximum of the likelihood function, so is $\tau(\boldsymbol{\theta}, \mathbf{p}) = (\theta_{\tau(1)}, ..., \theta_{\tau(k)}, p_{\tau(1)}, p_{\tau(k)})$ for every permutation $\tau \in \mathcal{P}_k$ the set of all permutations of $\{1, ..., k\}$. This makes maximization and even exploration of the posterior surface obviously harder because modes are separated by valleys that most samplers find difficult to cross [Marin and Robert, 2014]. This identifiability problem related to the exchangeability of the posterior distribution is called *label switching*.

One solution to the label switching problem is to adopt the *maximum a-posteriori* (MAP) estimator, which is equivalent to penalized maximum likelihood [Ciuperca et al., 2003]. Given an MCMC sample of size M, we can find a Monte Carlo approximation of the MAP estimator by taking $\boldsymbol{\theta}^{(i^*)}, \mathbf{p}^{(i^*)}$ such that

$$i^* = \arg \max_{i=1,...,M} p\left\{(\boldsymbol{\theta}, \mathbf{p})^{(i)} | \mathbf{X}\right\},$$

the simulated value that gives the maximal posterior density [Marin and Robert, 2014]. However, one of the main attractions of using a Bayesian approach is the ability to reflect the uncertainties related to our inference. This is of particular importance in mixture analysis (see Sperrin et al. [2010] for details and examples), and MAP does not allow this.

The problem is usually dealt with using methods to *relabel* the results of an MCMC run so that the components are in the same order at each iteration. Sperrin et al. [2010] divided the relabeling strategies into three categories:

- *Identifiability constraints* involve relabeling the output of the MCMC sampler so that the posterior satisfies a constraint on the component parameters. The constraint is chosen such that exactly one relabelling satisfies the constraint at each iteration of the sampler.

- *Deterministic relabeling algorithms* perform relabeling at each iteration of the MCMC output that minimizes the posterior expectation of some loss function.

- *Probabilistic* approaches acknowledge that there is uncertainty in the relabeling that should be selected on each iteration of the MCMC output.

The idea of the relabeling algorithms is as follows (see Stephens [2000], Sperrin et al. [2010], and Marin and Robert [2014]):

Let $\mathcal{P}_k$ be the k-permutation set and $\tau \in \mathcal{P}_k$. We suggest to minimize in τ an entropy distance summing the relative entropies between the $\mathbb{P}(z_t = j|\theta^{(i^*)}, \mathbf{p}^{(i^*)})$'s and the $\mathbb{P}(z_t|\tau\left\{\theta^{(i^*)}, \mathbf{p}^{(i^*)}\right\})$'s, namely

$$h(i,\tau) = \sum_{t=1}^{n}\sum_{j=1}^{k} \mathbb{P}(z_t = j|\theta^{(i^*)}, \mathbf{p}^{(i^*)}) \times \log\left\{\frac{\mathbb{P}(z_t = j|\theta^{(i^*)}, \mathbf{p}^{(i^*)})}{\mathbb{P}(z_t|\tau\left[(\theta^{(i)}, \mathbf{p}^{(i)})\right])}\right\}.$$

It follows for the selection of the permutations reordering the MCMC output:

Algorithm 4.2. *The Pivotal Reordering Algorithm*

1: **for** $i = 1, 2, ..., M$ **do**
2: Compute
$$\tau_i = \arg\min_{\tau in \mathcal{P}_k} h(i,\tau),$$
3: Set $(\theta^{(i)}, \mathbf{p}^{(i)}) = \tau_i\left\{(\theta^{(i)}, \mathbf{p}^{(i)})\right\}$.
4: **end for**

Due to this reordering, most iteration labels get switched to the same mode (the larger n, the more likely), and the identifiability problem is solved.

Part II

THE SHOWCASE

5 MCMC FOR STOCHASTIC PROFILING

Outline

In this chapter we will present an MCMC model for stochastic profiling of single-cell heterogenities. First, we will define the model and explain the priors and the samplers we used. We then present the results of our own experiments: we found the required sample sizes for different numbers of cells n in the tissue, check if the parameter n can be estimated, discuss finding the best n, and try various *wrong* n's to check its impact on estimation. Finally, we estimate some real data and compare it to the results of MLE. In the second section, we address the correlation of genes and present a model based on bivariate lognormal estimation. Again we compare it to MLE and analyze real data with it. In the last section, we try to generalize our model to multivariate lognormal estimation and will show the problems we had with this approach.

5.1 METHODS

Probabilistic Model

In our model, we used the same simple probabilistic model of gene expression as Bajikar et al. [2014]. We consider genes that exhibit two distinct regulatory states in a population of cells [Shalek et al., 2013], where each cell follows a lognormal distribution [Bengtsson et al., 2005]. We follow their definitions and modeled it as a mixture of two lognormal distributions:

$$f_{mix} = p \cdot f_1 + (1 - p) \cdot f_2, \tag{12}$$

where both f_1 and f_2 are PDFs of lognormal distributions as defined in (5) and share the same σ. For simplicity, we only considered a single gene[1]. Every random sample of the transcript is the sum of n independent single-cell expression measurements:

$$Y_i = \sum_{j=1}^{n} X_{ij} \tag{13}$$

where X_{ij} is the expression of the transcript in the j-th cell of the i-th random sample, occupying one of two regulatory states: one at a low

1 We therefore do not need the (g) superscript as in Bajikar et al. [2014], otherwise our definition are exactly the same.

level with log-mean μ_2, and one at a high level with log-mean μ_1; both share the same log-SD σ. Then the random sample Y_i characterizing the n-cell mixture and combining (12) and (13), has the PDF

$$f_n(y) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} f_{(j,n-j)}(y).$$

The binomial part $\binom{n}{j} p^j (1-p)^{n-j}$ represents the selection of cells with probabilities p and $1-p$ respectively. $f_{(j,n-j)}(y)$ is the density of a n-cell sum $\sum_{i=1}^{n} Z_i$ of iid random variables drawn from the mixture model:

$$Z_c = \begin{cases} \mathcal{LN}(\mu_1, \sigma^2) & \text{if} \quad 1 \leqslant c \leqslant j \\ \mathcal{LN}(\mu_2, \sigma^2) & \text{if} \quad j < c \leqslant n \end{cases}$$

The log-likelihood function for the model parameters given k random n-cell samples is

$$\ell(p, \mu_1, \mu_2, \sigma) = \sum_{i=1}^{k} \log \left[f_n \left(y_i | p, \mu_1, \mu_2, \sigma \right) \right]$$

We defined four parameter sets to address different kinds of gene expression, where three sets are based on real settings and one parameter set features more theoretical values (Figure 5.1). For MCMC sampling we used two different algorithms: the Griddy-Gibbs (GG) sampler and the Hit-and-Run-Metropolis (HARM) algorithm.

The Griddy-Gibbs Sampler

The Griddy-Gibbs (GG) sampler from Ritter and Tanner [1992] is an extension to the Gibbs sampler (Algorithm 2.1 and Algorithm 4.1) where the model is evaluated at a discrete set of points, the user-defined grid. It is a componentwise sampler with a 100% acceptance rate. The sampler is especially powerful when the model contains discrete parameters as in our case when we try to find the correct n, i.e., the number of cells in the tissue. We found a grid from -0.1 to 0.1 with 5 steps to work best; a larger grid dramatically raised the convergence rate. The GG produces fewer autocorrelated samples (Figure 5.2) but is one of the slower algorithms, performing at least 15 times slower than the HARM. We did not encounter any label switching when using the GG, and convergence can usually be achieved after as little as 10000 iterations.

The Hit-And-Run Metropolis

The Hit-And-Run Metropolis algorithm was introduced by Turchin [1971] and made popular by Smith [1984]. It is a general-purpose algorithm and a variation of the Random-Walk-Metropolis [Gelman

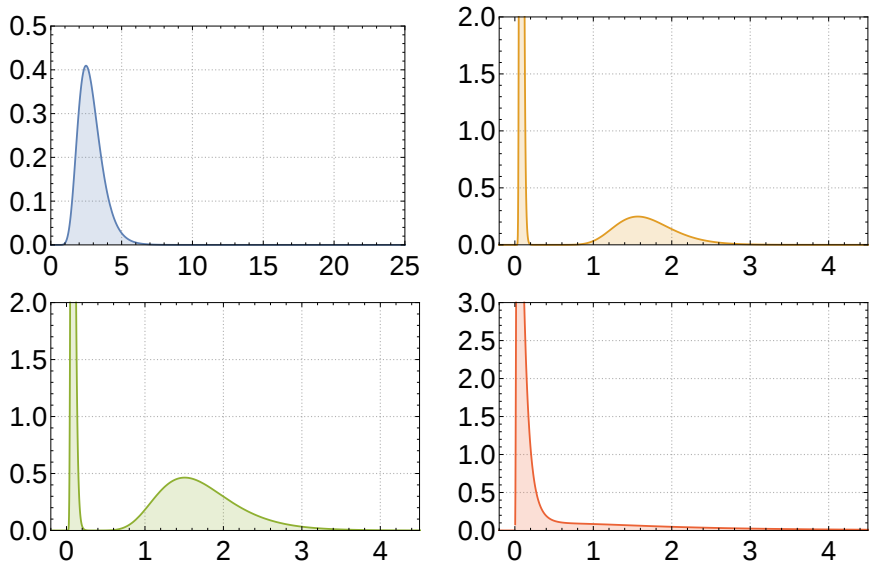


Figure 5.1: The four parameter sets we used as gene expression data. All are mixture distributions of two lognormal distributions: $f(x) = p \cdot \mathcal{LN}(\mu_1, \sigma) + (1-p) \cdot \mathcal{LN}(\mu_2, \sigma)$. (*Top left, blue*) $p = 0.8, \mu_1 = 1, \mu_2 = 6, \sigma = 0.3$. These are values that work well on a theoretical basis because the log-μs are well separated and the log-σ is small. It can be observed that the density function is almost unimodal. (*Top right, orange*) $p = 0.225, \mu_1 = 0.5, \mu_2 = -2.5, \sigma = 0.225$. These values are taken from Bajikar et al. [2014] and are based on realistic data. Both log-μs are less separated but the log-σ is still small. (*Bottom left, green*) $p = 0.55, \mu_1 = 0.5, \mu_2 = -2.5, \sigma = 0.3$. A p closer to 0.5 has been found to be problematic, so we test that as well. The remaining parameters are the same as in the previous set. (*Bottom right, red*) $p = 0.2, \mu_1 = 0.5, \mu_2 = -2.5, \sigma = 0.8$. We raised the log-σ to 0.8. The density function now becomes unimodal.

et al., 1996]. The HARM features multivariate proposals, so all parameters are proposed at once, and a Metropolis step for rejection. It is a very simple algorithm and one of the fastest. We managed to tune the acceptance rate to 15%, which is very low, but acceptable. Sometimes label switching occured with the HARM, but usually only when the sampling did not converge due to high n or if the sample size was too small.

The MCMC Model

We model our single-cell regulatory heterogenities as a mixture of two lognormal states. Our lognormal-lognormal (LN-LN) mixture is therefore charaterized by four parameters: the log-mean expression of the two regulatory states (μ_1 and μ_2), the log-SD[2] (σ), and the

2 It is assumed that both lognormal states share a common SD.

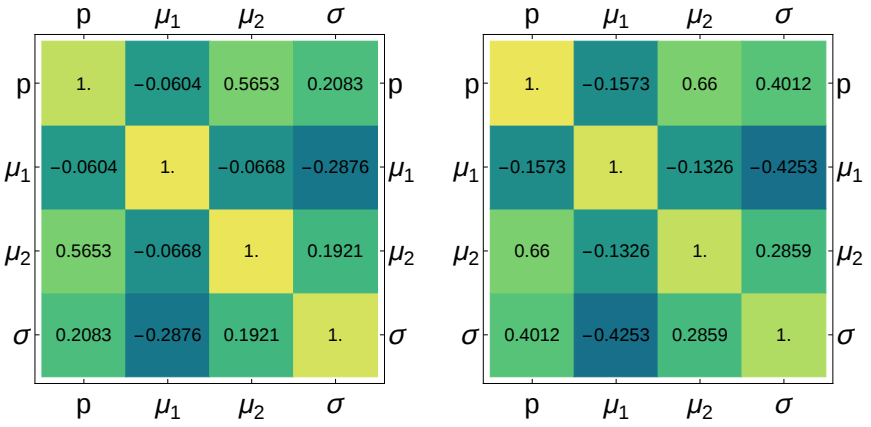


Figure 5.2: Correlation matrices for samples from the GG (*left*) and HARM (*right*) samplers. Very light and very dark colors mean a greater positve and negative correlation respectively. The GG produces fewer autocorrelated samples, but the HARM is very close. In both cases, the correlation between the samples is small. Both correlation matrices are average values of 100 runs of datasets with parameters $n = 10, p = 0.8, \mu_1 = 1, \mu_2 = 6, \sigma = 0.3$ and a sample size of 100.

expression frequency (p). Our tissue is a sample (sum) of n cells with these parameters. We defined the following hierarchical model:

$$y = \sum^{n} p \cdot \mathcal{LN}(\mu_1, \sigma) + (1 - p) \cdot \mathcal{LN}(\mu_2, \sigma)$$

$$n \sim \text{truncBin}(2n, a = n - 5, b = n + 5, p = 0.5)$$

$$p \sim \text{Beta}(0.9, 0.9)$$

$$\mu_1, \mu_2 \sim \mathcal{N}(0, 100)$$

$$\sigma \sim \mathcal{HC}(25)$$

For the parameter p we chose a $\text{Beta}(0.9, 0.9)$ distribution, which is a relatively flat prior in the $[0; 1]$ interval where values near 0.5 are less probable than those towards 0 or 1. For μ_1 and μ_2, we defined a normal distribution with a large variance of 100. The variance here is crucial for the acceptance rate in the HARM algorithm. For the scale parameter σ, we chose the Half-Cauchy distribution with parameter 25, a least informative and proper prior which has been proven to work better than the inverse-Gamma distribution [Polson et al., 2012]. If we also try to find the paramter n (section 5.2.2), we model it by a truncated binomial distribution that is centered at the estimated n and allows n to be off by ± 5. Model definitions for the `LaplacesDemon` package can be found in listing 1.

