Bayesian Inference in Large Data Problems

Matias Quiroz

Bayesian Inference in Large Data Problems

Matias Quiroz

Abstract

In the last decade or so, there has been a dramatic increase in storage facilities and the possibility of processing huge amounts of data. This has made large high-quality data sets widely accessible for practitioners. This technology innovation seriously challenges traditional modeling and inference methodology.

This thesis is devoted to developing inference and modeling tools to handle large data sets. Four included papers treat various important aspects of this topic, with a special emphasis on Bayesian inference by scalable Markov Chain Monte Carlo (MCMC) methods.

In the first paper, we propose a novel mixture-of-experts model for longitudinal data. The model and inference methodology allows for manageable computations with a large number of subjects. The model dramatically improves the out-of-sample predictive density forecasts compared to existing models.

The second paper aims at developing a scalable MCMC algorithm. Ideas from the survey sampling literature are used to estimate the likelihood on a random subset of data. The likelihood estimate is used within the pseudo-marginal MCMC framework and we develop a theoretical framework for such algorithms based on subsets of the data.

The third paper further develops the ideas introduced in the second paper. We introduce the difference estimator in this framework and modify the methods for estimating the likelihood on a random subset of data. This results in scalable inference for a wider class of models.

Finally, the fourth paper brings the survey sampling tools for estimating the likelihood developed in the thesis into the delayed acceptance MCMC framework. We compare to an existing approach in the literature and document promising results for our algorithm.

Keywords: Bayesian inference, Large data sets, Markov chain Monte Carlo, Survey sampling, Pseudo-marginal MCMC, Delayed acceptance MCMC.

©Matias Quiroz, Stockholm 2015
Cover image ©Alexander Danielsson, `www.vivalda.se`

ISBN 978-91-7649-199-7

Printed in Sweden by Holmbergs, Malmö 2015

Distributor: Department of Statistics, Stockholm University

To Lynn

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: Quiroz, M. and Villani, M. (2013), Dynamic mixture-of-experts models for longitudinal and discrete-time survival data, *Submitted*.

PAPER II: Quiroz, M., Villani, M. and Kohn, R. (2015), Speeding up MCMC by efficient data subsampling, *Submitted*.

PAPER III: Quiroz, M., Villani, M. and Kohn, R. (2015), Scalable MCMC for large data problems using data subsampling and the difference estimator, *Manuscript*.

PAPER IV: Quiroz, M. (2015), Speeding up MCMC by delayed acceptance and data subsampling, *Manuscript*.

Contents

Acknowledgements

It is with a lot of emotions that I realize that my journey as a doctoral student is approaching its end. A journey seemingly full of MCMC samplers failing to converge, programming bugs generated by some artificial intelligence mechanism (I did *not* do that!), and algebraic mistakes made by my inherent sloppiness. However, somewhere along the way, the samplers started to converge (most of the time), the artificial intelligence became less of a frequent visitor, and my sloppiness turned out to be not so inherent. For this progress I have my supervisor Mattias Villani to thank. Thank you from the bottom of my heart, Mattias. Your profound knowledge is a source of inspiration and your patience has been invaluable to me. You will doubtlessly remain the most important influence of my academic career. I also owe you a great deal for moving me away from the world of hypothesis testing. *It pays to be Bayes!*

I would also like to thank Robert Kohn who co-authored two of the papers in the thesis. I have learned a lot from you and I look forward to future collaborations. I thank my co-supervisor Gebrenegus Ghilagaber for your comments and suggestions regarding my work, and for always being kind and friendly. I thank Daniel Thorburn, Dan Hedlin and Olivia Ståhl for pointing me in the right directions in the survey sampling literature. I thank Paolo Giordani for showing interest in my work and for our fruitful discussions.

I would like to express my deepest gratitude to Kasper Roszbach, who suggested me to apply for a VINNOVA grant to pursue a Ph.D. and made important arrangements. I also thank Tor Jacobson and Jesper Lindé for their arrangements that made it possible for me to be part of the Research division of Sveriges Riksbank during my studies. In addition, I thank you for providing me with guidance regarding my future career.

I gratefully acknowledge all the financial support from VINNOVA, Stockholm University, Linköping University and Sveriges Riksbank.

During this journey, I have been surrounded by a lot of great people, both at Sveriges Riksbank and Stockholm University. Thank you all for making my work environment such a pleasant and friendly place. In particular, I would like to thank Mats Levander, my office mate and partner in crime since day one. Thank you for sharing the pain of being a Ph.D. student, but more importantly the same (sick) sense of humor, which has given me many good laughs over the years. *Oh yeah Scott!*

Special thanks to Richard Hager and Håkan Slättman at Stockholm University, and Lena Löfgren at Sveriges Riksbank, for all the help on administrative issues.

I thank my academic elder brothers: Bertil and Feng, and my younger ones: Per, Måns and Josef, for many great moments. The memories I share with some of you from the ISBA meetings in Kyoto and Cancún are some of the most valuable from my Ph.D. odyssey. Not to forget, the week I spent with Feng in Southampton. I hope we all travel to Sardinia for ISBA 2016 and have a big reunion (*Feng, if you read this, book your calendar!*).

A big thanks to all the silent heroes out there who develop and provide free software of outstanding quality. Writing this thesis in Windows with Microsoft products would certainly be an obstacle impossible to overcome.

I thank my good friend and talented artist Alexander Danielsson for making the cover image, a very nice drawing of the most beautiful theorem of them all.

Outside academia, there have been a number of people who by their love and support made this thesis possible. I would like to thank my friends Marcus, Mattias, Tomas and Rafael for the good portion of bromance and unconditional love you provide me with. The latter is really all you need from friends in times when the load of work consumes daily life. I thank my primo Mario for his love, wisdom and support, especially when life was tough. You are the elder brother I never had.

I am grateful to my parents, sister and brother. I love you with all my heart and I am proud that we always stay strong together, no matter what. Thank you mum and dad for always putting me and my siblings first. You left your family in Chile and moved to Sweden, only to provide a better future for your children. You are the best role models I could ever have.

Finally, I would like to thank my better half, soul mate, best friend and wife, Lynn. Your patience and understanding during times of intensive work have exceeded my expectations by far. Your ability of constantly making me laugh and smile makes you the source of energy required to complete this thesis. We did this together and therefore, to you, I dedicate this thesis.

Stockholm, July 2015 Matias Quiroz

1. Introduction

1.1 Background

This thesis develops inference and modeling techniques in the presence of the nowadays increasingly more available large data sets. Recent technological developments has made it possible to record high quality data on the micro level for many applications. Such detailed data opens up an interesting opportunity for so called complex models with abilities to capture a wide range of features in the data. Arguably, the Bayesian paradigm, with its inherent regularization properties and appealing modeling of uncertainty, provides an attractive framework for inference on complex models.

However, the increase in data volume challenges traditional inference methods. There is an increasing awareness of the importance of developing inference algorithms that can remain practical in the age of so called big data. This is in particular true for simulation algorithms commonly used in Bayesian inference, which typically are very computationally demanding, but are often worthwhile because they provide a full characterization of the uncertainty.

1.2 Aims of the thesis

The overall aim of the thesis is to develop statistical tools with the ability to scale to large data sets in terms of many observations. In doing so, it is of crucial importance to not only be faster with respect to computing time, but also preserve the statistical efficiency of existing tools as far as possible.

The thesis considers essentially two distinct approaches to handle large data problems.

1. The formulation of the model itself allows for efficient inference in the presence of many observations. This is the approach in Paper I, where we via a model assumption can handle relatively many observations that would not be feasible otherwise.

2. Existing estimation procedures are adapted to handle large data sets. This is the approach in Paper II-IV where scalable inference is achieved

by combining survey sampling methodology with so called Markov chain Monte Carlo algorithms.

1.3 Outline

This introduction aims to give an overview at an elementary level of the distinct topics considered in the thesis. The topics survey sampling and Markov chain Monte Carlo are first presented separately. Towards the end, we explain how they are merged to achieve the goal of developing inference tools for large data problems.

The thesis considers several statistical models. Some of the models are not treated in a standard textbook in statistics and we provide some basics and underlying intuition for these models in Chapter 2. In Chapter 3 we present the main elements of survey sampling. Chapter 4 motivates Bayesian inference and considers computational aspects, emphasizing Markov chain Monte Carlo simulation. Chapter 5 explains how to merge survey sampling and Markov chain Monte Carlo to achieve scalable inference. Finally, Chapter 6 summarizes the papers included in the thesis.

2. Statistical models

2.1 Survival models

What distinguishes survival analysis from standard statistical analysis is the type of data considered. In a standard statistical analysis the data is *complete*, which, loosely speaking, means that it has fully been observed by the scientist. In contrast, survival type of data may be *censored*, meaning that it does not provide full information about the phenomenon under study.

