Essays in Likelihood-Based Computational Econometrics

ISBN: 978 90 361 0358 9

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 562 of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Essays in Likelihood-Based Computational Econometrics

Over de rekenkundige aspecten van de
op aannemelijkheid gebaseerde econometrie

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 23 mei 2013 om 13:30 uur

door

Tim Salimans
geboren te Utrecht

Promotiecommissie

Promotoren:	Prof.dr. Richard Paap
	Prof.dr. Dennis Fok
Overige leden:	Prof.dr. John Geweke
	Prof.dr. Dick van Dijk
	Prof.dr. Siem Jan Koopman

Acknowledgments

Over the last four years, I was very fortunate to receive much help and support in performing my dissertation research. Writing a PhD thesis can be a difficult journey, and I would like to thank everyone who helped me successfully complete it.

The first task a prospective PhD student has to complete is perhaps also the most important one: selecting a supervisor. I am very happy with the choice I made. At our first meeting, Richard was immediately enthusiastic and supportive of my ideas, while offering his own insights for improving them. This proved to be typical for our working relationship during the rest of my dissertation research. Richard was selfless in helping me, which is why he does not appear as a co-author on any of my papers. However, his influence can be seen on every page of this dissertation, and especially in Chapter 4, which came out of my Master's thesis that he supervised.

I would also like to thank my other supervisor, Dennis. His door was always open for a chat and some advice when I needed it. I also really enjoyed working together on the research presented in Chapter 3, and while teaching the Econometrie 2 undergraduate course.

In addition to my supervisors, many other great people collaborated with me on the work in this dissertation. During the TI MPhil program, Jaap Abbring hired me as a research assistant and gave me my first taste of doing real econometric research. I learned a great deal from this experience, and the result of our collaboration can be seen in Chapter 2.

During my research I was fortunate to spend three months working at Microsoft Research Cambridge, on the invitation of Thore Graepel. Although a lot of the technical aspects of the research at Microsoft were similar to the work done at the Econometric Institute, looking at things from a different perspective was very refreshing. The work in

Chapter 6 is a result of my collaboration with Thore, as well as with Ulrich Paquet. In addition to being great collaborators, both also helped me feel at home in Cambridge and made this period very enjoyable.

During my time in Cambridge I also met my co-author for the work in Chapter 5, David Knowles. Besides being a great guy, David is also a well-known expert on variational approximations for Bayesian inference, and it was great to have the opportunity to work with him.

I am grateful to the committee for being at my defense and for providing feedback on my thesis. I have known each of the committee members for some time now, and each has contributed to my understanding and interest in econometrics. John Geweke and Herman van Dijk were an important source for my interest in Bayesian econometrics, and both have contributed useful advice on my work in the past couple of years. During the MPhil phase of the TI program, both Siem Jan Koopman and Dick van Dijk have lectured me about time series models, some of which I hope is reflected in the work of Chapter 3.

During my time in Rotterdam, I shared an office with many different people. I started out with Guangyuan and Xin, who I also studied with during the TI MPhil program. Next, I spend a long time with Jorn, who I would like to thank for being a great colleague and friend. After that, I spend brief but enjoyable periods sharing offices with Niels, Harwin, and Liesbeth. It has been nice to complete the circle by sharing an office with Guangyuan yet again these past few months.

I am also grateful to all other staff and fellow PhD students who helped make my stay at Erasmus university great, and I want to explicitly mention Sjoerd, Anne, Victor, and Eran. In addition, I would like to thank everyone who studied with me in the TI MPhil, specifically Berber and Roel for forming our first study group, and Thomas and Stephen for joining me in Shanghai on one of my first academic conferences.

I would also like to thank everyone else at the Tinbergen Institute and Econometric Institute for making my PhD research possible. Furthermore, I am grateful to The Netherlands Organisation for Scientific Research (NWO) for their financial support of this dissertation.

Het schrijven van een academisch proefschrift is vaak individualistisch, en soms frustrerend. Ook kan het behoorlijk eenzijdig zijn om de hele dag over een wiskundig vraagstuk na te denken. Daarom wil ik graag al mijn vrienden bedanken voor alle afleiding en aanmoediging die jullie mij de afgelopen jaren hebben gegeven. Dankzij jullie heb ik de balans tussen werk en leven kunnen behouden.

Bart is een van die vrienden die waarschijnlijk de meeste van mijn (voorbijgaande) frustraties aan heeft moeten horen, en altijd met de nodige aanmoediging klaar stond. Het is daarom passend dat hij mij als paranimf kan helpen het laatste stapje van mijn promotie te volbrengen. Bedankt hiervoor!

De belangrijkste persoon in mijn leven is Sanne. Ik had dit niet zonder jou kunnen doen! Bedankt voor al je liefde en support. Ik kijk er naar uit om samen met jou als paranimf mijn promotie te kunnen volbrengen, en om daarna nog lang en gelukkig samen te zijn.