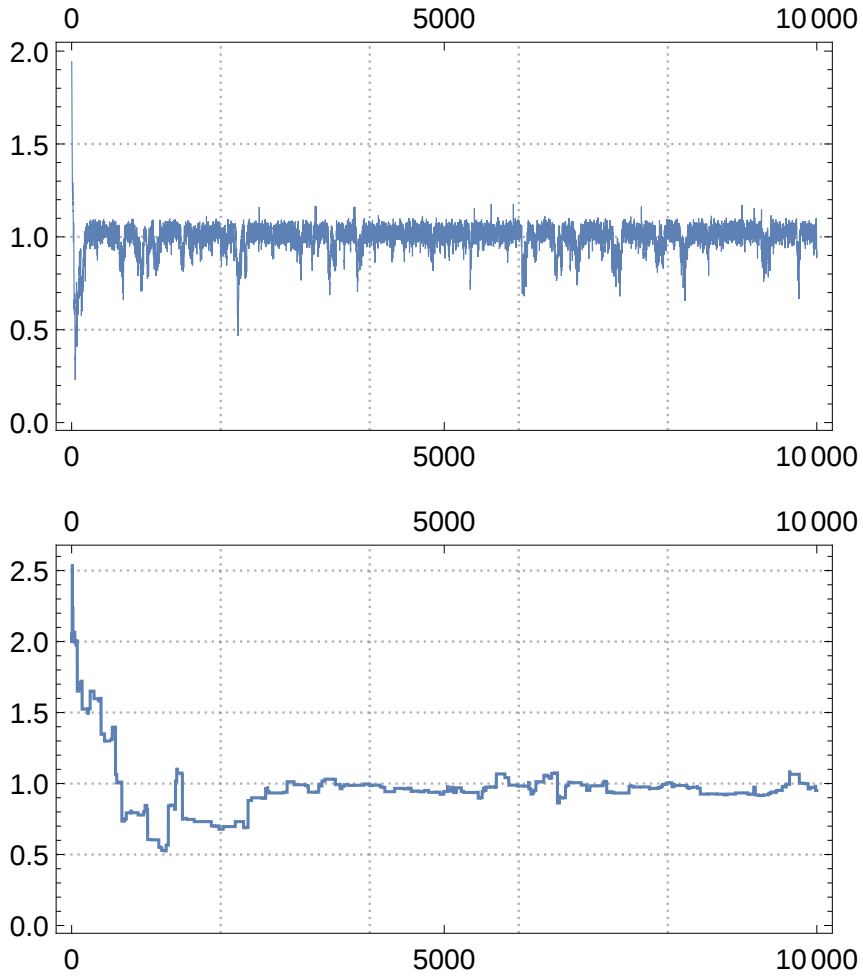


Figure 5.3: A comparison of the GG and HARM algorithms on the same dataset and the same priors. (*Top*) The first 10000 iterations of the GG algorithm. Convergence to the true value of 1 can be observed after a very small number of steps. (*Bottom*) The first 10000 iterations of the HARM algorithm. Because of its low acceptance rate, convergence is slower and the algorithm often stays at the same value for quite a long time.

5.2 RESULTS

5.2.1 *Required Sample Sizes*

To determine the minimum number of samples to achieve reasonable results (i.e., the true value being within the 95% credible interval) for each n (the number of cells), we defined sample sizes 10, 20, 30, 40, 50, 100, 200, 500, and 5000 and evaluated it for $n = 5$, 10, 20, 30, 40, 50, and 100 using the HARM sampler with 300000 iterations and a burn-in of 30000. Sums of more than 100 cells did not provide any meaningful results, no matter how large the sample size[3]. We calculated a

3 We tested up to a sample size of 200000.

95% credible interval and the sample mean (the parameter estimate) for every parameter p, μ_1, μ_2, and σ. We then took the exact same (generated) data and performed maximum likelihood estimation on them according to the method of Bajikar et al. [2014], implemented in the `stochprofML` package, where a 95% confidence interval around the estimated value is generated. All results are averages of 100 runs. Table 5.1 shows the required sample sizes for each n for MCMC and MLE methods to get results that are very close to the true values (within the 95% credible interval) for all four parameters p, μ_1, μ_2, and σ. Some parameters can even be estimated using fewer samples.

		number of cells					
	$n =$	5	10	20	30	50	100
Sample size	MCMC	10	30	50	100	500	5000
	MLE	20	30	50	100	500	5000

Table 5.1: The minimum sample size required for each n to get reasonable results for all four parameters p, μ_1, μ_2, σ (i.e., the true value being within the 95% interval).

See Appendix A for a detailed view of the credible intervals of every single parameter and sample size combination. We found that both methods, MCMC and MLE perform about equally well. Sometimes, the MCMC approach provides better results when the sample size is small. On the other hand, the confidence intervals for the MLE method are usually smaller than the credible intervals of MCMC, so this yields a more spot-on estimation if the problem is feasible. This is mainly because the MCMC samples have a greater variance, leading to greater intervals; see section 5.2.3 where this problem also appears.

5.2.2 *Estimating* n

Sometimes, the parameter n, the number of cells entering the sum, is not exactly clear. We investigate a case where we are *pretty sure* about the correct n, but allow our guess to be ± 5 off the true value. See section 5.1 for the prior definitions. In Figure 5.4, we show the first 30 iterations of the GG sampler trying to find the true value of $n = 10$, starting at $n = 14$. We see that the sampler converges incredibly fast (below 10 iterations) to the value of 10. We observed this throughout all combinations of sample size and n; the true value is always found after fewer than 20 iterations and stays there for the rest of the sampling procedure. The true n is even found when the estimation of the other parameters (e.g., if the sample size is too small) fails.

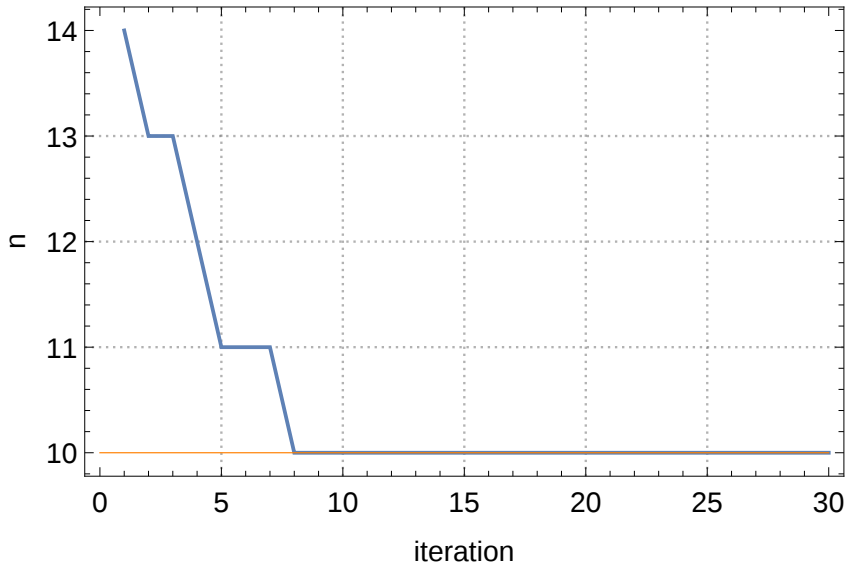


Figure 5.4: MCMC estimation of the parameter n using the GG sampler starting at $n = 14$. Convergence to the true value of 10 (orange line) happens after fewer than 10 iterations. The sampler stays there for the rest of the sampling procedure.

5.2.3 *Finding the Best* n

In a next step, we tried to find the best possible value for n for a fixed sample size of 100. According to Bajikar et al. [2014], samples with a value of $n = 10$ yield better results that one of single cell measurements ($n = 1$), when the sample size is limited. We now tried to find the best n when the sample size is feasible (i.e. equal or greater than what we found in section 5.2.1), by testing values for n from 1 to 30: we calculated the mean-squared-error (MSE) for each of these candidates and parameters, based on 100 calculations each, with both the MCMC and the MLE method (Figure 5.5, Figure 5.6, Figure 5.7 and Figure 5.8) on the same data. This was performed on all four parameter sets, as defined in Figure 5.1. MSE was calculated using the formula:

$$\text{MSE} = \text{Bias} + \text{Variance}^2.$$

For the first parameter set, the *ideal* parameters (Figure 5.5), we found that the MCMC approach outperforms the MLE, resulting in a smaller MSE for all values, except for μ_2 when n is closer to 30.

For the next parameter set, the more *realistic* ones (Figure 5.6), it is just the other way round; the MSE for the MLE method is smaller, the only exception being μ_2 again.

For the set with $p = 0.55$ (Figure 5.7), the MSE is smaller for the MCMC method for the parameters p and μ_2, whereas the MLE performs better on μ_1 and about equally as well on σ.

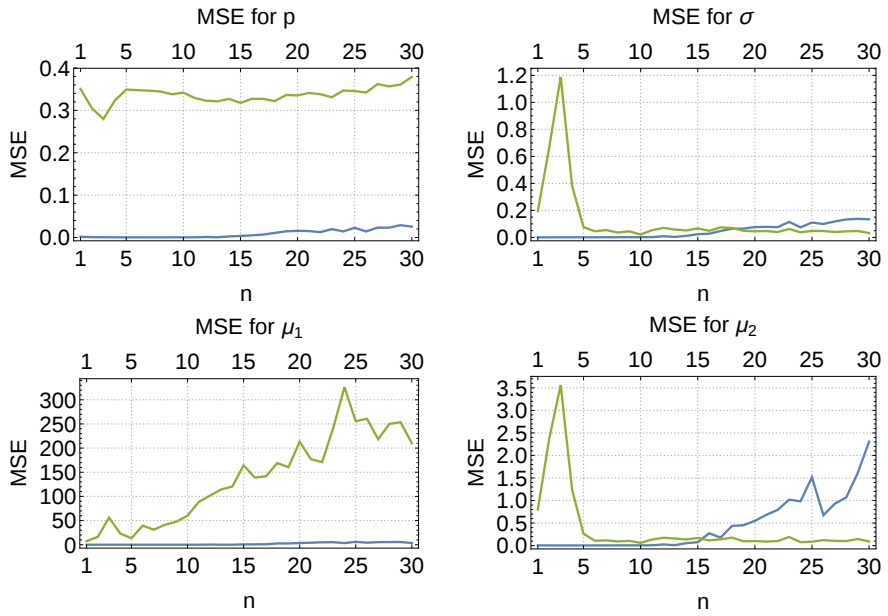


Figure 5.5: MSE for the parameter set $p = 0.8$, $\mu_1 = 1$, $\mu_2 = 6$, and $\sigma = 0.3$ for both MCMC (*blue*) and MLE (*green*).

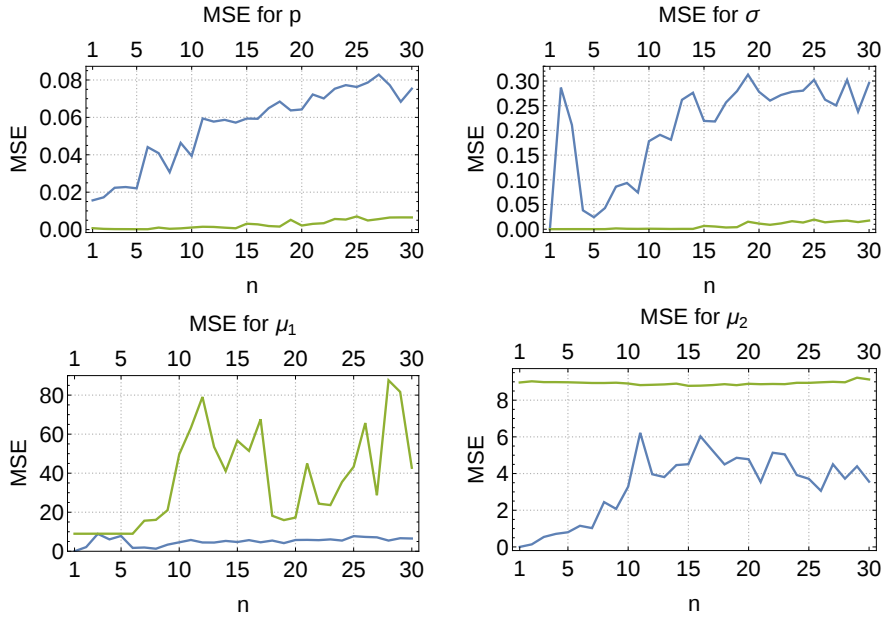


Figure 5.6: MSE for the parameter set $p = 0.225$, $\mu_1 = 0.5$, $\mu_2 = -2$, and $\sigma = 0.225$ for both MCMC (*blue*) and MLE (*green*).

The last parameter set, where σ is large (Figure 5.8), shows similar results: MLE performs better on p and μ_1, but worse on μ_2 and similar to MCMC on σ.

Obviously, it is strongly dependent on the values of p, μ_1, μ_2, and σ which method is to be preferred. We could not replicate the results given by Bajikar et al. [2014] where they noted a much higher MSE for $n = 1$ compared to $n = 10$ for the p parameter. But in our case, the

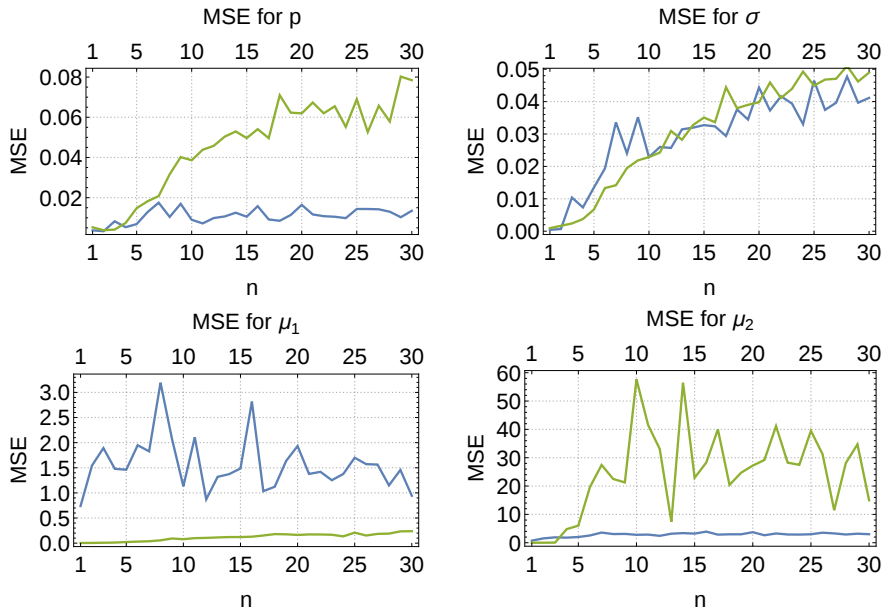


Figure 5.7: MSE for the parameter set $p = 0.55$, $\mu_1 = 0.5$, $\mu_2 = -2$, and $\sigma = 0.3$ for both MCMC (*blue*) and MLE (*green*).

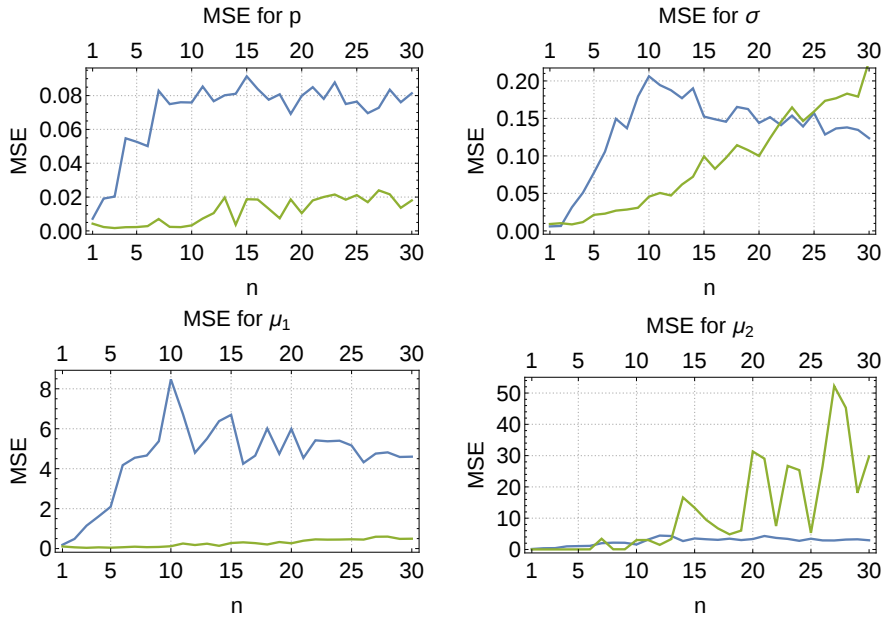


Figure 5.8: MSE for the parameter set $p = 0.2$, $\mu_1 = 0.5$, $\mu_2 = -2$, and $\sigma = 0.8$ for both MCMC (*blue*) and MLE (*green*).

sample size was much larger (100 compared with 10) so this is not a comparable situation. As the spike for $n = 3$ in Figure 5.5 shows, it is very much possible that, for certain parameter combinations, unpredictable behavior can occur.