To exemplify, suppose that a scientist makes an experiment with the objective to model the variable $t =$ "time to cure" for a certain medicament. To this end, the scientist conducts a study with duration time one week and collects data. This data is *right censored*, because t will only be observed for a fraction of the subjects. The remainder are only known to have $t > 1$ week as they may eventually be cured (but it has not been observed yet). The censoring clearly needs to be taken into account when making inference about t. There are other types of censoring but this thesis will only consider right censored subjects.

There is a vast literature on survival analysis when the "time to event" is continuous, see e.g. Miller *et al.* (1981) and Ibrahim *et al.* (2005). However, this thesis considers the discrete time case and we now turn to how this can be derived from a continuous time setting. For a thorough review of discrete time survival see Singer and Willett (1993).

2.1.1 Continuous time

We first introduce the continuous time model underlying the discrete time case considered in the next section. Let the continuous random variable T^c denote the time to some unrepeatable event. Assume that T^c has sample space $\{t : t \geq 0\}$ and a distribution $f(t|\lambda)$ parametrized by some parameter vector λ. Subject i is *right censored* if the event has not been observed before the *censoring time* T^*, and no further information about subject i is available after T^*. T^* is typically the end of the study (recall the example above), but some subjects may leave the study early so T^* may vary across subjects.

Introduce $c_i = 1$ if the ith subject is censored at time T_i^* and $c_i = 0$ otherwise. The *likelihood function*, i.e. the density of the data as a function of λ, if

the subjects are conditionally independent is

$$L(\lambda) = \prod_{i=1}^{n} f(t_i|\lambda)^{1-c_i} S(T_i^*|\lambda)^{c_i},$$

where $S(t|\lambda) = P(T^c > t|\lambda)$ is known as the *survival function*. A common way of representing the distribution of T^c is through the *hazard function*,

$$\begin{aligned} h(t|\lambda) &= \lim_{\Delta t \to 0} \frac{P(t \leq T^c < t + \Delta t | T^c \geq t, \lambda)}{\Delta t} \\ &= \frac{f(t|\lambda)}{S(t|\lambda)}, \end{aligned}$$

which is interpreted as the instantaneous rate of experiencing the event given that it has not been experienced yet. The survival function relates to the hazard function through

$$S(t|\lambda) = \exp\left(- \int_0^t h(u|\lambda) \mathrm{d}u \right). \tag{2.1}$$

The extension to regression is made by including dependence of the distribution of T^c on covariates x. As an example, the well known Cox regression model (Cox, 1972) is obtained by $h(t|\lambda, x) = h_0(t) \exp(x'\beta)$, where β includes the regression parameters and $h_0(t)$ is the *baseline hazard*.

2.1.2 Extensions to discrete time

Survival data are often observed in discrete time, for example weekly, monthly or yearly. It is therefore natural to model the time to event discretely. In addition, there are some advantages compared with the continuous time setting. The most important are that time-varying covariates can easily be incorporated and that the proportional hazards assumption in Cox regression model can be relaxed (Singer and Willett, 1993).

Assume that a study is observed over J periods which are divided as

$$(0, t_1], (t_1, t_2], \ldots, (t_{J-1}, t_J].$$

Let $T \in \{1, 2, \ldots\}$ be the (discrete) random variable recording the time period where the event occurs, such that $T = j$ if $T^c \in (t_{j-1}, t_j]$, with T^c as in Section 2.1.1.

We will express the likelihood in terms of the hazard probabilities which we denote $h_j = P(T = j | T \geq j)$. The hazard probability relates to the survival function as

$$h_j = \frac{P(T > j-1) - P(T > j)}{P(T > j-1)} = 1 - \frac{S(j)}{S(j-1)},$$

where S is computed by Equation (2.1) with $h(t|\lambda)$ being interpreted as the hazard *rate* of the underlying continuous random variable. The likelihood for a complete observation is

$$P(T = j) = \left(\prod_{k=1}^{j-1} (1 - h_k) \right) h_j,$$

and for a censored observation,

$$P(T > j) = \prod_{k=1}^{j} (1 - h_k).$$

Let the ith subjects' hazard probability at period j be denoted $h_{ij}(x_{ij})$. Assuming n independent subjects, the likelihood of all data is expressed as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{n_i} h(x_{ij})^{y_{ij}} (1 - h(x_{ij}))^{1-y_{ij}},$$

where

$$y_{ij} = \begin{cases} 0, \text{ if subject } i \text{ does not experience the event at period } j, \\ 1, \text{ if subject } i \text{ does experience the event at period } j. \end{cases}$$

From an inferential point of view, it is interesting to note that this likelihood has the same form as in the case of regression for binary data with link function h^{-1}. Note that the recording of data is done through the binary representation of the response; if subject i has n_i periods then the observation is recorded as a sequence $\{y_{ij}, x_{ij}\}_{j=1}^{n_i}$. Whenever the subject is censored y consist of only zeros, and if the event is experienced the sequence is terminated by 1 at the time period where the event took place.

2.2 Finite mixture models

When applying any parametric model to a data set, the scientist implicitly assumes that the data distribution is in agreement with the shape of the assumed model. There exist parametric models that can capture very flexible features of the data, such as e.g. a highly skewed distribution or fat tails. However, even the most flexible parametric families are often unimodal. In many applications the scientist may face data that indeed seem to be generated from a multimodal distribution. Finite mixture models offer a way to model such data by combining parametric models. Essentially any complex distributional shape can be obtained by combining simple distributions. Two examples are illustrated in Figure 2.1.

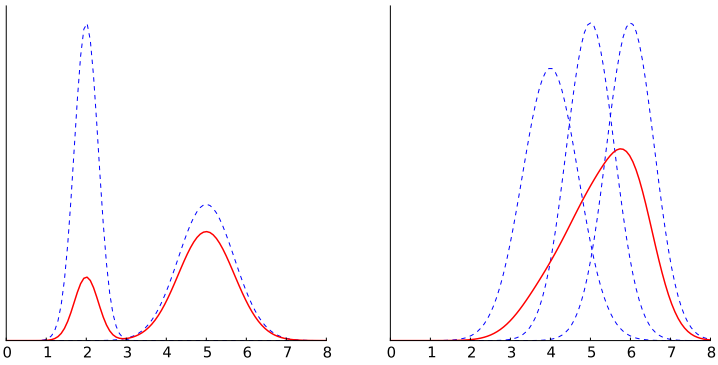

Figure 2.1: Examples of a finite mixture - The figure shows how Gaussian densities (dashed blue lines) can be combined to represent a more complex density (solid red line). The left panel shows a finite mixture of two components modeling a bimodal density. The right panel shows a finite mixture with three components modeling a skewed density.

We start from the most basic case of a finite mixture followed by a brief review of the so called mixture-of-experts model. This idea is then extended to a *longitudinal data* setting, which is the topic explored in Paper I. See Frühwirth-Schnatter (2006) for a comprehensive introduction to finite mixture models.

2.2.1 The basic case

The most simple case of a finite mixture model with K components assumes that the data y has density

$$p(y|\theta) = \sum_{k=1}^{K} w_k p_k(y|\theta_k), \text{ with } \quad \theta = (\theta_1, \ldots, \theta_K), \tag{2.2}$$

where θ_k are the parameters corresponding to the kth component density. Since $\sum_{k=1}^{K} w_k = 1$ it follows that $p(y|\theta)$ is a proper density. The probability w_k is interpreted as the prior probability that the observation belongs to component k.

Finite mixture models are also useful for model based clustering. The idea here is to probabilistically classify which cluster (component) an observation belongs to. Consider the alternative formulation of the model in (2.2),

$$\begin{aligned} p(y|\theta, s) &= p_s(y|\theta_s), \\ p(s = k) &= w_k, \end{aligned} \tag{2.3}$$

where s denotes the component membership. The classification is based on the posterior probability,

$$p(s = k|y) \propto w_k p_k(y|\theta_k),$$

which combines the prior probability with the probability of the data under the given component. One typically classifies y to the component with the highest posterior probability.

An advantage of writing the model in its augmented form (2.3) is that it facilitates the estimation through the Gibbs sampler (see Section 4.3.1).

2.2.2 Mixture-of-experts

A mixture-of-experts model is an extension of the model in Section 2.2.1 to a regression setting, where we also assume that the weights w_k are covariate dependent. Early references are Jacobs *et al.* (1991) and Jordan and Jacobs (1994).

A common choice of w_k, known as the *gating* or *mixing* function, is the multinomial logit

$$w_k \quad = \quad \frac{\exp(x'\gamma_k)}{\sum_{l=1}^{K} \exp(x'\gamma_l)}, \quad \text{with} \quad \gamma_1 = 0$$

for identification. The mixture-of-experts models the regression density (suppressing dependence on parameters)

$$p(y|x) = \sum_{k=1}^{K} w_k(x) p_k(y|x).$$

It is straightforward to derive the model's augmented form as in (2.3).

There are some results in the literature that a mixture-of-experts model can approximate certain densities arbitrarily well as $K \to \infty$. Jiang and Tanner (1999) proves this fact under the following two important assumptions. First, the component densities belong to the Generalized Linear Model (GLM) family (Nelder and Wedderburn, 1972) with a linear predictor. Second, the target density (what we want to approximate) is also a GLM but with a flexible predictor. Norets (2010) has a more general proof that does not restrict the target density to GLM, although requires it to be continuous.

2.2.3 A longitudinal generalization

When facing longitudinal data one has to rethink the concept of a finite mixture model. Since a subject is now followed through time, we can think about the finite mixture in at least two ways:

1. Each subject has a single component indicator s over its entire life span. In other words, the subject belongs to the same component for all time periods. We refer to this as a *static* mixture.

2. In each period, the subject belongs to component s_t, $t = 1,\ldots,T$, where T is the total number of time periods. The component indicator becomes $s = (s_1,\ldots,s_T)$. . We refer to this as a *dynamic* mixture.

The main contribution of Paper I is a model for the second point above which we now discuss in some depth.

To generalize to a longitudinal data setting we introduce the following notation. For notational clarity we here assume a single subject that is observed for T time periods. Let

$$y_{1:T} = (y_1,\ldots,y_T)' \in \mathbb{R}^{T\times 1}, \qquad x_{1:T} = (x_1,\ldots,x_T)' \in \mathbb{R}^{T\times p_x} \quad \text{and}$$
$$z_{1:T} = (z_1,\ldots,z_T)' \in \mathbb{R}^{T\times p_z},$$

where both x and z are covariate vectors of length p_x and p_z, respectively. The regression density is now the multivariate $p(y_{1:T}|x_{1:T})$, which we say is *p-lag longitudinal* if it factorizes as

$$p(y_{1:T}|x_{1:T}) \quad = \quad \prod_{j=1}^{T} p(y_j|y_{j-p:j-1},x_j), \tag{2.4}$$

under the usual assumption that p pre-sample observations $y_0, y_{-1}, ..., y_{-(p-1)}$ are available. The dynamic model we propose is a finite mixture for each of the conditional distribution in (2.4), i.e.

$$p(y_j|y_{j-p:j-1},x_j) = \sum_{k=1}^{K} w_j^{(k)}(z_j) p_k(y_j|y_{j-p:j-1},x_j),\ j = 1, ..., T,$$

where

$$w_j^{(k)}(z_j) \quad = \quad \frac{\exp(z_j'\gamma_k)}{\sum_{l=1}^{K}\exp(z_j'\gamma_l)}, \quad \text{with } z_j = (x_j, y_{j-p:j-1})' \text{ and } \gamma_1 = 0.$$

It is straightforward to write the model on the augmented form as in (2.3) with $s = s_{1:T}$.

It is illustrative to think about the following example to understand the dynamic model. Suppose that one is interested in predicting firm bankruptcy. For simplicity, assume that we have two components ($K = 2$) and a single covariate, say cash holdings, which only enters in the mixing function. Furthermore, suppose that one component is interpreted as a low risk component, whereas

the other is a high risk one. The model then assumes that the observables $z_{1:T}$ (cash holdings) are determining the overall risk of a firm and, importantly, that its risk classification can vary through time.

A natural approach to model the unobserved $s_{1:T}$ is to assume it follows a Markov model. The sampling of $s_{1:T}$ (for each subject) is then performed by sequential Monte Carlo methods (Doucet *et al.*, 2000). This is computationally infeasible in our application due to the large number of subjects. We therefore make the assumption that $s_{1:T}$ is an independent sequence *conditional* on the path of time-varying covariates $z_{1:T}$, i.e.

$$P(s_{1:T} = k_{1:T}|z_{1:T}) \quad = \quad \prod_{j=1}^{T} P(s_j = k_j|z_j),$$

with $k_{1:T} = (k_1, \ldots, k_T)$ and $1 \leq k_j \leq K$ for $j = 1, \ldots, T$. The temporal dependence of $s_{1:T}$ is thus induced by the path of the time series for the covariates in $z_{1:T}$. This assumption makes it straightforward to sample $s_{1:T}$ independently for all subjects within the Gibbs sampler.

In Paper I the result of Jiang and Tanner (1999) (see Section 2.2.2) is generalized for the joint distribution $p(y_{1:T}|x_{1:T})$ of the longitudinal data under the p-lag assumption in (2.4). In other words, the proposed dynamic mixture model can approximate certain $p(y_{1:T}|x_{1:T})$ arbitrarily well if enough components are used in the model.

3. Survey sampling

Survey sampling deals with the problem of estimating quantities of a *finite population* based on a random sample. The variable of interest is recorded for the sampled elements and the goal is to infer, for example, the total of the population. As in any statistical analysis, the scientist wants accurate statements and, since the survey may be costly, this should ideally be achieved with a relatively small sample. Arguably, the key ingredients of a survey is the *sampling scheme* (how the random sample is selected) and the choice of *estimator*.

This chapter reviews the sampling schemes and estimators used in the thesis. For a comprehensive introduction to survey sampling see Särndal *et al.* (2003).

3.1 Notations and preliminaries

We consider a population consisting of n elements with the index set

$$F = \{1, \ldots, k, \ldots, n\}.$$

Let Y denote a variable and let Y_k be the value for the kth element in the population. For any set $A \subseteq F$, let

$$Y_A := \{Y_k; k \in A\}$$

and let $\bar{Y}_A$ be the mean of the collected elements.

The main problem in this chapter is the estimation of the population total

$$t = \sum_{k \in F} Y_k \tag{3.1}$$

based on a subsample S with $|S| = m$. An important distinction between sampling schemes is whether they use *without replacement* or *with replacement* of the sampling elements. In general, with replacement sampling gives a higher variance of any estimator because selecting the same element more than once does not add any new information about the (finite) population. We note that with replacement gives sampling elements that are independent.

In the case of without replacement sampling, we define the selection indicators

$$u_k = \begin{cases} 1, \text{ if element } k \text{ is included in the sample,} \\ 0, \text{ otherwise.} \end{cases} \tag{3.2}$$

Let $\pi_k = \Pr(u_k = 1)$ denote the inclusion probability of element k. Another important probability is the *second order* inclusion probability

$$\pi_{kl} = \Pr(u_k u_l = 1),$$

i.e. the probability of two given elements appearing in S. We also note that $\pi_{kk} = \pi_k$.

For with replacement sampling, a single observation can appear more than once and we redefine u and π,

$$\begin{aligned} u_i &= k, \text{ if element } k \text{ is selected in the } i\text{th draw, } i = 1, \ldots, m, \\ \pi_k &= \Pr(\text{element } k \text{ appears at least once in } m \text{ draws}). \end{aligned} \tag{3.3}$$

If we define

$$p_k = \Pr(\text{element } k \text{ is selected in a given draw}),$$

it is clear that $\pi_k = 1 - (1 - p_k)^m$. We stress the fact that the u's are dependent in (3.2) whereas independent in (3.3).

Both types of sampling schemes are used in the thesis. The without replacement is not considered in advanced sampling schemes for computational reasons as discussed in the next section. Paper II and III use with replacement sampling as the independence of the u's is crucial for the development of the theory. Paper IV, however, does not rely on this theory and therefore a without replacement scheme is implemented. It should also be noted that when $m << n$ the schemes are approximately equal.

3.2 Sampling schemes

This section begins with the simplest possible sampling scheme and then considers a more advanced one. We postpone an explanation of the superiority of the advanced sampling until Section 3.3 when the problem of estimating t in (3.1) is addressed. The thesis considers sampling schemes with a fixed sample size m.

3.2.1 Simple random sampling

Simple random sampling assigns the same probability to any possible subset S (with m elements), i.e.

$$p(S) = \begin{cases} 1/\binom{n}{m}, & \text{for without replacement,} \\ 1/n^m, & \text{for with replacement.} \end{cases} \tag{3.4}$$

$p(S)$ is referred to as the sampling design. We avoid to explicitly define the sample space of S and rather think directly of $p(S)$, which is simple to understand for the schemes considered here.

We use the abbreviation SI and SIR for without and with replacement sampling, respectively. It is clear that

$$\begin{aligned} \pi_k &= m/n \quad \text{(SI)}, \\ p_k &= 1/n \quad \text{(SIR)}. \end{aligned}$$

Finally, it can be shown that

$$\pi_{kl} = \frac{m(m-1)}{n(n-1)} \quad \text{(SI)}.$$

3.2.2 Probability proportional-to-size sampling

The idea in proportional-to-size sampling is to choose π_k (or p_k) proportional to a measure of the element Y_k's size which we denote W_k. It is assumed that W_k is available for all $k \in F$ prior to the sampling. Hence,

$$\pi_k = \frac{W_k}{\sum_{k \in F} W_k} \tag{3.5}$$

is known for the full population, and similar for p_k.