We could not determine the *perfect* n, but values from 5 to 10 seem to work best. In general, a higher n usually leads to a higher MSE. Recall that $\text{MSE} = \text{Bias} + \text{Variance}^2$, so one reason for the higher MSE of

MCMC estimation may rely on the higher variance. Figure 5.9 shows boxplots of 100 estimates of parameter p for each n. Although both methods produce estimates of comparable quality regarding bias (as indicated by the mean and median lines), it can be seen that the variance in MCMC estimation is much larger, leading to a higher MSE.

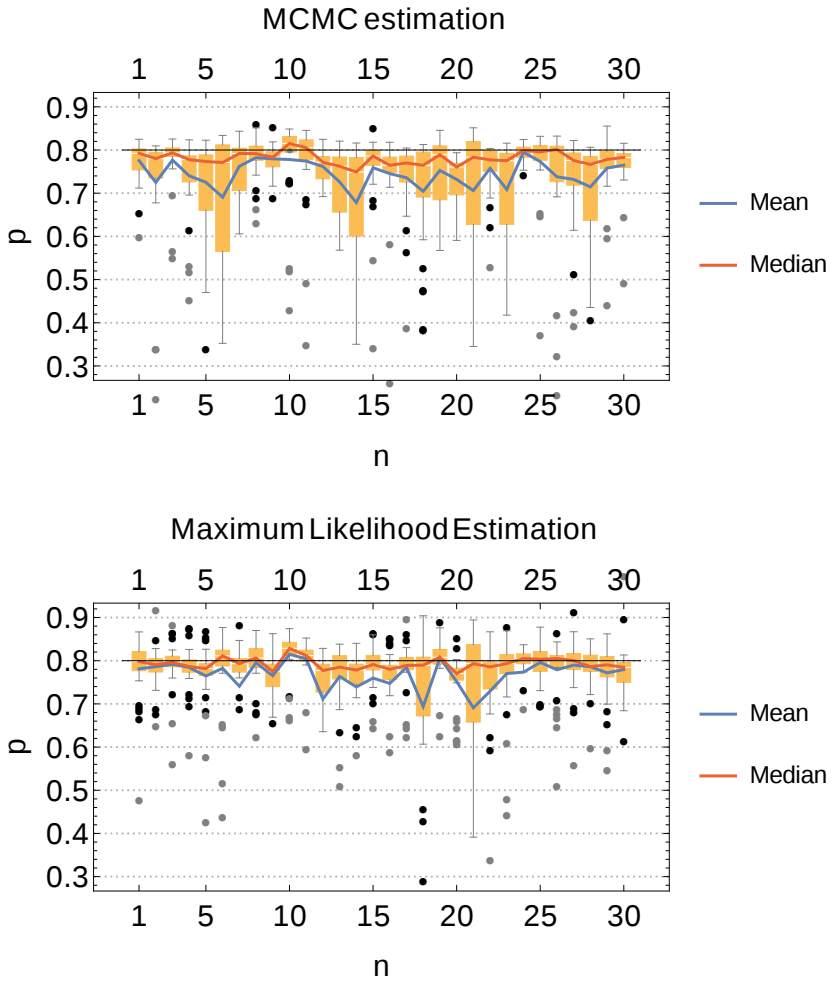


Figure 5.9: Boxplots of MCMC and maximum likelihood estimation of the parameter p for $n = 1, ..., 30$ of 100 runs each. The black line at $p = 0.8$ represents the true values to be estimated. Although the bias is low in both methods (indicated by the mean and median lines), the variance in MCMC estimation is much larger.

5.2.4 *Fixed* n

Sometimes it can be hard to get the right number of cells in a tissue sample, so we investigated how much influence the correct n has on the estimation. We generated data with $n = 10$ but performed the estimation with a fixed n of 5, 7, 9, 10, 11, 13, and 15. We used the

GG sampler with a sample size of 100 and averaged over 100 runs. Figure 5.10 shows the estimates and credible intervals for each parameter. Obviously, an n that is only off by 1 ($n = 9, 11$) has little impact on the estimation. Interestingly, $n = 5$ produces a worse estimate than $n = 15$ although both are off by 5. For the parameter p, we achieved a better estimation with $n = 11$ than with the true value of $n = 10$.

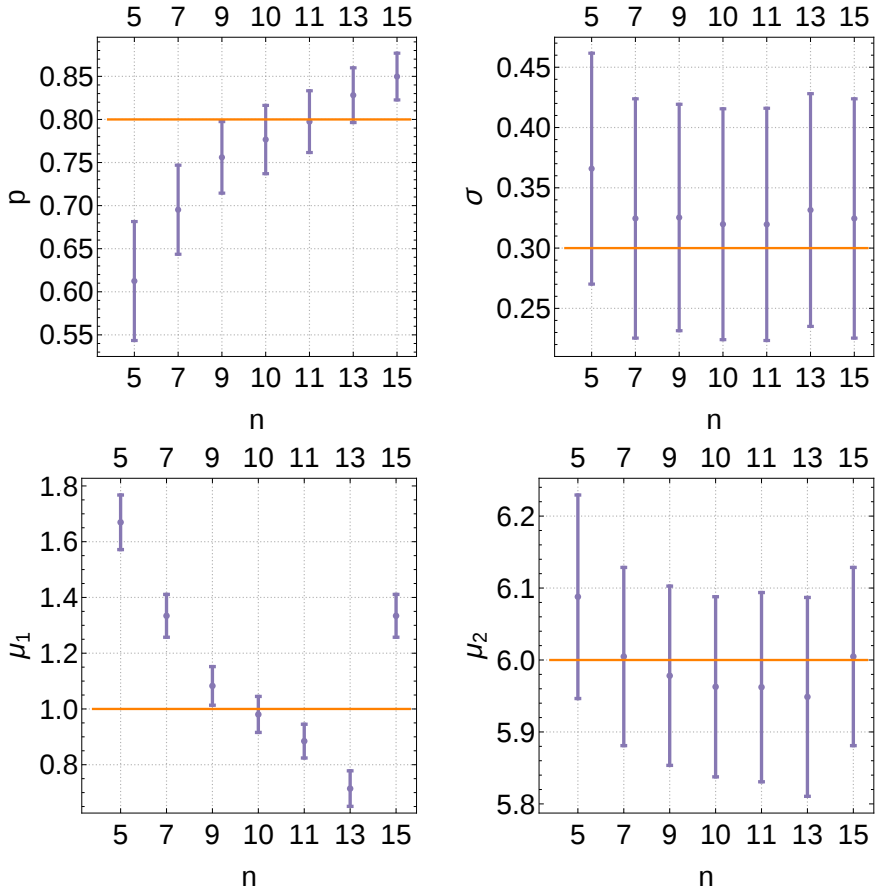


Figure 5.10: Intervals for MCMC estimation based on data with $n = 10$ but performed with different n. The orange line represents the true value. An n that is only off by 1 ($n = 9, 11$) has little impact on the estimation. For parameter p, we achieved better results with $n = 11$ than with the true value of $n = 10$.

5.2.5 *Estimation of Real Data*

We now perform MCMC and MLE on a real data set, consisting of expression of the detoxifying enzyme superoxide dismutase 2 (SOD2). It consists of 81 measurements and is part of the `stochprofML` package for `R`. MCMC estimation was performed using the GG sampler with 30000 iterations and a burn-in of 10000. Table 5.2 shows the results of both MCMC and MLE on this data set. It can be seen that the

results of MCMC and MLE are pretty similar; each method's estimate is within the interval of the other method (or, in the case of p and μ_2, pretty close to it). We were able to reproduce the exact estimation values as in Bajikar et al. [2014], where the quality of the MLE estimation is discussed..

parameter	MCMC	MLE
p	**0.37** [0.14; 0.64]	**0.23** [0.13; 0.34]
μ_1	**-1.72** [-2.68; -0.81]	**-1.06** [-1.68; -0.42]
μ_2	**-4.61** [-4.89; -3.36]	**-4.13** [-4.43; -3.84]
σ	**1.02** [0.65; 1.42]	**0.74** [0.54; 1.10]

Table 5.2: Comparison of the estimates for each parameter of MCMC and MLE on a real data set, the SOD2 data. Values in brackets represent the 95% credible interval (MCMC) and the 95% confidence interval (MLE) respectively. All estimates are very similar in both methods, being within or very close to each method's interval.

5.3 BIVARIATE LOGNORMAL ESTIMATION

Genes are often correlated, i.e., the proportion of variance that two traits share due to genetic causes is very high. Some traits, e.g., height and eye color, are likely to depend on different genes, whereas others, e.g., height and weight, are likely to depend on many overlapping genes [Kim, 2009]. When modeling two genes by a mixture of bivariate lognormals, each gene occupies one of two regulatory states: a transcript at a low level with log-mean μ_2 and one at a high level with μ_1; both share the same log-SD σ. The probability of drawing a single cell from the high regulatory states of both genes (μ_{11} and μ_{12}) is characterized by the parameter p. Thus, the probability of drawing a single cell from the low regulatory states (μ_{21} and μ_{22}) is characterized by $(1-p)$. ρ expresses the correlation between each of these states.

Correlation is not taken into account when estimating the parameters by the maximum likelihood method of Bajikar et al. [2014], where independence (in the case of lognormals that is $\rho = 0$) is assumed. We now study two test cases, one with two genes that share a low correlation ($\rho = 0.2$, figure 5.11) and two highly correlated ones ($\rho = 0.8$, figure 5.12). In this case, the single-cell expression follows the PDF:

$$f_{mix} = p \cdot f_1 + (1-p) \cdot f_2,$$

where f_1 and f_2 are defined as in (7). Again (see section 5.1), the i-th random sample of transcript g, $g \in \{1,2\}$, is the sum of n measurements ($n = 10$ in our case) of independent single-cell expressions:

$$Y_i^{(g)} = \sum_i^n X_{ij}^{(g)},$$

with $X_{ij}^{(g)}$ representing the expression of transcript g in the j-th cell of the i-th random sample. Then, the random sample $Y_i^{(g)}$, $g \in \{1,2\}$, characterizing the n-cell mixture has the PDF:

$$f_n(y_1, y_2 | p, \mu_1, \mu_2, \sigma, \rho) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} f_{(j,n-j)}(y_1, y_2).$$

As in section 5.1, the binomial part $\binom{n}{j} p^j (1-p)^{n-j}$ represents the selection of cells with probabilities p and $1-p$ respectively. $f_{(j,n-j)}(y_1, y_2)$ is the density of a n-cell sum $\sum_{i=1}^{n}$ of independent random variables drawn from the mixture model:

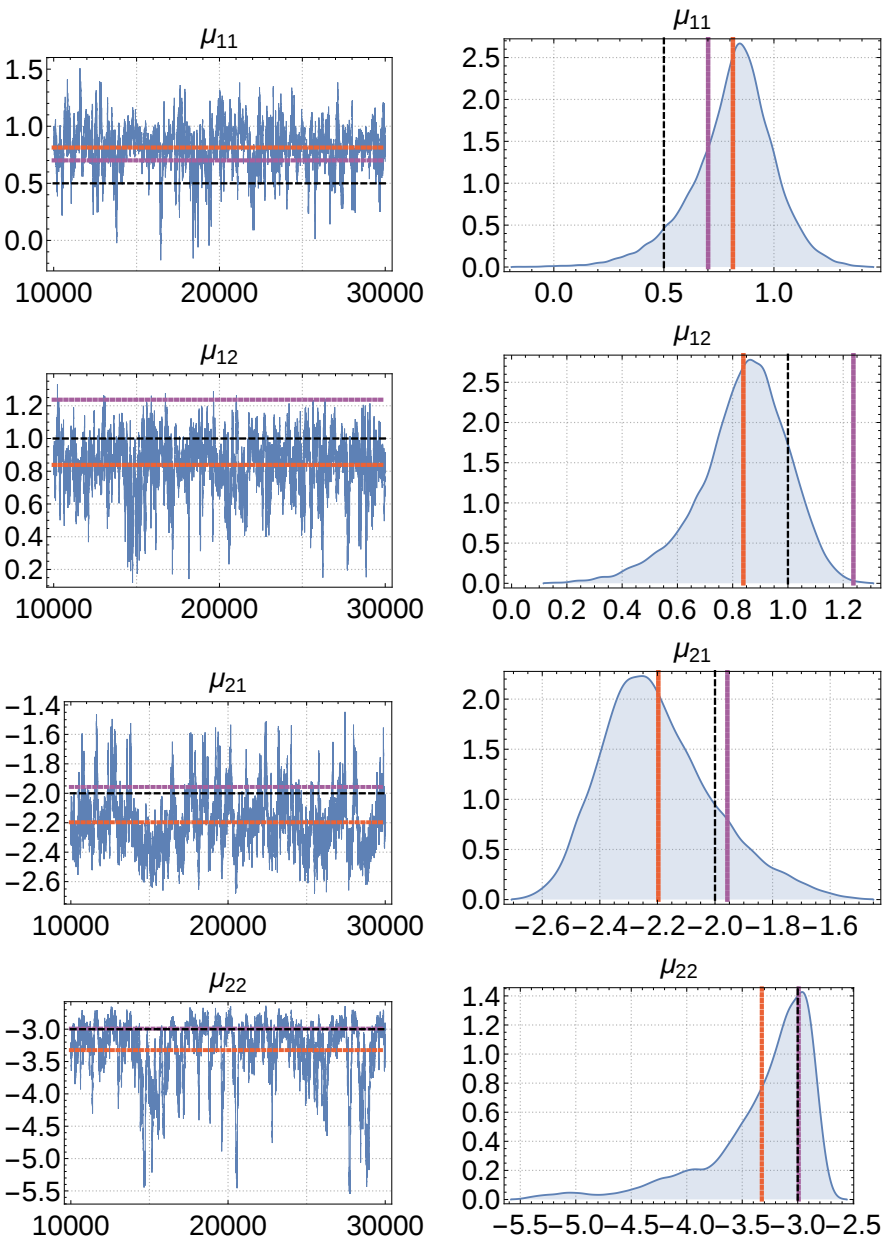
$$Z_c = \begin{cases} \mathcal{LN}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}) & \text{if} \quad 1 \leqslant c \leqslant j \\ \mathcal{LN}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma}) & \text{if} \quad j < c \leqslant n \end{cases},$$

where $\boldsymbol{\mu_1} = \begin{pmatrix} \mu_{11} \\ \mu_{12} \end{pmatrix}$, $\boldsymbol{\mu_2} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$.
The log-likelihood function for the model parameters given k random n-cell samples is

$$\ell(p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma, \rho) = \sum_{i=1}^{k} \log\left[f_n\left(x_i, y_i | p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma, \rho\right)\right].$$

For the MCMC model, we used the same priors as in section 5.1 with all μs following a normal distribution $\mathcal{N}(0, 100)$. We defined $\rho \sim \text{Beta}(2, 2)$ where values around 0.5 are more likely than those toward 0 or 1. Model definitions for the `LaplacesDemon` package can be found in listing 2.

In Figure 5.11, we see that, for a low ρ ($\rho = 0.2$ here), taking the correlation into account does not improve the overall results. The MLE, where $\rho = 0$ is assumed, yields similar results to the MCMC estimation where ρ is also estimated. On the other hand, we observe that, for a higher ρ (we used $\rho = 0.8$), the estimation of all parameters improves significantly when using the bivariate model of the MCMC estimation, compared with MLE (Figure 5.12).