The proportional-to-size scheme without replacement, often called πPS sampling, has an intractable sampling design $p(S)$. Furthermore, it is very tedious to compute π_{kl} and consequently the implementation is computationally slow. We will therefore instead consider PPS sampling, which is the equivalent when sampling is with replacement. The sampling design is

$$p(S) = p_{u_1} \cdot \cdots \cdot p_{u_m},$$

with u_i as in (3.3). It is straightforward to implement this scheme.

3.3 Estimators

Once the random sample is obtained it is used to estimate the population total. We will consider two different estimators; one that is typically used for without replacement schemes (Horvitz-Thompson) and the other for with replacement schemes (Hansen-Hurwitz). Some basic properties of the estimators are presented. We then consider an estimator that incorporates auxiliary information about the elements in the population to achieve a more precise estimate.

3.3.1 Horvitz-Thompson

For with replacement sampling the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) is considered. This estimator has the form

$$\begin{aligned} \hat{t}_{HT} &= \sum_{k\in S} \frac{Y_k}{\pi_k} \\ &= \sum_{k\in F} \frac{Y_k}{\pi_k} u_k \quad \text{with } u_k \text{ in (3.2)}, \end{aligned} \tag{3.6}$$

and is an unbiased estimate of t in (3.1), i.e.

$$E[\hat{t}_{HT}] = t.$$

Furthermore, the variance of the estimator is

$$V[\hat{t}_{HT}] = \sum_{k\in F}\sum_{l\in F} (\pi_{kl} - \pi_k\pi_l)\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}, \tag{3.7}$$

and an unbiased estimate of the variance is

$$\hat{V}[\hat{t}_{HT}] = \sum_{k\in S}\sum_{l\in S} (1 - \frac{\pi_k\pi_l}{\pi_{kl}})\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}. \tag{3.8}$$

In the case of SI, we obtain

$$\hat{t}_{HT}^{SI} = \frac{1}{f}\sum_{k\in S} Y_k,$$

where $f = m/n$ is the sampling fraction. The variance expressions become

$$V[\hat{t}_{HT}^{SI}] = n^2\frac{(1-f)}{m}s_F^2, \quad \hat{V}[\hat{t}_{HT}^{SI}] = n^2\frac{(1-f)}{m}s_S^2, \tag{3.9}$$

where

$$s_F^2 = \frac{1}{n-1}\sum_{k\in F}(Y_k - \bar{Y}_F)^2 \quad \text{and} \quad s_S^2 = \frac{1}{m-1}\sum_{k\in S}(Y_k - \bar{Y}_S)^2.$$

The choice of π_k is crucial for the efficiency. We now compare πPS and SI from Section 3.2 on a hypothetical problem. In particular, we aim to highlight the weakness of SI. Suppose that a scientist conducts a survey for estimating the total production (in some unit) of all 100 industrial firms in a given municipality. He will survey 20 firms in detail. As a first try, the scientist thinks of SI with the HT estimator in (3.6). He then realizes that the population of firm productions is probably highly skewed because there are 3 giant firms in the municipality. This hypothetical population of

$$Y_k = \text{The production of firm } k, \quad k = 1, \ldots, 100$$

is illustrated in Figure 3.1(a). Studying the form of (3.6), the scientist realizes that this is a poor strategy because the estimator inflates each sampled element with $\pi_k = m/n$. This is only likely to work if all the Y_k's are roughly of the same size. For example if, by chance, he surveys all the giant firms the estimator will inflate all the large values, thereby overestimating the true total quite heavily. Indeed, Figure 3.1(b) shows a simulation experiment where t is estimated repeatedly based on $m = 20$ firms with SI. It is seen that the worst outcome is when all three giant firms are sampled (and inflated). To avoid this, the scientist finds a register with firms' sizes (in some measure), which he believes are correlated with the production volumes. He proceeds to design a πPS scheme which samples a firm proportional to its size. The inflated population elements Y_k/π_k accounting for the firm size are depicted in Figure 3.1(c). It is evident that they are more homogeneous than the original Y_k population (compare with Figure 3.1(a)). Consequently, the estimator of the total is more accurate as shown in the simulation experiment in 3.1(d).

More formally, suppose that the sampling probabilities are such that $\pi_k \propto Y_k$ for every element, i.e.

$$\frac{Y_k}{\pi_k} = c, \quad k \in F,$$

where c is a constant. Then, $\hat{t}_{HT}$ in (3.6) is constant and therefore $V[\hat{t}_{HT}] = 0$ for any sample size m. This is not possible in practice, because it requires knowledge of Y_k for the full population and then there is no point in sampling. However, if we set $\pi_k \propto W_k$ (the size measure) as in (3.5), and W_k is correlated with Y_k, we would expect the estimator to have a low variability.

3.3.2 Hansen-Hurwitz

For with replacement sampling, it is possible to use the Horvitz-Thompson estimator with $\pi_k = 1 - (1 - p_k)^m$. However, it is more common with the

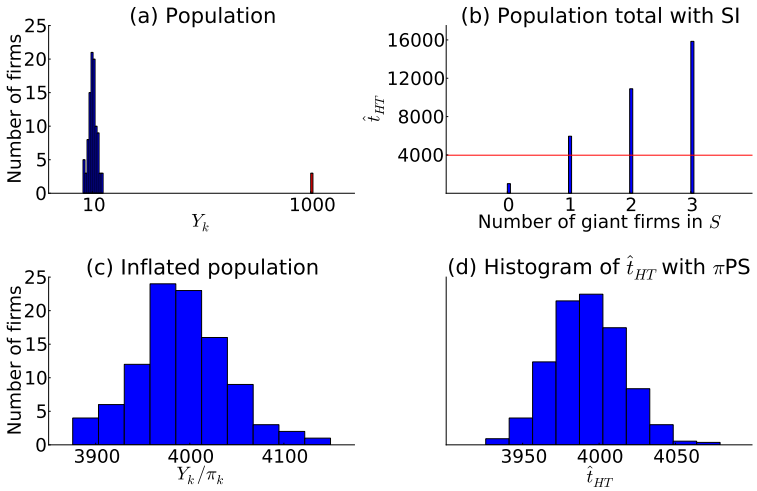

Figure 3.1: Estimating the total production of firms - The total number of firms is $n = 100$ with the total $t \approx 4000$. The estimates are based on $m = 20$ firms. Panel (a) shows the histogram of (synthetic) data of the firms' production Y_k. The blue part corresponds to regular sized firms whereas the red denotes the giant firms. Panel (b) shows a simulation of the HT estimator with SI, where the outcome is divided into the number of giant firms included in the sample. The true total is marked by the horizontal red line. Panel (c) shows the inflated population Y_k/π_k, where $\pi_k \propto W_k$ and W_k is a measure of the firm size. Finally, panel (d) shows a simulation of the HT estimator with πPS (with weights W_k).

Hansen-Hurwitz (HH) estimator (Hansen and Hurwitz, 1943),

$$\hat{t}_{HH} = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_{u_i}}{p_{u_i}} \quad \text{with } u_i \text{ in (3.3).} \tag{3.10}$$

This estimator is unbiased for t and the variance is

$$V[\hat{t}_{HH}] = \frac{1}{m}\sum_{k\in F}\left(\frac{Y_k}{p_k} - t\right)^2 p_k. \tag{3.11}$$

An unbiased estimator of the variance is

$$\hat{V}[\hat{t}_{HH}] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{Y_{u_i}}{p_{u_i}} - \hat{t}^{HH}\right)^2. \tag{3.12}$$

Furthermore, the choice of p_k is identical to that of π_k in Section 3.3.1.

3.3.3 Incorporating auxiliary information

The use of auxiliary information about the population for improving the precision of estimates is characteristic of survey sampling. We have already seen one example through proportional-to-size sampling in Section 3.2.2, where a measure W_k of the size of Y_k was used in the sampling stage. In this section, the auxiliary information is instead used in the estimation stage.

Let Y_k^0 denote an approximation of Y_k and decompose the population total as

$$t = \sum_{k\in F} Y_k^0 + \sum_{k\in F} D_k, \quad D_k = Y_k - Y_k^0.$$

It is assumed that $\sum_{k\in F} Y_k^0$ is known prior to sampling. The *Difference Estimator* (DE) is obtained by estimating $\sum_{k\in F} D_k$ with some sampling scheme coupled with an estimator. One example is SI combined with the Horvitz-Thompson, in which case

$$\hat{t}_{DE} = \sum_{k\in F} Y_k^0 + \hat{d}_{HT}, \quad \hat{d}_{HT} = \frac{1}{f}\sum_{k\in S} D_k.$$

It is straightforward to show that

$$E[\hat{t}_{DE}] = t, \quad V[\hat{t}_{DE}] = V[\hat{d}_{HT}] \quad \text{and} \quad \hat{V}[\hat{t}_{DE}] = \hat{V}[\hat{d}_{HT}],$$

where the variance expressions are given by (3.9) but for the D_k population.