μ_{11}
μ_{11}
μ_{12}
μ_{12}
μ_{21}
μ_{21}
μ_{22}
μ_{22}

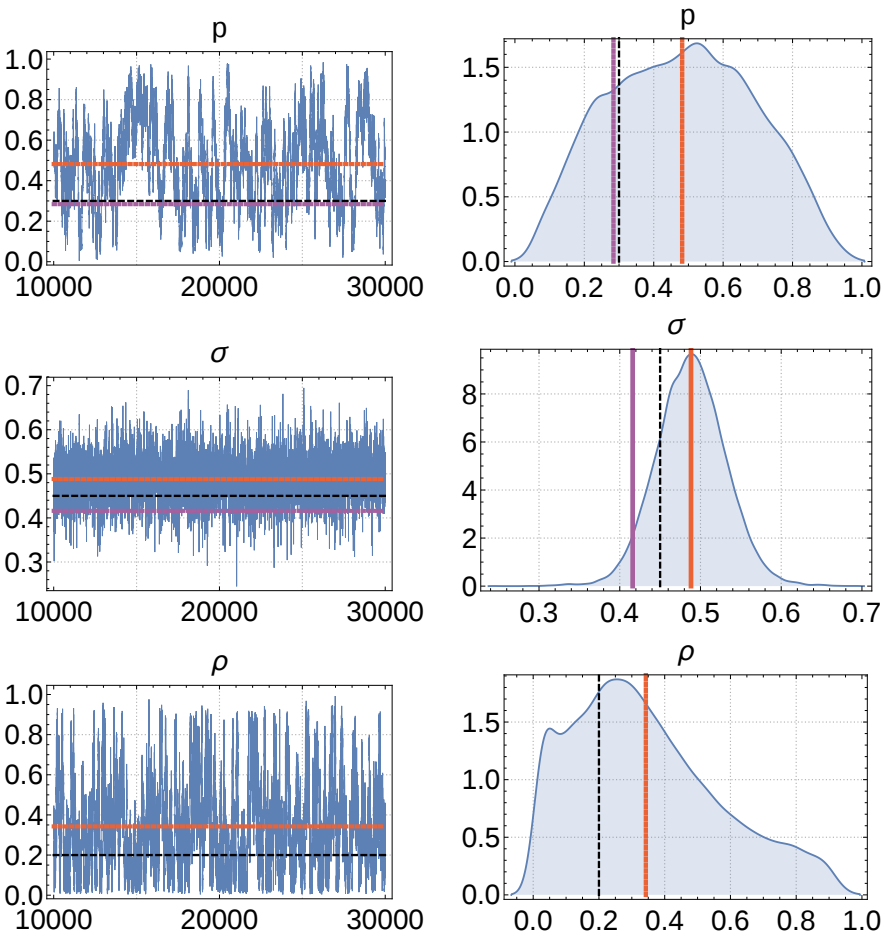
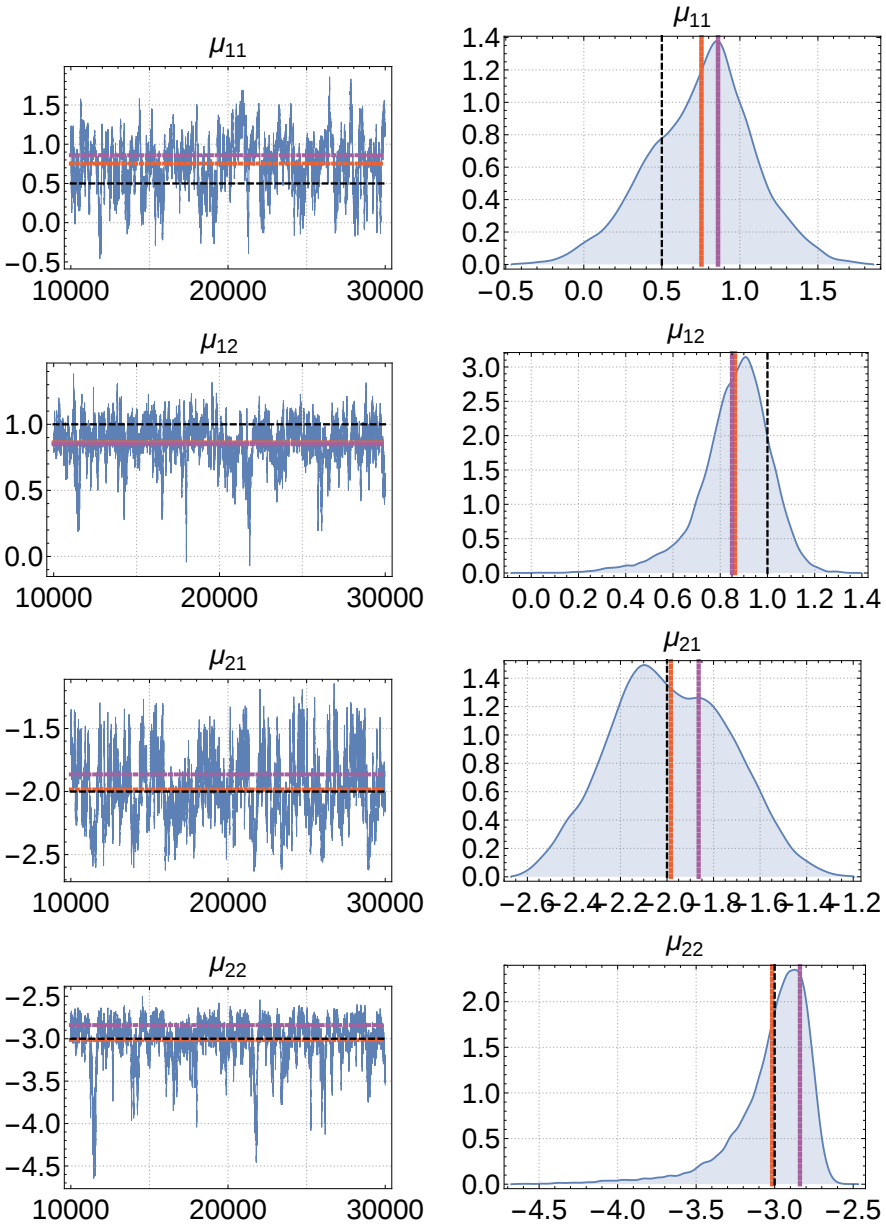


Figure 5.11: Line- and density plots of MCMC estimation using the GG sampler for all parameters $p, \mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}, \sigma$, and ρ. The *orange* line represents the MCMC estimate, the *purple* line the estimate of the MLE. The black dashed line marks the true value. It can observed that, for a small correlation between two genes ($\rho = 0.2$ here), MLE yields similar results to MCMC even though independence ($\rho = 0$) is assumed in MLE .


μ_{11}
μ_{11}
μ_{12}
μ_{12}
μ_{21}
μ_{21}
μ_{22}
μ_{22}
10000
20000
30000

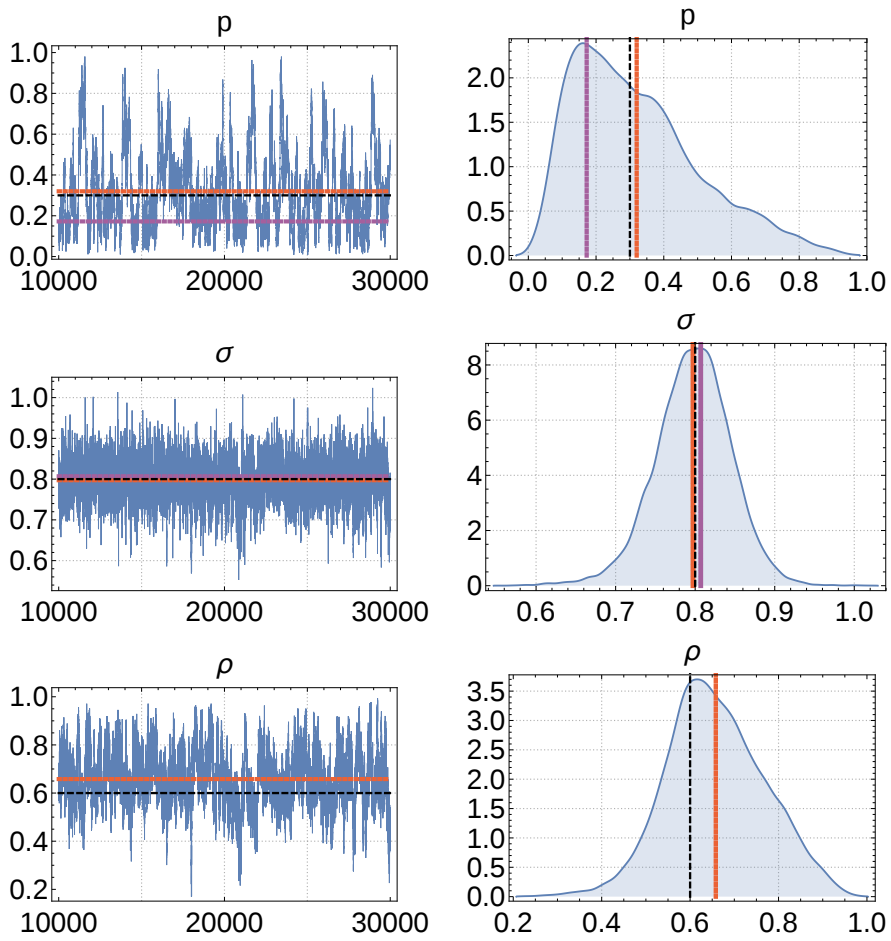


Figure 5.12: Line- and density plots of MCMC estimation using the GG sampler for all parameters $p, \mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}, \sigma$, and ρ. The *orange* line represents the MCMC estimate, the *purple* line the estimate of the MLE. The black dashed line marks the true value. It can be observed that, for a high correlation between two genes ($\rho = 0.8$ here), MCMC estimation, where ρ is estimated, yields better results than MLE where independence ($\rho = 0$) is assumed.

Correlation Estimation with Real Data

We now estimate the expression of two genes using real data. We analyze three data sets, consisting of only 16 measurements each. We picked the two most correlated genes and performed MCMC and MLE on them. MCMC was performed using the GG sampler with 30000 iterations and a burn-in of 10000. Figure 5.13 shows the correlation matrix heatmap of all genes in the first data set; after that, we picked gene 94 and gene 10. Table 5.3 shows the results of both MCMC and MLE for each parameter for this first data set. Because of the small sample size, these results have to be treated with caution. Especially in the case of μ_{22}, we did not achieve convergence in the

MCMC sampler, but neither does MLE achieve reasonable results for this parameter. Overall, the estimates of MCMC and MLE are quite different.

Table 5.4 shows the results of the second dataset. Again, we performed MCMC and MLE on the two most correlated genes, genes 903 and 642. In this case, we achieved pretty good convergence on all parameters (reaching a stationary distribution), and the results of MCMC and MLE are pretty close, i.e., both method's estimates are within each other's intervals and even differ only by a decimal fraction.

For the last data set, we picked genes 363 and 286 as the moste correlated ones. Estimation results are not as similar as those for data set 2, but again they are pretty close for MCMC and MLE, see Table 5.5.

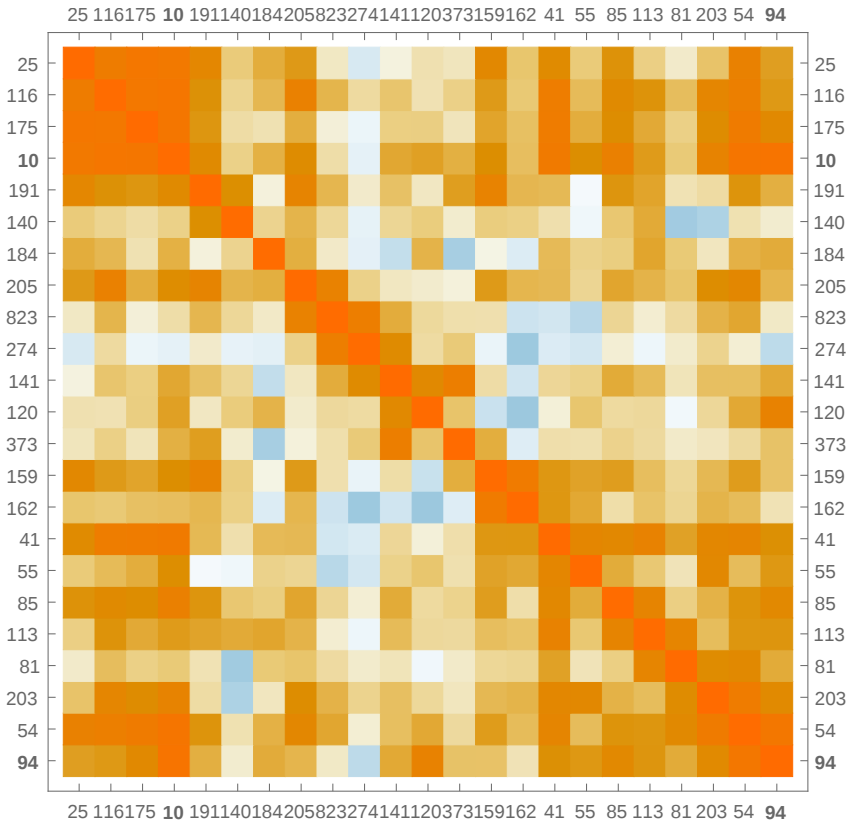


Figure 5.13: Correlation heatmap for all 23 genes in a real data set, data set 1. Darker orange squares represent a higher positive correlation, darker blue squares a higher negative correlation. The two genes with the highest correlation (gene 94 and gene 10) are marked in bold.

parameter	MCMC	MLE
p	**0.47** [0.28; 0.67]	**0.37** [0.32; 0.42]
μ_{11}	**1.88** [1.37; 2.37]	**-0.12** [-0.49; 0.24]
μ_{21}	**1.27** [0.82; 1.87]	**1.67** [1.63; 1.71]
μ_{12}	**0.54** [0.07; 1.04]	**1.91** [1.84; 1.99]
μ_{22}	**-4.58** [-11.73; -1.28]	**-22.05** [-23.01; -21.11]
σ	**0.48** [0.27; 0.73]	**0.12** [0.08; 0.17]
ρ	**0.48** [0.13; 0.88]	NA

Table 5.3: Comparison of the estimates of MCMC and MLE on the two highest correlated genes in a real data set, data set 1. Bold values represent the estimated value, values in brackets represent the 95% credible interval (MCMC) and the 95% confidence interval (MLE) respectively. Because of the small sample size, only 16 measurements, the results are not reliable. MCMC and MLE results are quite different.

parameter	MCMC	MLE
p	**0.032** [0.008; 0.11]	**0.006** [0.001; 0.025]
μ_{11}	**2.23** [0.78; 3.58]	**2.28** [0.55; 4.02]
μ_{21}	**1.32** [0.82; 1.87]	**1.32** [-0.4; 3.05]
μ_{12}	**-3.41** [-3.58; -3.25]	**-3.37** [-3.56; -3.19]
μ_{22}	**-3.68** [-3.84; -3.52]	**-3.37** [-3.83; -3.46]
σ	**0.86** [0.73; 1.0]	**0.81** [0.67; 0.98]
ρ	**0.27** [0.06; 0.50]	NA

Table 5.4: Comparison of the estimates of MCMC and MLE on the two highest correlated genes in a real data set, data set 2. Bold values represent the estimated value, values in brackets represent the 95% credible interval (MCMC) and the 95% confidence interval (MLE) respectively. Despite the small sample size, only 16 measurements, we achieved good convergence for the MCMC samples (reaching a stationary distribution), and the results of MCMC and MLE are pretty close.

parameter	MCMC	MLE
p	**0.13** [0.05; 0.22]	**0.09** [0.05; 0.15]
μ_{11}	**-1.71** [-2.46; -1.20]	**-1.33** [-1.87; -0.80]
μ_{21}	**-0.54** [-1.25; -0.03]	**-0.27** [-0.67; 0.12]
μ_{12}	**-3.22** [-3.42; -3.06]	**-3.12** [-3.28; -2.96]
μ_{22}	**-3.26** [-3.57; -3.05]	**-3.16** [-3.33; -2.99]
σ	**0.63** [0.39; 0.88]	**0.46** [0.28; 0.76]
ρ	**0.62** [0.29; 0.91]	NA

Table 5.5: Comparison of the estimates of MCMC and MLE on the two highest correlated genes in a real data set, data set 3. Bold values represent the estimated value, values in brackets represent the 95% credible interval (MCMC) and the 95% confidence interval (MLE), respectively. For this third data set, we again achieved good convergence for the MCMC samples; the results of MCMC and MLE are still quite close.

5.4 MULTIVARIATE LOGNORMAL ESTIMATION

In the last section, we estimated mixture models of two bivariate lognormal distributions. To assess the correlation of more than two genes in one cell, a next logical step is to generalize to multivariate lognormal distributions. That is, we estimate a mixture

$$f_{mix} = p \cdot f_1 + (1-p) \cdot f_2,$$

where f_1 and f_2 are defined as

$$f_\nu(\mathbf{y}) = \frac{1}{\prod_{i=1}^{m} y_i (2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y}-\boldsymbol{\mu})\}, \quad \mathbf{y} > 0, \nu \in \{1,2\}.$$

$|\Sigma|$ denotes the determinant of the covariance matrix Σ, $\mathbf{y} = (y_1, ..., y_m)$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)$. The random sample Y_i and the likelihood are modeled analogous to section 5.3, as are the priors for MCMC estimation.

We then used the three data sets from section 5.3 for estimation. Each set contains 16 measurements of 23, 10, and 13 genes respectively. So we had to estimate three mixtures f_{mix} with $m = 23$, $m = 10$, and $m = 13$.

As in section 5.3, we used the GG sampler with 30000 iterations and a burn-in of 10000. However, we were not able to produce any meaningful results; the sampler did not converge at all, not reaching a stationary distribution. Figure 5.14 shows the estimation steps for parameter p for data set 1. Even after raising the iterations to 500000, which takes almost a day of computation on a modern computer, the sampler did not converge. Switching to the HARM sampler and 5 Mio. iterations did not change that. This may be because of the small sample size and the huge number of parameters: for $m = 23$, there are 49 parameters to be estimated. But even with simulated data and a sample size of 500, convergence did not improve for the m's specified above.

However, based on simulated data, we were able to determine the required amount of samples to have the sampler converge. When m is becoming larger, estimation is getting very difficult and we could not get reasonable results, i.e., the sampler does not converge to a stationary distribution when m is greater than 10. Table 5.6 shows the required sample sizes and the amount of parameters that did not converge for $m = 2, 3, 4, 5, 6, 7, 8, 9$ and 10. The values in this table are based on experience of 50 runs for each m. It is possible to achieve convergence with lower samples sizes in some cases. Parameters that did not converge are only the μs. When $m > 10$, parameters p and σ are unlikely to converge and thus estimation becomes infeasible (MCMC samples look like the one in Figure 5.14).

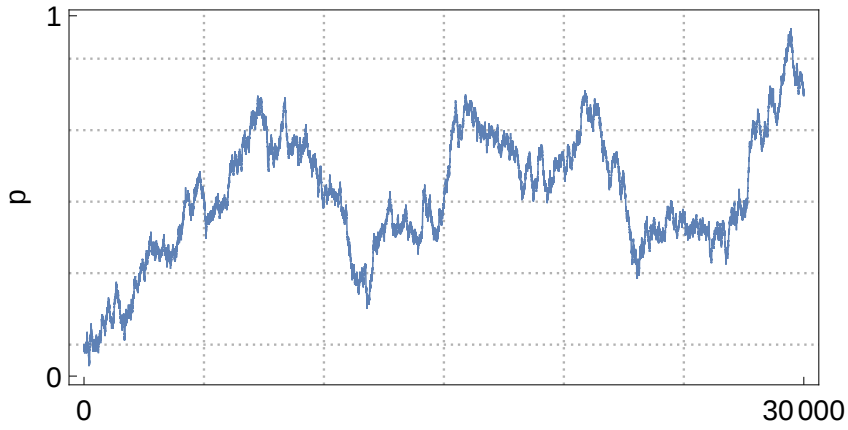


Figure 5.14: Lineplot of a multivariate estimation ($m = 23$) of parameter p of data set 1, which contains 16 measurements of 23 genes in a single cell. No covergence is achieved and thus no estimation of p can be produced.

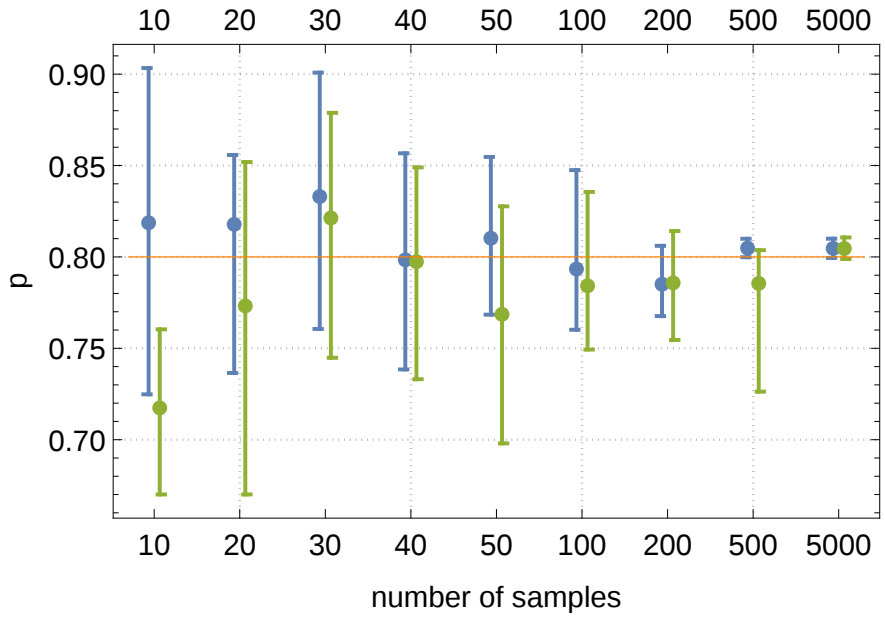
m	required sample size	# of parameters that did not converge
2	50	0
3	100	0
4	200	1-3
5	200	2-4
6	300	4-6
7	300	5-9
8	300	12-14
9	500	11-15
10	1000	11-19

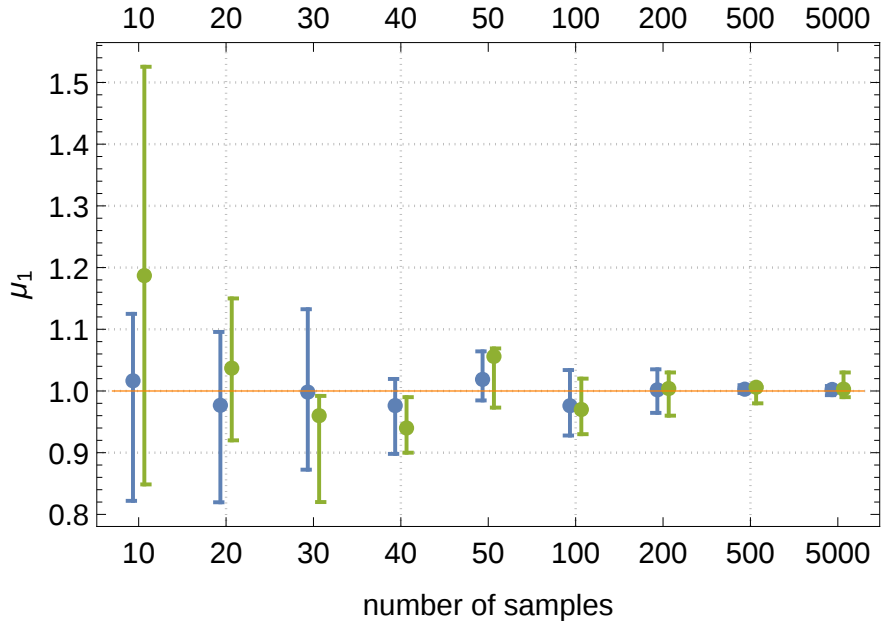
Table 5.6: Required sample sizes and number of paramters not converging when estimating a mixture of m-dimensional multivariate lognormals using the GG sampler. The parameters not converging are always some of the μs. The values in this table are based on experience of 50 runs for each m. It is possible to achieve convergence with lower samples sizes in some cases. Estimation becomes infeasible when $m > 10$.

6 CONCLUSION

In this thesis, we have presented an MCMC approach to the problem of *Stochastic Profiling*. We have shown that this approach enables accurate single-cell expression characteristics for a variety of n-cell measurements. We found the required sample sizes for the estimation to be feasible for a variety of different ns. If the true n is unknown, i.e., one does not know how many cells the measurement consists of, it can be easily estimated. When comparing the MCMC method with MLE, we have shown that it strongly depends on the parameter sets which method yields better results. We have further shown that the estimation can be improved when modeling a bivariate lognormal mixture, where correlation is taken into account as well. However, this cannot be extended to the general case of an m-dimensional multivariate lognormal mixture, as we found parameters for $m > 10$ impossible to estimate.

APPENDIX


10 20 30 40 50 100 200 500 5000
0.90
0.85
0.80
0.75
0.70
p
10 20 30 40 50 100 200 500 5000
number of samples


10 20 30 40 50 100 200 500 5000
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
μ_1
10 20 30 40 50 100 200 500 5000
number of samples

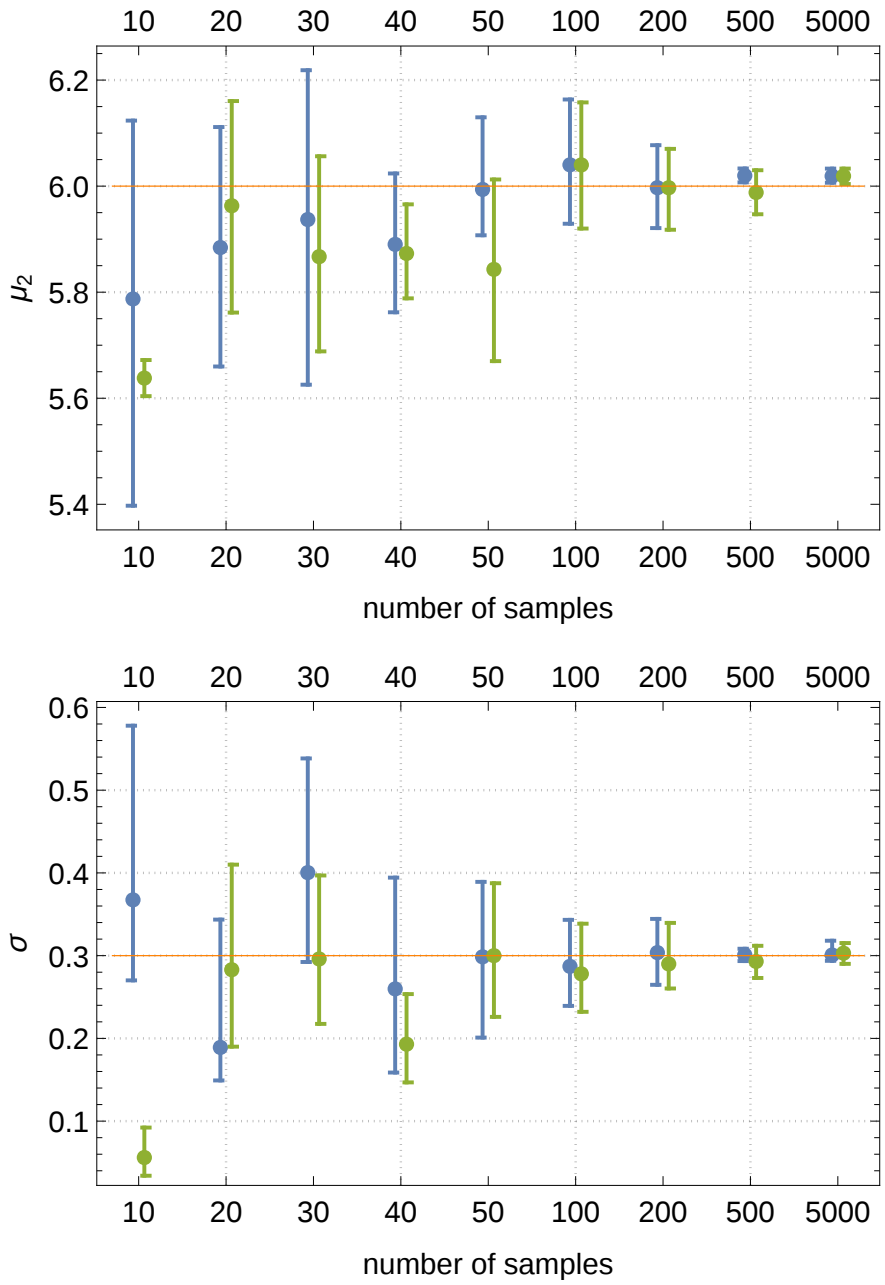
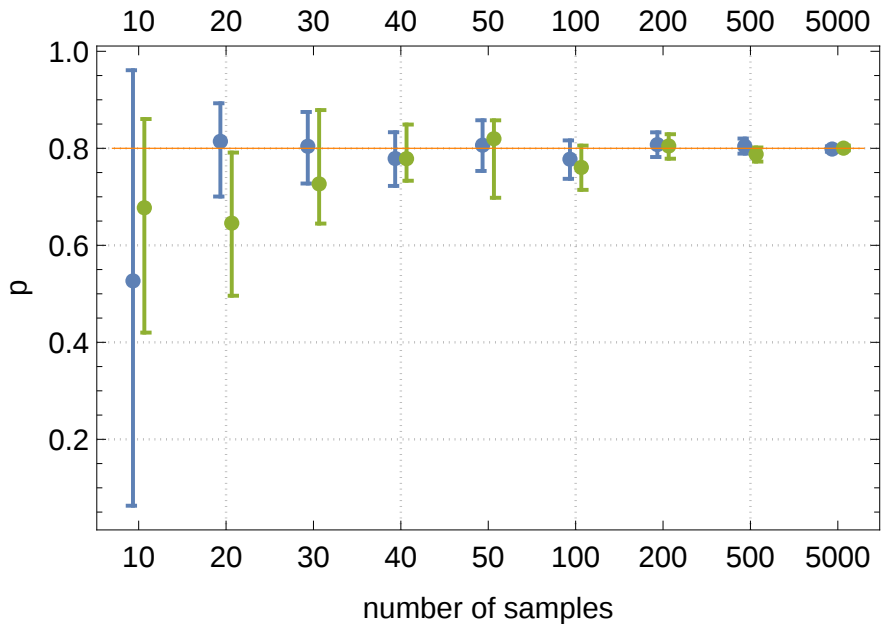
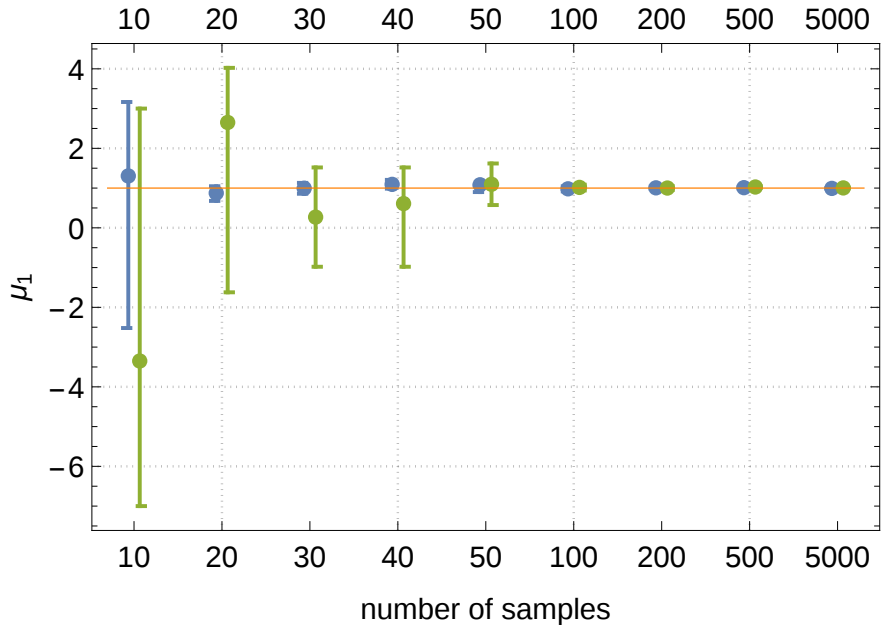


Figure A.1: Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 5$. The orange line represents the true value.