In Section 3.3.1 we argue that simple random sampling is an inefficient strategy because the population can be heterogeneous (recall the firm production example). This problem is solved by the difference estimator, because it transforms the original population into one where the elements are roughly of the same size, thereby avoiding the need for proportional-to-size sampling.

4. Inference

The main topic of the thesis is inference in the presence of many observations. This is achieved by modern Markov Chain Monte Carlo (MCMC) methods combined with survey sampling techniques. This chapter serves the purpose of describing the basics of MCMC and some recent advances, which in the next chapter are merged with ideas from survey sampling.

Throughout the thesis the focus is on Bayesian inference and therefore the first section is devoted to the *Bayesian paradigm*.

4.1 The Bayesian paradigm

In the Bayesian approach to inference, a scientist regards the *state of nature* θ as a random variable. θ can represent, for example, a hypothesis, a model parameter or a future prediction of the system under study. The state of nature does not have to be random per se, but is treated so because it is unknown to the scientist. In the Bayesian paradigm, the belief about unknown quantities are expressed by *subjective* probability measures. Typically, the scientist has some belief about θ before conducting an experiment, which is expressed by $p(\theta)$, known as *the prior* distribution. The prior distribution reflects a *subjective* belief and can vary between scientists depending on e.g. previous experience. To learn about θ, the scientist collects data y and summarizes the information about θ given by the data by the *likelihood* function $p(y|\theta)$. The scientist now *objectively* updates his belief about θ by combining the likelihood and the prior via *Bayes theorem* to obtain the *posterior* distribution $\pi(\theta) = p(\theta|y)$,

$$\pi(\theta) \quad = \quad \frac{p(y|\theta)p(\theta)}{p(y)} \quad \text{with } p(y) = \int p(y|\theta)p(\theta)d\theta. \tag{4.1}$$

A major distinction from classical inference, besides the belief that θ is random, is that Bayesian inference has a *conditional* viewpoint. This means that all statements regarding θ are conditional on the *actual observed data*. This is in sharp contrast to classical inference, where inferential statements about θ are considered over *all possible data* that could have been observed (but were not observed!). After the data de facto have been observed, statements based on data that were *not* observed can give quite peculiar and non-intuitive results (Berger, 1985, Section 1.6.3).

Some people discard Bayesian methods because of the subjectivity imposed by the prior distribution. The ideal Bayesian analysis incorporates a proper prior investigation and, moreover, as more data becomes available the likelihood dominates the influence of the prior. It can also be argued that perhaps the most subjective component in a model is the choice of the model itself. Scientists would then defend themselves by saying that one makes sure that the model is consistent with the data. The same reasoning should then also apply for a prior distribution because it is a part of the model.

Finally, it should be mentioned that Bayesian inference is intimately connected with decision theory. The need of subjective probabilities to quantify uncertainty is a consequence of some axioms on rational behavior. See for example Bernardo and Smith (2009).

4.2 Bayesian computations

Computations in Bayesian statistics are often very complex for the following reasons. First, the posterior $\pi(\theta)$ is a known distribution only for a few choices of $p(y|\theta)$ and $p(\theta)$, so called *conjugate priors*. A conjugate prior for the likelihood gives a closed-form posterior which has the same distribution as the prior (with updated parameters). Second, the analysis post estimation often involves averaging a function $h(\theta)$ over the posterior, a task that can be very tedious. One example of averaging over the posterior appears when a new observation $\tilde{y}$ is to be predicted, through the *predictive distribution*

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)p(\theta|y)d\theta. \tag{4.2}$$

When analytical derivations are not possible there are a number of ways to proceed. The simplest is to approximate the posterior, usually by a Gaussian distribution. Other, nowadays increasingly more popular, methods are Variational Bayes (VB) (Ormerod and Wand, 2010) and Approximate Bayesian Computations (ABC) (Marin *et al.*, 2012). The first is an optimization method whereas the latter is a simulation method. A great advantage with VB is that it is usually very fast in finding the posterior mode, however, it often gives a poor approximation of the posterior spread. ABC is useful when the likelihood is intractable but one drawback is that there are currently no theoretical results to assess the error in the approximate posterior. Finally, simulations methods through Markov Chain Monte Carlo (MCMC) algorithms are widely common.

The use of MCMC methods is desirable as it gives the full posterior distribution without approximations. Furthermore, model diagnostics, model selection, and model regularization through variable selection, etc. may be obtained

within a single run of the sampler which enhances its attractiveness. The drawback of MCMC methods is that they are computationally expensive and this is particularly true for large data sets. One of the main contribution of this thesis is to develop MCMC algorithms and a theoretical framework to deal with this specific problem.

4.3 Markov chain Monte Carlo

Posterior samples $\{\theta^{(j)}\}_{j=1}^{N}$ are often used for computing high dimensional integrals, e.g. posterior moments or predictions as in (4.2). By the strong law of large numbers, if the samples are iid.,

$$\frac{1}{N}\sum_{j=1}^{N} h(\theta^{(j)}) \xrightarrow{a.s.} E[h(\theta)]. \tag{4.3}$$

Here a.s. denotes almost sure convergence which is the strongest type of convergence for random variables. However, it is only in very low dimensions that it is possible to sample iid draws. With more than a couple of parameters one must resort to MCMC methods which is the topic of this section.

MCMC generates a sequence of parameters with *Markov dependence* so that

$$\{\theta^{(j)}\}_{j\geq J}^{N} \text{ is distributed according to } \pi(\theta) \text{ for large enough } J.$$

The iid property is sacrificed in exchange to be able to deal with higher dimensions. Equation (4.3) still holds, although, the statistical efficiency of the estimate is reduced because of the dependence in the sequence. The efficiency of MCMC is usually evaluated by computing the variance of $\bar{\theta}$, which is the left hand side of (4.3) when h is the identity function. If the sequence of samples is weakly stationary, it can be shown that

$$NV[\bar{\theta}] \to \sigma^2(1+2\sum_{k=1}^{\infty}\rho_k),$$

where $\sigma^2 = V[\theta^{(j)}]$ and ρ_k is the auto-correlation function at lag k of the sequence. Notice that for independent sequences this implies that $V[\bar{\theta}] = \sigma^2/N$. This motivates the definition of the Effective Sample Size (ESS),

$$ESS = \frac{N}{1+2\sum_{k=1}^{\infty}\rho_k}. \tag{4.4}$$

ESS tells us how many equivalent iid draws our algorithm is sampling from the posterior. The denominator in (4.4) is called the *inefficiency factor* or the *integrated auto-correlation time*.

We now review some basic MCMC algorithms. See Brooks *et al.* (2011) for a modern treatment of MCMC.

4.3.1 The Gibbs sampler

Consider a model with a parameter vector that can be divided into K blocks,

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K).$$

The blocks can be of different sizes and scalar blocks are also allowed with this definition. In many models the full parameter vector θ does not have a closed-form posterior, but, after proper conditioning, the conditional distributions can be of known form. Other models can be augmented with parameters so that the model is simplified after conditioning on these parameters. We have already seen such an example in Section 2.2.1 for the finite mixture, where the component memberships s were introduced.

The Gibbs sampler (Geman and Geman, 1984) is a convenient solution for the problem above. It proceeds by sampling from each conditional distribution sequentially. We start the chain at some arbitrary point

$$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_K^{(0)}).$$

Then, for $j = 1, \ldots, N$, we repeat the following steps

1. $\theta_1^{(j)} \sim \pi(\theta_1^{(j-1)} | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_K^{(j-1)})$,
2. $\theta_2^{(j)} \sim \pi(\theta_2^{(j-1)} | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots \theta_K^{(j-1)})$,

   $\vdots$

K. $\theta_K^{(j)} \sim \pi(\theta_K^{(j-1)} | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{K-1}^{(j)})$.

In many models, some of the blocks will not have a closed-form solution. In the next section the Metropolis-Hastings algorithm is introduced, which is a powerful tool when a closed-form solution does not exist. The Metropolis-Hastings algorithm can be used within the Gibbs sampler for any block that cannot be straightforwardly simulated. The sampler then goes under the name *Metropolis-within-Gibbs*.

4.3.2 Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm was first proposed by Metropolis *et al.* (1953) and further extended in Hastings (1970). As previously mentioned, the M-H is useful when a closed-form solution for $\pi(\theta)$ does not exist, not even after proper conditioning.

We now consider the *random walk* Metropolis-Hastings which constructs the Markov chain as follows. Start the chain at $\theta_c = \theta^{(0)}$, where θ_c denotes the current state of the chain. Now, for $j = 1, \ldots, N$, repeat

1. Generate a candidate θ_p from a *proposal* distribution $q(\theta|\theta_c)$.

2. Compute

$$\alpha = \min\left(1, \frac{\pi(\theta_p)/q(\theta_p|\theta_c)}{\pi(\theta_c)/q(\theta_c|\theta_p)}\right). \tag{4.5}$$

3. Set

$$\theta^{(j)} = \begin{cases} \theta_p, \text{ with probability } \alpha, \\ \theta_c, \text{ with probability } 1-\alpha, \end{cases} \quad \text{and} \quad \theta_c = \theta^{(j)}.$$

It should be noted that, fortunately, the normalizing constant of the posterior cancels in (4.5) and is therefore not required to be tractable.