10 20 30 40 50 100 200 500 5000
1.0
0.8
0.6
0.4
0.2
p
10 20 30 40 50 100 200 500 5000
number of samples


10 20 30 40 50 100 200 500 5000
4
2
0
−2
−4
−6
μ_1
10 20 30 40 50 100 200 500 5000
number of samples

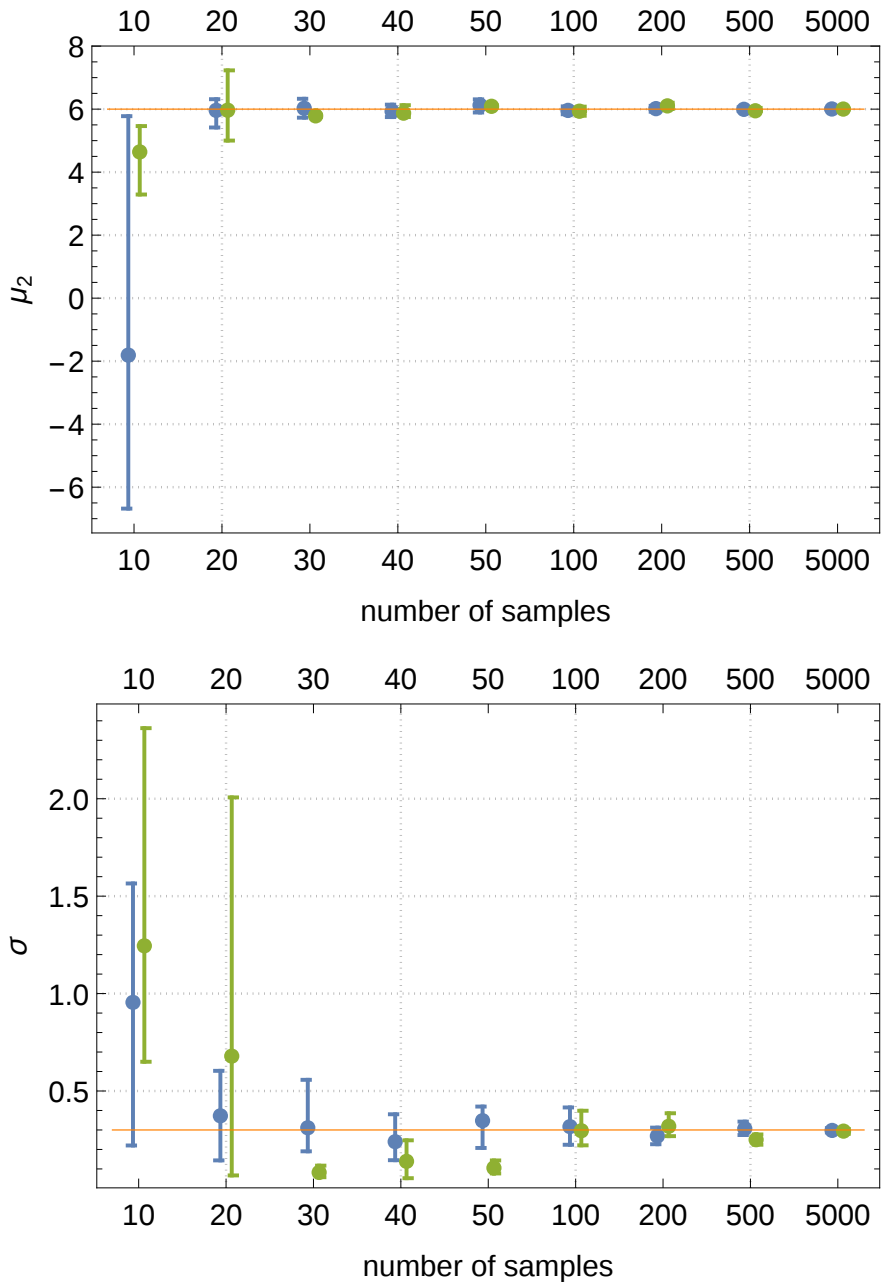
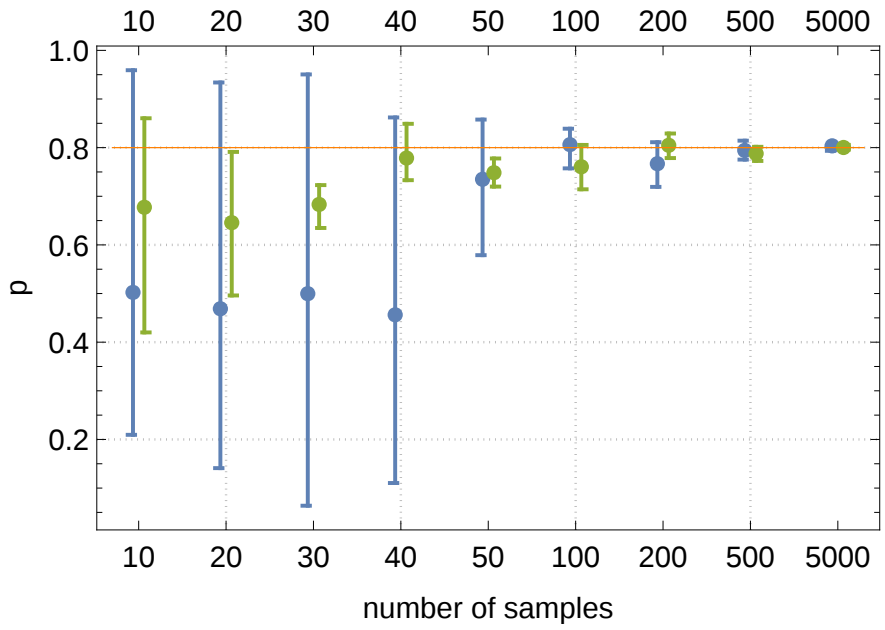
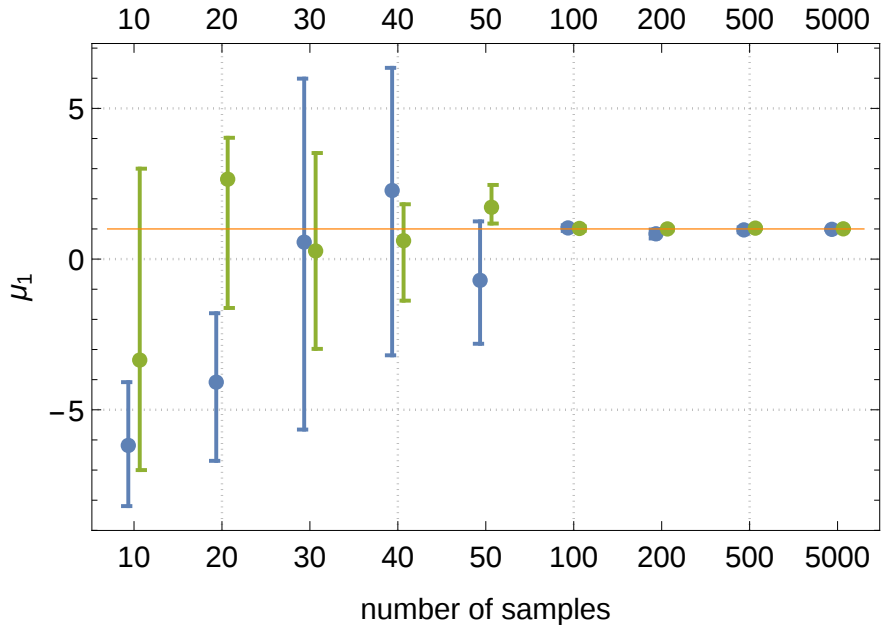


Figure A.2: Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 10$. The orange line represents the true value.


10
20
30
40
50
100
200
500
5000
1.0
0.8
0.6
0.4
0.2
p
number of samples


10
20
30
40
50
100
200
500
5000
5
0
−5
μ_1
number of samples

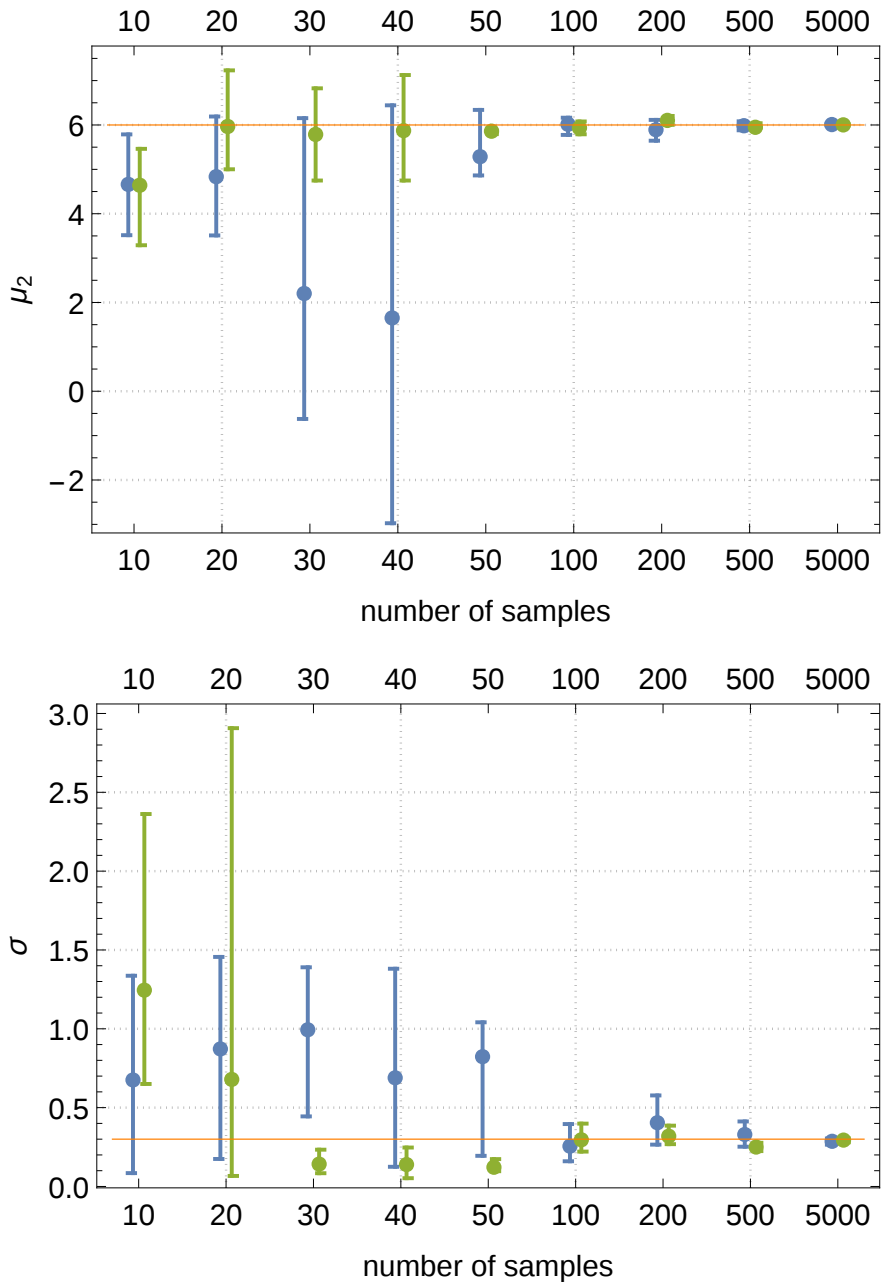
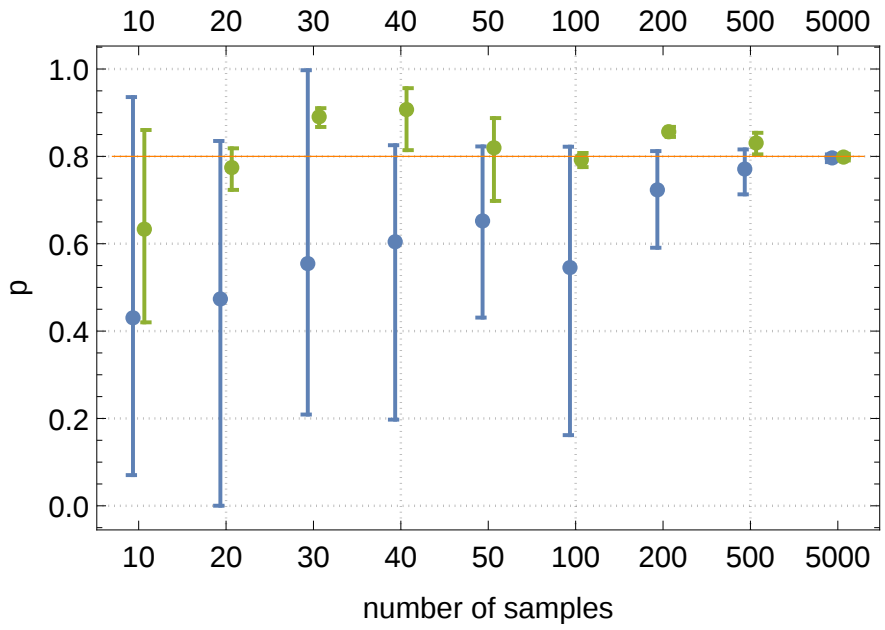
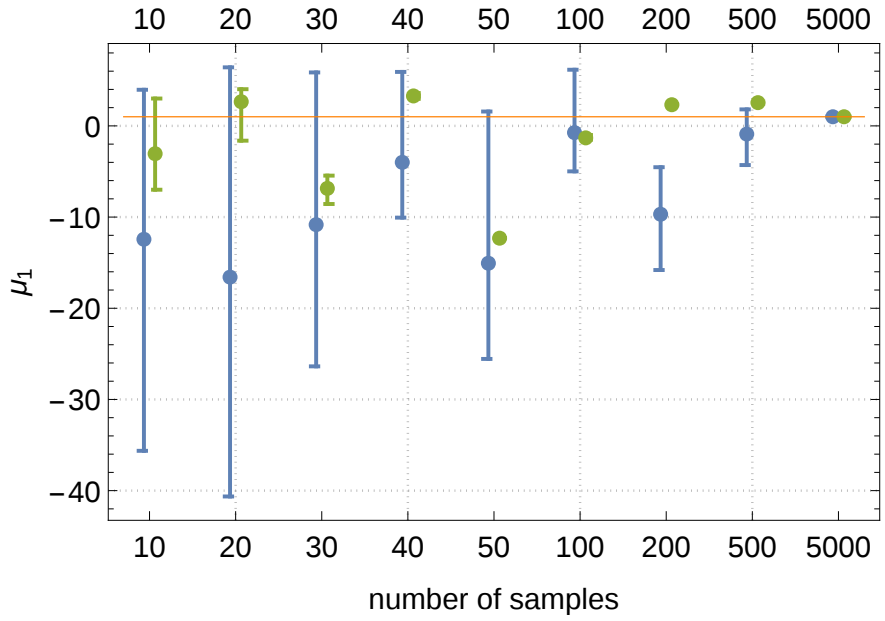


Figure A.3: Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 20$. The orange line represents the true value.