It can be proved that the draws obtained from the M-H algorithm have $\pi(\theta)$ as stationary distribution. This is in fact also true for the Gibbs sampler because it is a special case of a M-H, namely a block-at-a-time M-H (Chib and Greenberg, 1995), where a proposed sample for a given block is accepted with probability 1.

The efficiency of the M-H (recall the discussion in Section 4.3) depends crucially on the proposal distribution q. Figure 4.1 illustrates this fact for an inefficient proposal (a) and an efficient one (b). It is seen that the key feature of an efficient algorithm is that it generates less dependent draws, thereby allowing for a more effective exploration of the parameter space. To construct an efficient algorithm, it is possible to consider a *tailored* proposal. Broadly speaking, such a proposal mimics the posterior and therefore, intuitively, the sampler will often propose draws where the posterior mass is located. This strategy is explored in Paper I.

4.4 Advanced topics in Markov chain Monte Carlo

This section present recent advances in MCMC methods. We introduce two classes of MCMC algorithms which will be among the key ingredients in our approach to developing algorithms for the large data setting.

4.4.1 Pseudo-marginal MCMC

The M-H algorithm requires that we can evaluate $\pi(\theta) \propto p(y|\theta)p(\theta)$ in the acceptance ratio (4.5). For many models it is not possible to compute $p(y|\theta)$ directly. The pseudo-marginal algorithm (Andrieu and Roberts, 2009) provides a tool to handle such situations. The idea is to augment the parameter space with auxiliary variables which serve the purpose of *estimating* the intractable likelihood. The Markov chain is then constructed on the *augmented*

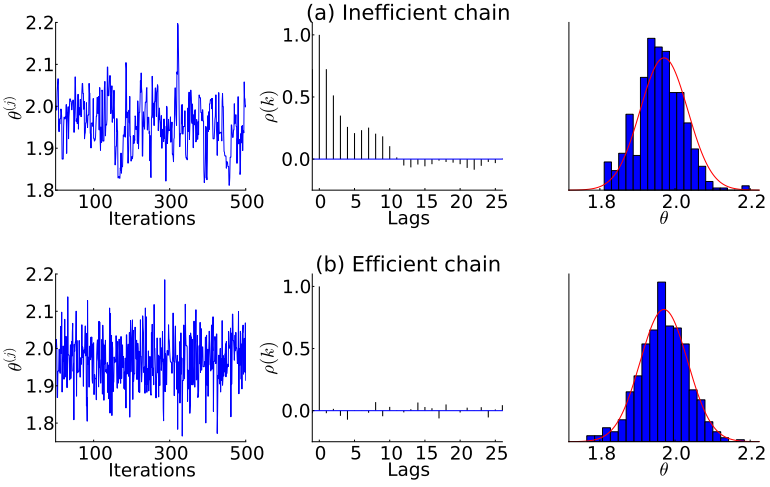

Figure 4.1: Illustrating efficiency of MCMC algorithms - Estimating a one parameter toy model with an inefficient proposal (panel (a)) and an efficient proposal (panel (b)). Panel (a): For the inefficient proposal, the left figure shows the draws from the posterior for 500 iterations. The middle figure shows the autocorrelation function of the draws. The right figure shows a normalized histogram of the draws together with the true posterior (red line). Panel (b) shows the same quantities for the efficient proposal case.

space to obtain an algorithm with similar properties as the *marginal algorithm*, i.e. the standard M-H algorithm for θ. Andrieu and Roberts (2009) prove that if the likelihood estimator is *unbiased* then the true posterior is the stationary distribution of the algorithm.

Assume the existence of an unbiased likelihood estimator $\hat{p}(y|\theta,u)$ such that

$$p(y|\theta) = \int \hat{p}(y|\theta,u)p(u)du, \tag{4.6}$$

where $u \sim p(u)$ are the auxiliary random variables used to estimate the likelihood. Define

$$\tilde{\pi}(\theta,u) = \frac{\hat{p}(y|\theta,u)p(u)p(\theta)}{p(y)}, \text{ with } \quad p(y) = \int_\theta \int_u \hat{p}(y|\theta,u)p(u)p(\theta)dud\theta,$$

on the augmented space (θ,u). By the unbiasedness condition in (4.6),

$$\begin{aligned} \int \tilde{\pi}(\theta,u)du &= \frac{(\int_u \hat{p}(y|\theta,u)p(u)du)\, p(\theta)}{\int_\theta (\int_u \hat{p}(y|\theta,u)p(u)du)\, p(\theta)d\theta} \\ &= \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} \\ &= \pi(\theta). \end{aligned}$$

The augmented posterior is thus a proper density with the correct marginal (if the estimator is unbiased). Commonly used methods to estimate the likelihood unbiasedly are importance sampling, or the particle filter in state space models (Andrieu *et al.*, 2010). In both cases u is interpreted as the random variables generating the particles.

Draws from $\tilde{\pi}(\theta,u)$ are obtained similarly as in the M-H algorithm in Section 4.3.2, but on the augmented space (θ,u). That is, for $j = 1,\ldots,N$, repeat

1. Generate candidates $u_p \sim p(u)$ and $\theta_p \sim q(\theta|\theta_c)$.

2. Compute

$$\alpha = \min\left(1, \frac{\hat{p}(y|\theta_p,u_p)p(\theta_p)/q(\theta_p|\theta_c)}{\hat{p}(y|\theta_c,u_c)p(\theta_c)/q(\theta_c|\theta_p)}\right). \tag{4.7}$$

3. Set

$$(\theta^{(j)},u^{(j)}) = \begin{cases} (\theta_p,u_p), \text{ with probability } \alpha, \\ (\theta_c,u_c), \text{ with probability } 1-\alpha, \end{cases}$$

and $(\theta_c,u_c) = (\theta^{(j)},u^{(j)})$.

The expression in (4.7) is similar to (4.5), but with an estimated likelihood in place of the true likelihood.

Note that a highly variable likelihood estimator in (4.7) can easily produce extreme overestimates of the true likelihood which may cause the M-H sampler to get stuck for long spells. More generally, the efficiency (recall Figure 4.1) of the resulting Markov chain when the true likelihood is replaced by an estimate will depend on the variance of the estimator. Increasing the variance (by less particles) results in reduced computing time per iteration, but gives a Markov chain with higher inefficiency. On the other hand, decreasing the variance (by more particles) increases the computing time, but gives a more efficient chain. The trade-off between computing time and efficiency has been investigated in Pitt *et al.* (2012) and Doucet *et al.* (2015), arriving at the conclusion that one should choose the number of particles such that the standard deviation of the log-likelihood estimator is around 1.

4.4.2 Delayed acceptance MCMC

The delayed acceptance MCMC is a two stage sampler which was initially developed to solve computationally expensive inverse problems (Christen and Fox, 2005). Let $p^*(y|\theta)$ denote a computationally cheap approximation of the expensive likelihood. The approximation can, for example, be obtained by linearizing the model. The delayed acceptance MCMC uses the approximate likelihood to first test if a proposal has a good chance of being accepted. If so, the expensive likelihood is evaluated, otherwise a new proposal is tested. The algorithm speeds up on the standard MCMC because it avoids the likelihood evaluation if the draw is unlikely to be accepted.

The algorithm proceeds by repeating, for $j = 1, \ldots, N$,

1. Generate a candidate $\theta' \sim q(\theta|\theta_c)$.

2. Compute

$$\alpha_1 = \min\left(1, \frac{p^*(y|\theta')p(\theta')/q(\theta'|\theta_c)}{p^*(y|\theta_c)p(\theta_c)/q(\theta_c|\theta')}\right).$$

3. Set

$$\theta_p = \begin{cases} \theta' \text{ with probability } \alpha_1, \\ \theta_c \text{ with probability } 1-\alpha_1, \end{cases}$$

4. Compute

$$\begin{aligned} \alpha_2 &= \min\left(1, \frac{p(y|\theta_p)p(\theta_p)/q_2(\theta_p|\theta_c)}{p(y|\theta_c)p(\theta_c)/q_2(\theta_c|\theta_p)}\right), \\ q_2(\theta|\theta_c) &= \alpha_1 q_1(\theta|\theta_c) + r(\theta_c)\delta_{\theta_c}(\theta), \end{aligned} \tag{4.8}$$

where δ is the Dirac delta function and

$$r(\theta_c) = 1 - \int \alpha_1 q_1(\theta|\theta_c) d\theta.$$

5. Set

$$\theta^{(j)} = \begin{cases} \theta_p, \text{ with probability } \alpha_2, \\ \theta_c, \text{ with probability } 1-\alpha_2, \end{cases} \quad \text{and} \quad \theta_c = \theta^{(j)}.$$

Note that when step 3 rejects, then $\alpha_2 = 1$ and the expensive likelihood is not evaluated.

It can be shown that the delayed acceptance generates draws from the $\pi(\theta)$, although, it is less efficient than the standard M-H. Intuitively, the efficiency of the algorithm should increase as the approximation of the likelihood improves. We provide more insights on this point when we take the delayed MCMC to the large data setting.

5. Towards scalable inference

Now the time has come to outline how survey sampling and modern MCMC algorithms are combined to solve our main problem. We define the computational bottleneck that is of main interest in the thesis. We then explain the general idea and make the connections to Chapter 3 and Chapter 4 transparent.

Before starting, we defend one obvious criticism regarding MCMC as an inferential tool for big data. The perhaps greatest advantage of MCMC is the appealing assessment of the uncertainty of the parameters through the full posterior distribution. One may argue that, if the data is very large, this uncertainty is not of any practical value (it is negligible). Certainly, the mode is still of interest but this can be achieved using faster algorithms (e.g. variational Bayes). We argue that large data will always allow for more complex models in terms of more variables. Therefore, given enough structure of the model, posterior uncertainty will never vanish. Another point we make, that is often missed, is that what is considered as large data should really be model dependent. For example, for a computationally complex model, then perhaps $10,000$ observations should be regarded as large, due to the time complexity of the estimation. Posterior uncertainty will surely be of interest in this case.

5.1 The computational bottlenecks of MCMC

As we have seen, MCMC algorithms require a likelihood evaluation in each iteration. In practice, the number of iterations can be quite large, as the algorithm needs to converge to the stationary distribution and, moreover, a reasonable effective sample size is desired. For these reasons MCMC methods are widely considered to be computationally intensive.

If we consider a large data set and restrict our attention to a single iteration another computational bottleneck emerges. Suppose we have conditionally iid. data, then the likelihood

$$p(y|\theta) = \prod_{k=1}^{n} p(y_k|\theta) \tag{5.1}$$

is a massive product. Moreover, if $p(y_k|\theta)$ is costly to compute, the likelihood can be very expensive even for moderately large n. The product in (5.1) re-

quires n evaluations, hence it is said that the computational *complexity* is $O(n)$. The thesis develops algorithms to improve on this complexity.

5.2 The general idea

Paper I approaches the large data problem differently. It is the formulation of the model itself, as discussed in Section 2.2.3, that allows MCMC inference in the presence of many subjects. However, it was even then only possible to estimate the model with a fraction of the data. This raised and motivated the problem addressed in the rest of the thesis (Paper II-IV), namely: *how do we adapt the MCMC algorithm to deal with large data sets?* We now turn our attention to this specific problem.

Consider (5.1) in log-scale, i.e. the log-likelihood

$$\log p(y|\theta) = \sum_{k=1}^{n} \log p(y_k|\theta),$$

and define,

$$l(\theta) = \sum_{k=1}^{n} l_k(\theta) \quad \text{with} \quad l_k(\theta) = \log p(y_k|\theta).$$

For a given θ,

$$l_k(\theta) = \text{The log-likelihood contribution of observation } k, \quad k = 1, \ldots, n$$

is a finite population. The $l_k(\theta)$ correspond to the Y_k in Section 3.1, with a population total $t = l(\theta)$. The log-likelihood contributions may vary significantly across observations, as was the case with the productions of firms in Section 3.3.1. It is therefore crucial to find proxies for the $l_k(\theta)$'s to employ in the sampling stage (Paper II) or in the estimation step (Paper III and IV). A significant portion of the thesis is devoted to developing computationally efficient proxies.

The overall idea to speed up computations should now be obvious: we aim to use efficient sampling schemes and estimators to estimate $l(\theta)$ based on a sample of m observations. We highlight that the methods developed in the thesis are more general than the case of iid. observations in (5.1). We require that the likelihood decomposes as a product where each term depends on a unique piece of data. This is typically the case in many models, at least after conditioning on parameters of interest. For example, in a panel data model with individual specific random effects the likelihood factorizes as a product of the individuals. The individuals are then considered to be the population elements.

The next two sections outlines how the estimated log-likelihood is used in the MCMC methods presented in Section 4.4.

5.2.1 The pseudo-marginal MCMC approach

The pseudo-marginal algorithm in Section 4.4.1 was initially developed to deal with an intractable likelihood. In the large data case, the likelihood is indeed tractable but is considered to be too expensive to evaluate. We therefore propose to estimate the likelihood based on a subsample and use the estimate within a pseudo-marginal framework.

The analogy to the material as presented in Section 4.4.1 is as follows. The u's representing the particles in the pseudo-marginal algorithm now become the u's underlying our sample scheme in Section 3.1. The theoretical results regarding the efficiency of the pseudo-marginal MCMC suggested that the number of particles should be chosen so that the standard deviation of the log-likelihood estimator is around 1. In our setting, this translates to choosing the sample size m to achieve this target.

We will focus the estimation effort on the log-likelihood rather than the likelihood. This has several advantages, the most important are that we can adopt well-studied survey sampling methodology and, moreover, we can use the simple rules in Pitt *et al.* (2012) and Doucet *et al.* (2015) to choose the optimal sample size.

We use the Hansen-Hurvitz (HH) estimator in (3.10) and with replacement sampling to estimate the log-likelihood. This greatly facilitates the derivation of the theoretical framework compared to the case of sampling without replacement. Denote the HH estimator of the log-likelihood by $\hat{l}_m$ and note the dependence on θ, which we often suppress for notational reasons. By the properties of HH, $E[\hat{l}_m] = l$, however

$$E[\exp(\hat{l}_m)] \neq \exp(l),$$

thus not fulfilling the unbiasedness condition in (4.6). Suppose that $\hat{l}_m$ is normally distributed with

$$E[\hat{l}_m] = l \quad \text{and} \quad V[\hat{l}_m] = \sigma^2.$$

Then, since $\exp(\hat{l}_m)$ is log-normally distributed,

$$E[\exp(\hat{l}_m)] = \exp(l + \sigma^2/2).$$

This motivates to introduce the *bias-corrected* estimator of the likelihood,

$$\hat{p}_m(y|\theta,u) = \exp\left(\hat{l}_m - \sigma^2/2\right). \tag{5.2}$$

The computation of σ^2 is through (3.11), but it is not of any value in our setting as it requires evaluation of all n observations. However, we have access to an unbiased estimator $\hat{\sigma}^2$ in (3.12), which we plug into (5.2) to obtain

$$\hat{p}_m(y|\theta,u) = \exp\left(\hat{l}_m - \hat{\sigma}^2/2\right). \tag{5.3}$$

The estimator in (5.3) is slightly biased for the true likelihood. Nevertheless, we can think of it as an unbiased estimator for a *perturbed* likelihood $p_m(y|\theta)$, i.e.

$$p_m(y|\theta) = \int \hat{p}_m(y|\theta,u)p(u)du.$$

Similar to Section 4.4.1,

$$\tilde{\pi}_m(\theta,u) = \frac{\hat{p}_m(y|\theta,u)p(u)p(\theta)}{p_m(y)}, \text{ with } \quad p_m(y) = \int \hat{p}_m(y|\theta,u)p(u)du,$$

is a proper density on the augmented space. However, the marginal posterior

$$\pi_m(\theta) = \frac{p_m(y|\theta)p(\theta)}{p_m(y)}, \tag{5.4}$$

is perturbed.

The approach taken in the thesis can be said to be *exact approximate*; we are doing *exact inference* by pseudo-marginal MCMC, but for an *approximate problem* because of the perturbation. We propose a theoretical framework to carry out MCMC in this setting for the class of estimators in (5.3). We remark that the results can straightforwardly be extended for the difference estimator (with replacement). It is proved that the perturbed posterior in (5.4) can become arbitrarily close to the true posterior $\pi(\theta)$ by increasing m. Moreover, we derive upper bounds for the approximation error of $p_m(y|\theta)$. In a number of applications we illustrate that this upper bound is small. Hence, our algorithms sample from a posterior which is very similar to the true posterior distribution.

5.2.2 The delayed acceptance MCMC aproach

The delayed acceptance approach was initially developed to deal with costly likelihood evaluations and therefore suits well for the large data setting. In the first stage of the sampler, a fraction of the data is used to determine if the proposal passes the first stage. If so, the second stage computes the full data likelihood to determine upon final acceptance. Otherwise, the costly full data likelihood evaluation is avoided.

This idea is originally proposed in Payne and Mallick (2014). However, their solution use simple random sampling, without incorporating any auxiliary information about the likelihood. We have already seen that this can be detrimental for the variance of the estimator. We explore the difference estimator in Section 3.3.3 and find a dramatic decrease in the variance compared to the estimator in Payne and Mallick (2014).