10 20 30 40 50 100 200 500 5000
1.0
0.8
0.6
0.4
0.2
0.0
p
10 20 30 40 50 100 200 500 5000
number of samples


10 20 30 40 50 100 200 500 5000
0
−10
−20
−30
−40
μ_1
10 20 30 40 50 100 200 500 5000
number of samples

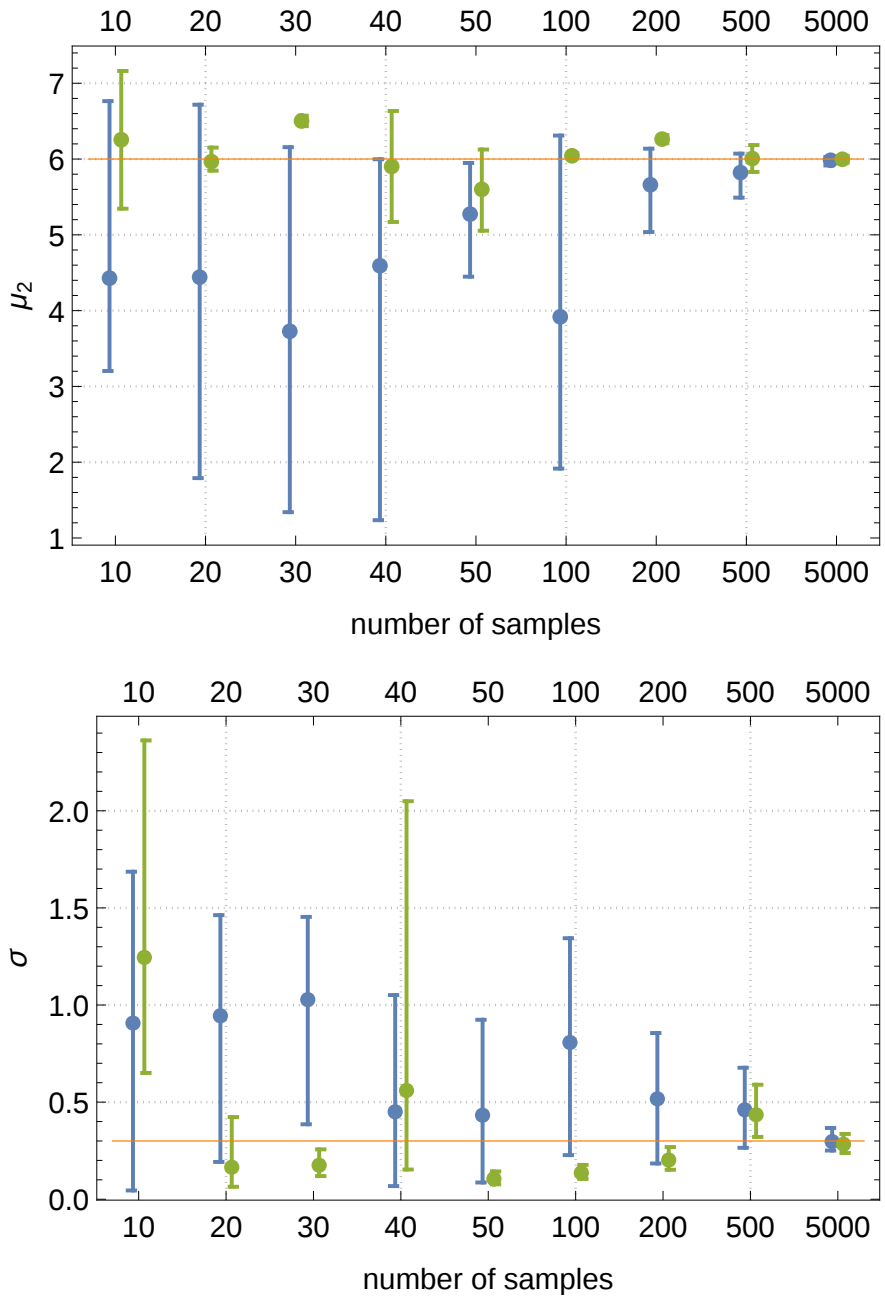
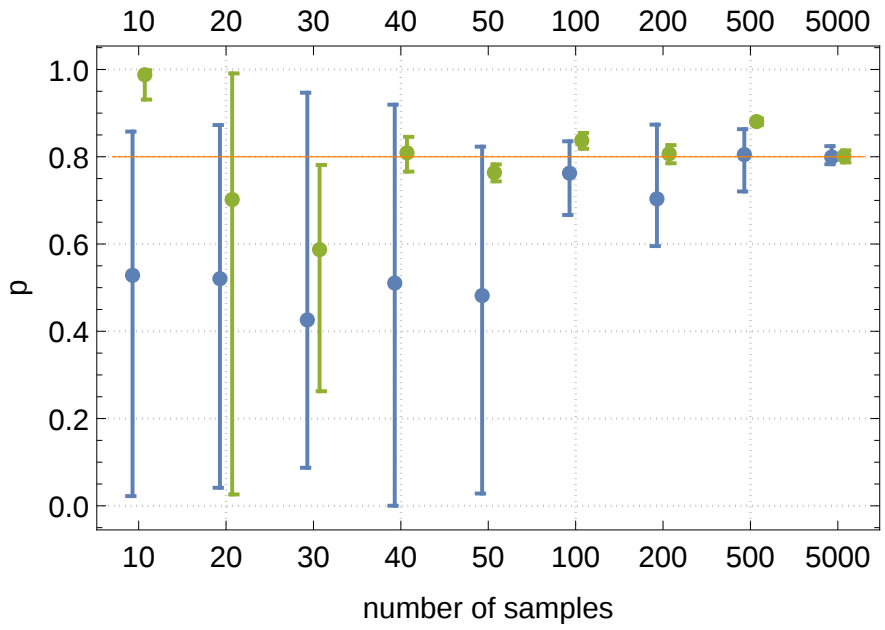
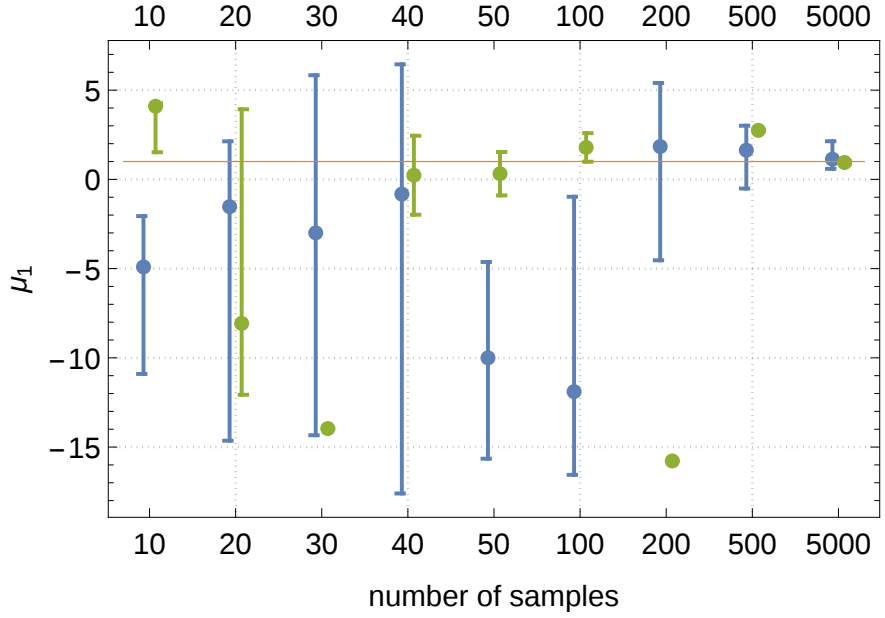


Figure A.4: Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 30$. The orange line represents the true value.


10 20 30 40 50 100 200 500 5000
1.0
0.8
0.6
0.4
0.2
0.0
p
10 20 30 40 50 100 200 500 5000
number of samples


10 20 30 40 50 100 200 500 5000
5
0
−5
−10
−15
μ_1
10 20 30 40 50 100 200 500 5000
number of samples

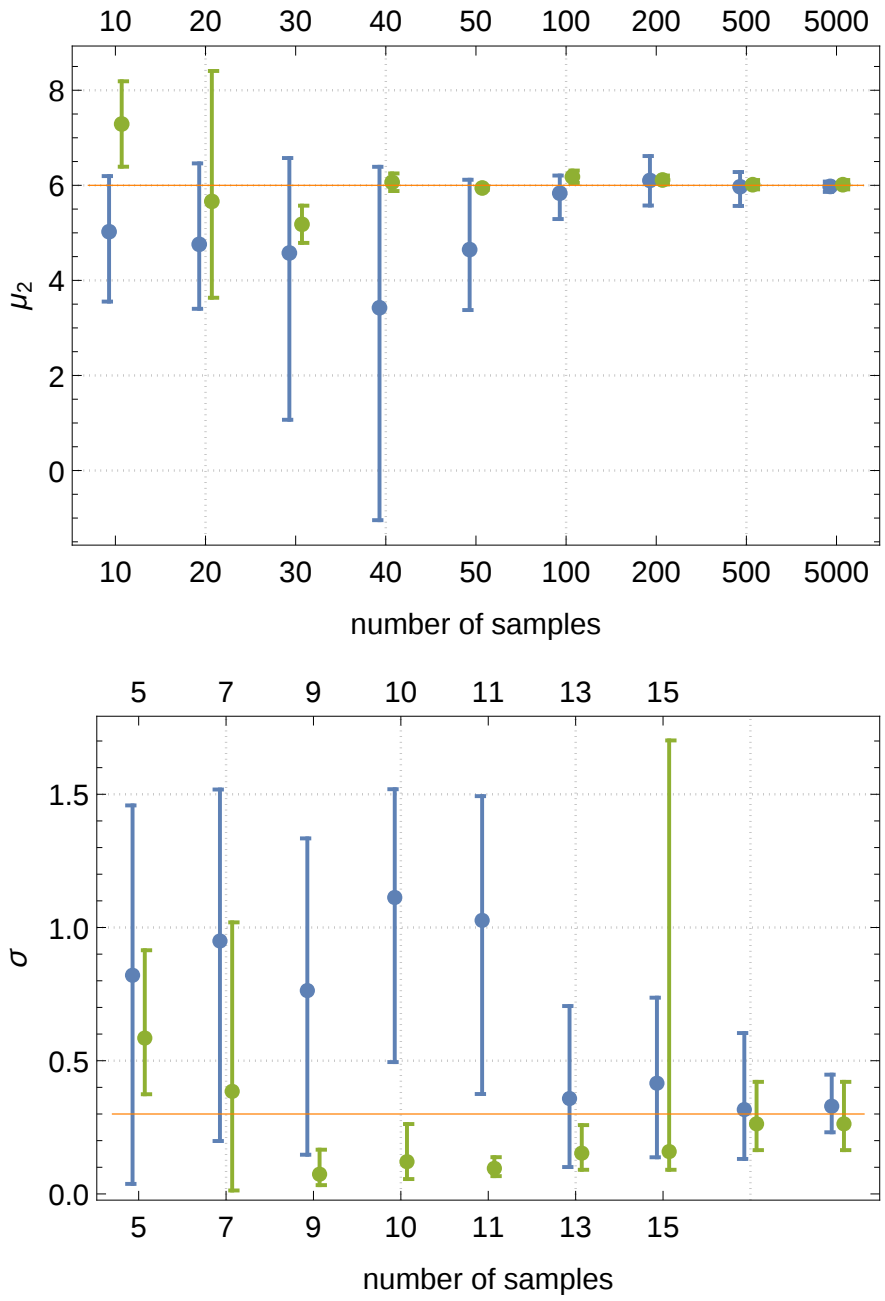
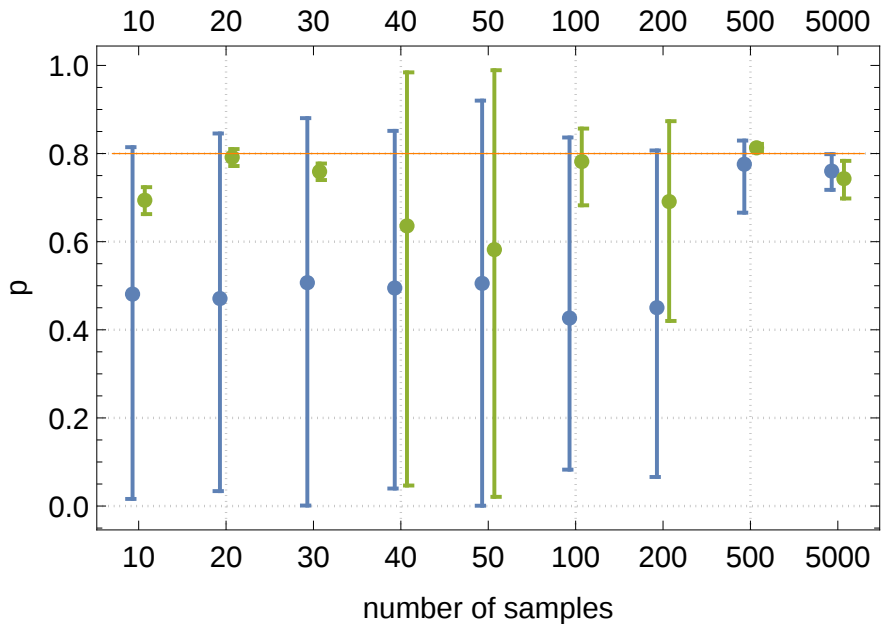
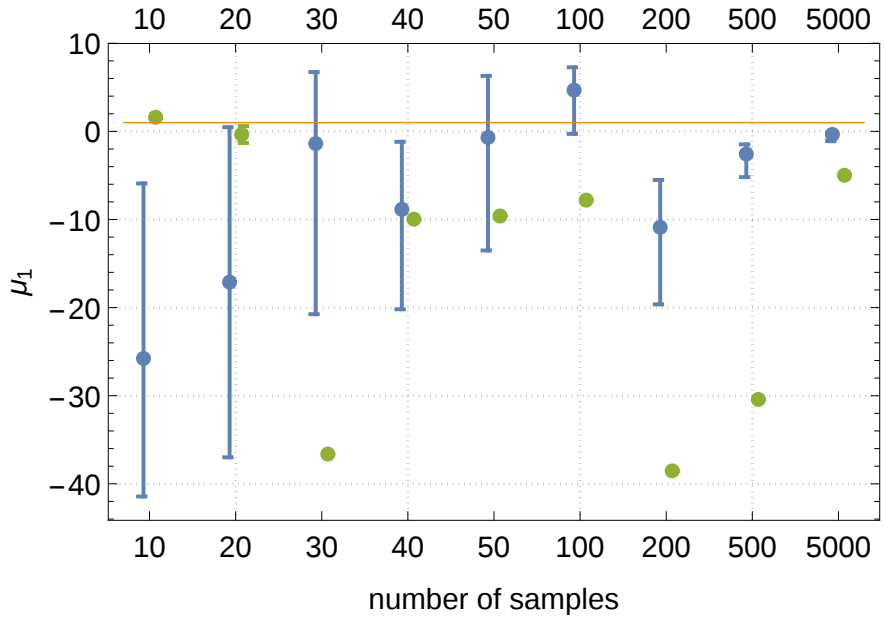


Figure A.5: Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 50$. The orange line represents the true value.


10 20 30 40 50 100 200 500 5000
1.0
0.8
0.6
0.4
0.2
0.0
p
10 20 30 40 50 100 200 500 5000
number of samples


10 20 30 40 50 100 200 500 5000
10
0
−10
−20
−30
−40
μ_1
10 20 30 40 50 100 200 500 5000
number of samples

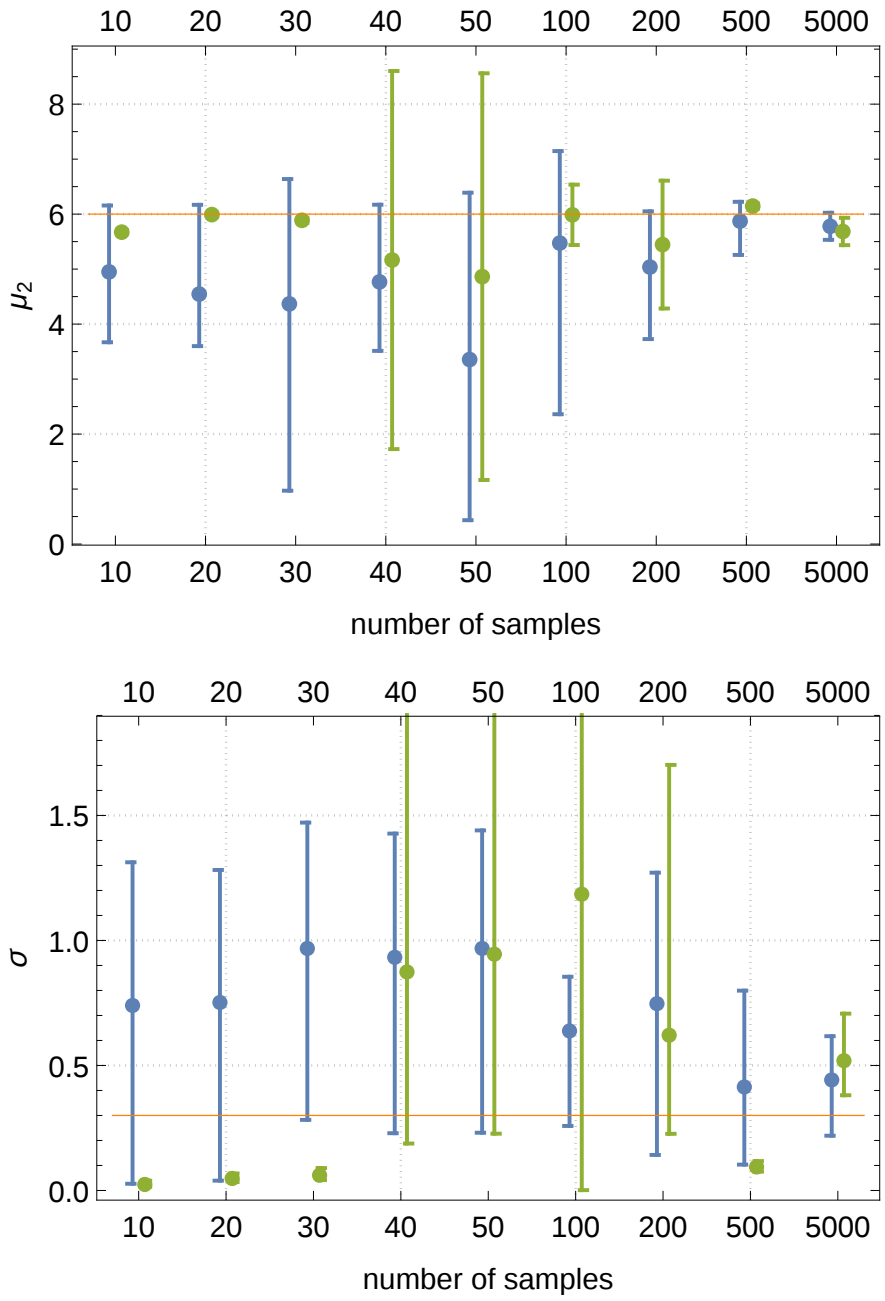


Figure A.6: Resulting estimate and intervals of parameter estimation by MCMC (*blue*) and MLE (*green*), $n = 100$. The orange line represents the true value.

Listing 1: MCMC model definition for the univariate case

```
model.non <- function(parm, data) {  # parm = p,mu1,mu2,sigma
## Parameters
p <- interval(parm[1],0,1)
parm[1] <- p
mu1 <- parm[2]
mu2 <- parm[3]
sigma <- interval(parm[4], 1e-100, Inf)
parm[4] <- sigma
## Log-Prior
p.prior <- dbeta(p, shape1=0.9,shape2=0.9, log=TRUE)
mu1.prior <- dnorm(mu1, mean=0, sd=100, log = TRUE)
mu2.prior <- dnorm(mu1, mean=0, sd=100, log = TRUE)
sigma.prior <- dhalfcauchy(sigma, 25, log=TRUE)
## Loglikelihood
LL <- sum(d.sum.of.mixtures(y=data$y, n=30, p.vector=c(p,1-p),
      mu.vector=c(mu1,mu2),
      sigma.vector=c(sigma,sigma), logdens = TRUE))
## Log-Posterior
LP <- LL + p.prior + mu1.prior + mu2.prior + sigma.prior
ModelOut <- list(LP = LP, Dev = -2 * LL, Monitor = LP, yhat = NA,
            parm = parm)
return(ModelOut)
}
```

Listing 2: MCMC model definition for the bivariate case

```
model.mv <- function(parm, data) {  # parm = p,mu11,mu12,mu21,
    mu22,sigma,rho
        ## Parameters
        p <- interval(parm[1],0,1)
        parm[1] <- p
        mu11 <- parm[2]
        mu12 <- parm[3]
        mu21 <- parm[4]
        mu22 <- parm[5]
        sigma <- interval(parm[6], 1e-100, Inf)
        parm[6] <- sigma
        cv <- interval(parm[7],0,1)
        parm[7] <- cv
        ## Log-Prior
        p.prior <- dbeta(p, shape1=0.9,shape2=0.9, log=TRUE)
        mu11.prior <- dnorm(mu11, mean=0, sd=100, log = TRUE)
        mu12.prior <- dnorm(mu12, mean=0, sd=100, log = TRUE)
        mu21.prior <- dnorm(mu21, mean=0, sd=100, log = TRUE)
        mu22.prior <- dnorm(mu22, mean=0, sd=100, log = TRUE)
        sigma.prior <- dhalfcauchy(sigma, 25, log=TRUE)
        cv.prior <- dbeta(cv, shape1=2,shape2=2, log=TRUE)
        ## Loglikelihood
        LL <- sum(data.list$d.sum.of.mixtures.mv(y=data$y, n=10,
            p.vector=c(p,1-p), mu1.vector=c(mu11,mu21), mu2.
            vector=c(mu12,mu22)
        , sigma.vector=c(sigma,sigma), cor=cv, logdens = TRUE))
        ## log posterior
        LP <- LL + p.prior + mu11.prior + mu12.prior + mu21.prior
             + mu22.prior + sigma.prior + cv.prior
        ModelOut <- list(LP = LP, Dev = -2 * LL, Monitor = LP,
            yhat = NA,
        parm = parm)
        return(ModelOut)
}
```

BIBLIOGRAPHY

J. Aitchison and J.A. Brown. *The Lognormal Distribution*. Department of Applied Economics Monographs. Cambridge University Press, 1957. ISBN 9780521040112.

Jim Albert. *Bayesian Computation with R*. Use R! Springer, 2009. ISBN 9780387922980.

Tomohiro Ando. Predictive bayesian model selection. *American Journal of Mathematical and Management Sciences*, 31(1-2):13–38, 2011. doi: 10.1080/01966324.2011.10737798.

F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 03 1963. doi: 10.1214/aoms/1177704255.

Adriano Azevedo-Filho and Ross D. Shachter. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In R. Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence*, volume 57, pages 97–109. Morgan Kauffman, 1994.

Sameer S. Bajikar, Christiane Fuchs, Andreas Roller, Fabian J. Theis, and Kevin A. Janes. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proceedings of the National Academy of Sciences*, 111(5):E626–E635, 2014. doi: 10.1073/pnas.1311647111.

N. Balakrishnan and W.W.S. Chen. *Handbook of Tables for Order Statistics from Lognormal Distributions with Applications*. Springer, 1999. ISBN 9780792357124.

Thomas Bayes and Richard Price. An Essay Towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. doi: 10.1098/rstl.1763.0053.

Isabel Beichl and Francis Sullivan. The metropolis algorithm. *Computing in Science and Engg.*, 2(1):65–69, January 2000. ISSN 1521-9615. doi: 10.1109/5992.814660.

Martin Bengtsson, Anders Ståhlberg, Patrik Rorsman, and Mikael Kubista. Gene expression profiling in single cells from the pancreatic islets of langerhans reveals lognormal distribution of mrna levels. *Genome research*, 15(10):1388–1392, 2005. doi: 10.1101/gr.3820805.

James Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 09 2006. doi: 10.1214/06-BA115.

Josè M Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, West Sussex, England, 2009. ISBN 9780470317716.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2011. ISBN 9781584886983.

Gabriela Ciuperca, Andrea Ridolfi, and Jérome Idier. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30(1):45–59, 2003.

Edwin L. Crow and Kunio Shimizu, editors. *Lognormal Distributions: Theory and Applications*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1987. ISBN 9780824778033.

B. De Finetti. Probabilismo. *Erkenntnis*, 31:169–223, 1931.

Lawrence Fenton. The sum of log-normal probability distributions in scatter transmission systems. *Communications Systems, IRE Transactions on*, 8(1):57–67, 1960.

Ronald Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4):3–32, 1921.

A. Gelman, G. Roberts, and W. Gilks. Efficient metropolis jumping hules. *Bayesian statistics*, 5:599–608, 1996.

Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, 2008. ISSN 1097-0258. doi: 10.1002/sim.3107.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596.

Michael Goldstein. Subjective bayesian analysis: Principles and practice. *Bayesian Analysis*, 1(3):403–420, 09 2006. doi: 10.1214/06-BA116.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97.

Joseph G. Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 02 2000. doi: 10.1214/ss/1009212673.

Kevin A. Janes, Chun-Chao Wang, Karin J. Holmberg, Kristin Cabral, and Joan S. Brugge. Identifying single-cell molecular programs by stochastic profiling. *Nature Methods*, 7(4):311–317, 2010. doi: 10.1038/nmeth.1442.

Harold Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, 1961.

Y.K. Kim. *Handbook of Behavior Genetics*. Biomedical and Life Sciences. Springer, 2009. ISBN 9780387767277.

John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial Introduction using R and BUGS*. Academic Press, Oxford, Second edition, 2010. ISBN 9780123814869.

Pierre-Simon Laplace. *A Philosophical Essay on Probabilities.* Oxford, 1814. ISBN 1602063281.

A.M. Law and W.D. Kelton. *Simulation modeling and analysis*. McGraw-Hill series in industrial engineering and management science. McGraw-Hill, 2000. ISBN 9780070592926.

Jean-Michel Marin and Christian P. Robert. *Bayesian Essentials with R (2nd ed.)*. Springer, Berlin, 2014. ISBN 978-1-4614-8686-2.

N.A. Marlow. A normal limit theorem for power sums of independent random variables. *Bell System Technical Journal*, 46(9):2081–2089, 1967. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1967.tb04244.x.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *jcp*, 21:1087–1092, June 1953. doi: 10.1063/1.1699114.

James R. Norris. *Markov Chains*. Number Nr. 2008 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 9780521633963.

J.K. Patel and C.B. Read. *Handbook of the Normal Distribution, Second Edition*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1996. ISBN 9780824793425.

Martyn Plummer. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008.

Nicholas G. Polson, James G. Scott, et al. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012. doi: 10.1214/12-BA730.

Howard Raïffa and Robert Schlaifer. *Applied statistical decision theory*. Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University, 1961. ISBN 9780875840178.

Frank P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. Harcourt, Brace and Company, New York, 1926.

Guinness World Records, editor. *Guinness World Records 2013*. Guinness World Records, London, 2012. ISBN 9781904994879.

Christian Ritter and Martin A. Tanner. Facilitating the gibbs sampler: the gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, 87(419):861–868, 1992. doi: 10.1080/01621459.1992.10475289.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, Second edition, 2007. ISBN 9780387715988.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, 2005. ISBN 9780387212395.

Christian P. Robert, Jean-Marie Cornuet, Jean-Michel Marin, and Natesh S. Pillai. Lack of confidence in approximate bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117, 2011. doi: 10.1073/pnas.1102900108.

Donald B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4): 1151–1172, 12 1984. doi: 10.1214/aos/1176346785.

L. J. Savage. *The Foundations of Statistics*. Wiley publications in statistics. John Wiley & Sons, West Sussex, England, 1954.

Alex K. Shalek, Rahul Satija, Xian Adiconis, Rona S. Gertner, Jellert T. Gaublomme, Raktima Raychowdhury, Schragi Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498:236–240, 2013. doi: 10.1038/nature12172.

Robert L. Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984. doi: 10.1287/opre.32.6.1296.

Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20(3):357–366, 2010.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00353.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Stephen M. Stigler. Who discovered bayes's theorem? *The American Statistician*, 37(4):290–296, 1983.

Valentin F. Turchin. On the computation of multidimensional integrals by the monte-carlo method. *Theory of Probability & Its Applications*, 16(4):720–724, 1971. doi: 10.1137/1116083.

Angelika van der Linde. DIC in variable selection. *Statistica Neerlandica*, 59(1):45–56, 2005. ISSN 1467-9574. doi: 10.1111/j.1467-9574.2005.00278.x.

S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009. ISBN 9780521864671.

Wikipedia. Thomas bayes — Wikipedia, the free encyclopedia, 2015a. URL http://en.wikipedia.org/wiki/Thomas_Bayes. [Online; accessed 22-January-2015].

Wikipedia. Andrey markov — Wikipedia, the free encyclopedia, 2015b. URL http://en.wikipedia.org/wiki/Andrey_Markov. [Online; accessed 22-January-2015].

Roger I. Wilkinson. The combination of probability curves in engineering. *Electrical Engineering*, 61(12):953–962, 1942.

DECLARATION

Ich erkläre, dass ich die Arbeit selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benützt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

München, February 2015

Christoph Franz Kurz