To understand the importance of the quality of the likelihood approximation, we can express α_2 in (4.8) as

$$\alpha_2 \quad = \quad \min\left(1, \frac{p^*(y|\theta_c)/p(y|\theta_c)}{p^*(y|\theta_p)/p(y|\theta_p)}\right),$$

where p^* denotes the likelihood estimate. If the approximations are very accurate then α_2 is close to 1, hence the full data likelihood is only evaluated for the draws with a very high probability of being accepted. We show in Paper IV that how close α_2 is to 1 only depends on the variance of the estimator. The lower the variance, the closer α_2 is to 1. Consequently, since our estimator has lower variability, it is also more effective in only computing the full data likelihood for draws that are likely to be accepted.

6. Summary of papers

Paper I

The motivation for the model developed in the first paper of the thesis comes from the aim of predicting firm bankruptcy using a large microeconomic data set of Swedish firms. The model of choice is a mixture-of-experts model which is generalized to a longitudinal data setting. We propose a general class of such models, with a special emphasis on discrete-time survival data suitable for our data set. The subjects are allowed to move between the different mixture components through time. This is achieved by modeling the time-varying probabilities of component membership as a function of subject-specific time-varying covariates. This allows for an interesting model interpretation and, more importantly, for manageable computations in the presence of a large number of subjects.

Each parameter in the component densities and in the mixing function is connected to its own set of covariates through a link function. We estimate the model by a Bayesian approach through a highly efficient MCMC algorithm. The proposals of the MCMC are tailored and, in addition, Bayesian variable selection is performed on all sets of covariates. This allows for model parsimony and gives insights on the importance of covariates in different parts of the model.

The proposed *dynamic mixture* dramatically improves the out-of-sample predictive density forecasts compared to a model with time-invariant mixture probabilities, i.e. a *static mixture*. Moreover, we use previous approximation results for standard (non-longitudinal) mixture-of-experts models to derive an approximation result for the dynamic mixture. This result states that the approximation error of the dynamic mixture can be made arbitrarily small by increasing the number of components.

Paper II

The main problem we are confronted with in the second paper is the computational burden of carrying out MCMC for large data sets. We propose to speed up the algorithm by estimating the log-likelihood unbiasedly on a random sub-

set of data, resulting in substantially fewer density evaluations. The random subsets are selected using efficient probability proportional-to-size schemes. The selection probability of an observation is proportional to an approximation of its contribution to the log-likelihood function. We outline three different methods to obtain the approximations, all sharing the property of being a computationally fast and accurate proxy of the log-density.

The likelihood estimate, which is slightly biased, is used within the pseudo-marginal framework. We provide a theoretical framework to carry out MCMC with a general class of approximately unbiased likelihood estimators. This theory ensures that the posterior targeted by our algorithm is within $O(m^{-1/2})$ of the true posterior, as are the posterior moments. Furthermore, the constant of proportionality in the $O(m^{-1/2})$ error bound of the likelihood is shown to be small. Moreover, the approximation errors in the posterior are illustrated to be negligible in our applications.

We propose to adapt the sample size of the algorithm along the MCMC iterations. The adaptation is such that the sampling efficiency of the MCMC is optimized for a fixed computational budget.

The method is evaluated on two examples. First, a bivariate probit model to analyze the endogenous treatment effect of holding cash on bankruptcy while controlling for other variables. Second, a Weibull regression model with random effects for discrete-time survival data. In both examples we find that our algorithm improves on the standard Metropolis-Hastings algorithm in terms of effective sample size scaled with respect to computing time.

Paper III

The drawback of the approach in Paper II is that the probability proportional-to-size schemes require an approximation of the log-likelihood for every observation. Thus, the method only improves on the standard Metropolis-Hastings for models with costly log-density evaluations. Otherwise, it is likely that the computational burden of the proxy is similar to that of evaluating the log-densities.

In the third paper we overcome this issue. Our algorithm improves on the $O(n)$ complexity of regular MCMC by operating over local data clusters instead of the full sample when computing the likelihood. We propose a simple algorithm to obtain N_C clusters, where typically $N_C << n$. A key feature of our approach is the use of the highly efficient difference estimator. The proxy for each observation in the difference estimator is obtained via a Taylor series approximation of the (log) data density around the centroid in a local data cluster. The approximation is derived for the class of generalized linear models (Nelder and Wedderburn, 1972). The local data clusters, together with the difference

estimator, allows us to estimate the likelihood with complexity $O(N_C) + O(m)$, where the latter is attributed to the subsample used for estimation.

The likelihood estimate is used in the pseudo-marginal framework. It is straightforwardly shown that the theory developed in Paper II applies for the difference estimator. Therefore, our algorithm samples from a perturbed posterior which is within $O(m^{-1/2})$ of the true posterior.

The algorithms are used for estimating a logistic regression model to predict firm bankruptcy. We document a significant speed up compared to the standard Metropolis-Hastings in terms of sampling efficiency accounting for the number of density evaluations.

Paper IV

The fourth and final paper of the thesis brings the tools developed in Paper II and III for efficient likelihood estimation into the delayed acceptance MCMC framework. Payne and Mallick (2014) propose the delayed acceptance approach to handle large data sets. However, simple random sampling without the use of auxiliary information is used in the first stage of their algorithm.

We propose to use the efficient difference estimator which incorporates auxiliary information about the full data likelihood while only operating on a sparse set of the data. We show that the resulting delayed acceptance MCMC is asymptotically more efficient compared to that of Payne and Mallick (2014).

The second stage of the delayed acceptance algorithm evaluates the full data likelihood. This makes it infeasible for data sets to large to fit in Random-Access Memory (RAM). Payne and Mallick (2014) propose to combine their approach with the consensus Monte Carlo algorithm (Scott *et al.*, 2013). As an alternative, we propose an algorithm that estimates the likelihood in the second stage based on a subsample of the data. The algorithm becomes a delayed acceptance pseudo-marginal MCMC, and we can use the theory developed in Paper II and III to quantify the errors produced in the approximate posterior. This is an interesting feature that the consensus Monte Carlo method currently lacks.

Sammanfattning

I denna avhandling studeras modeller samt skattningsmetoder som är lämpliga för stora datamängder. Avhandlingen inkluderar fyra artiklar inom dessa områden, där tonvikten läggs på att effektivisera skattningsmetoder som är vanligt förekommande inom Bayesiansk statistik.

I första artikeln utvecklas en modellklass som appliceras för att prediktera svenska företags konkurssannolikheter givet bokslutsdata. Modellen antar att ett företag tillhör ett utav flera kluster, men tillåts byta kluster under dess livstid beroende på företagets finansiella utveckling över tid. Företagens klustertillhörighet samt modellens parametrar skattas utifrån data. Traditionella skattningsmetoder tillåter inte inferens för klustertillhörighet i ett datamaterial med ett stort antal företag inom en rimlig tid. Genom ett förenklade modellantagande så möjliggörs detta.

Resten av avhandlingen utvecklar ett ramverk för att estimera modeller med s.k. Markov Chain Monte Carlo (MCMC) metoder baserat på ett mindre urval av observationer. Statistiska metoder från urvalsmetodik används för att estimera den s.k. likelihoodfunktionen på en delmängd av observationer. Ett stort fokus läggs på att implementera effektiva urvalsscheman så att likelihoodfunktionen estimeras med så liten varians som möjligt. Skattningen av likelihoodfunktionen tillämpas i kombination med moderna MCMC metoder för att uppnå skalbar inferens för stora datamängder. Ett av avhandlingens huvudbidrag är att utveckla ett teoretiskt ramverk samt implementera diverse algoritmer för att genomföra detta.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, , 697–725.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, vol. 405. John Wiley & Sons.

Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–335.

Christen, J. A. and Fox, C. (2005). MCMC using an approximation. *Journal of Computational and Graphical Statistics*, **14**, 795–810.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220.

Doucet, A., Godsill, S. and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, **10**, 197–208.

Doucet, A., Pitt, M., Deligiannidis, G. and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *To appear in Biometrika*, .

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer-Verlag.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, , 721–741.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, **14**, 333–362.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Ibrahim, J., Chen, M. and Sinha, D. (2005). *Bayesian survival analysis*. Wiley Online Library.

Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.

Jiang, W. and Tanner, M. (1999). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics*, **27**, 987–1011.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, **22**, 1167–1180.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Miller, R., Gong, G. and Muñoz, A. (1981). *Survival analysis*. Wiley New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, pp. 370–384.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, **38**, 1733–1766.

Ormerod, J. and Wand, M. (2010). Explaining variational approximations. *The American Statistician*, **64**, 140–153.

Payne, R. D. and Mallick, B. K. (2014). Bayesian big data classification: A review with complements. *arXiv preprint arXiv:1411.5653*, .

Pitt, M. K., Silva, R. d. S., Giordani, P. and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, **171**, 134–151.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model assisted survey sampling*. Springer.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E. and McCulloch, R. (2013). Bayes and big data: the consensus monte carlo algorithm. In *EFaBBayes 250 conference*, vol. 16.

Singer, J. and Willett, J. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, **18**, 155–195.