Verder wil ik mijn familie bedanken voor alle liefde en aanmoediging die jullie mij hebben gegeven. Pracho, Anton, Mathilde, Eva, Inge, Marcha, Magda, oma, opa, ooms en tantes, neven en nichten, ik ben er trots op dit boekje aan jullie te kunnen presenteren.

Ten slotte wil ik graag nog even stil staan bij degenen die dit proefschrift helaas niet meer zullen lezen. Jacques, Annie, Arend, meer dan wie dan ook hebben jullie mij altijd laten merken hoe trots jullie op mij waren, en hoe geïntereseerd jullie waren in mijn promotie onderzoek. Het is treurig dat ik het afgelopen jaar van jullie afscheid heb moeten nemen, en dat we nu niet samen mijn promotie kunnen vieren.

Tim Salimans
Rotterdam, 2013

Contents

Chapter 1

Introduction and outline

> The theory of probabilities is basically
> only common sense reduced to a calculus.
>
> Pierre Simon Laplace, 1812

The quote above is from Pierre Simon Laplace's introduction to his seminal work *Théorie analytique des probabilités*, in which he lays the groundwork for what is currently known as Bayesian analysis. He proceeds to describe probability theory, and statistical inference, as a method that *makes one estimate accurately what right-minded people feel by a sort of instinct, often without being able to give a reason for it.* (translation from French: Dale, 1995) This statement contains a profound truth and insight: Probability theory offers a clean and simple recipe for reasoning under uncertainty which I experienced as eye-opening when I first learned about it. As my knowledge of probability theory increased, however, I also realized that in isolation this quote presents things to be much simpler than they actually are: Reducing common sense to a calculus is extremely difficult to do well in practice. Translating our common sense into the language of probabilities takes a lot of practice, and if done accurately it often leads to a calculus without any exact solutions. It is therefore the task of statisticians and econometricians to find practical ways of reducing our common sense to calculus, and to devise smart new methods for efficiently doing the resulting calculations. This work represents my contribution towards these goals.

1.1 Probabilistic Modeling

Econometrics concerns itself with using data to learn about economic phenomena. In practice, econometricians usually study *data sets* consisting of multiple observations of a particular *dependent variable* y that is of our main interest, for example the GDP of a country, and a number of *explanatory variables* x, for example the population density and natural resource level of that country. The econometrician's goal is then to learn about the economic relationship between x and y. In itself, a list of x's and y's is not very useful: it does not give us any economic insights and it will not allow us to generalize to situations for which we do not yet have data. Learning can therefore only occur if we first provide some context to the data. In the field of econometrics, this context usually takes the form of a probabilistic model.

A model is a mathematical description of a set of (economic) hypotheses about the relationship between the variables in our data set. Economic theory might for example provide us with an idea about how the variables in x will influence the variable y. Models therefore encode our assumptions about the world before we have seen the data, and they allow us to make the data interpretable: Rather than having a list of numbers that could have arisen in an infinite number of ways, our data can now be interpreted in the context of the model. In economics, such models describe at best a rough approximation of reality. Economies are so complex that we can never hope to observe all relevant information, or to understand fully all processes that are at work. By allowing uncertainty into our models we admit that they are imperfect. Probability theory is a language for making this uncertainty more precise; it allows us to specify exactly how much and what type of uncertainty we want to incorporate in our models.

Uncertainty typically enters an econometric model in two different ways: The first is due to our uncertainty about the economic relationship between x and y, and is often called *parameter-* or *model uncertainty*. We may for example be in doubt as to whether this relationship is linear on nonlinear. Even if we are fairly certain that the relationship is linear we may not know the slopes of this linear relationship. When constructing a probabilistic model this type of uncertainty is typically captured by making the model dependent on a set of unknown *parameters* which we denote by θ. The second type of

uncertainty captures the fact that the model is fallible: Even a very good model will not capture all factors relevant to an economic process, or flawlessly describe its relationship with these factors. By allowing additional uncertainty into the model we are leaving open the possibility that something might influence y that the model does not capture. This type of uncertainty is typically represented by an *error term* denoted by ϵ and sometimes a number of *latent random variables* denoted by s. These latent variables often describe important but unobserved aspects of the countries or persons in our data set, and they are used, for example, in cases where our data is incomplete. The uncertainty in ϵ and s is made explicit by assigning to them a *distribution*, which captures how much and what type of uncertainty we place in these variables.

It is important to realize that this approach is not the only way to learn from data. In many particular cases people have used other solutions that are not based on probability theory or on any formal model. The great advantage of probabilistic modeling however is that it provides a general framework for reasoning under uncertainty. Probabilistic models are interpretable and their structure is modular, allowing researchers to mix and match elements for different applications. An additional advantage is that this approach separates the knowledge and assumptions encoded in the model from the algorithm that learns from the data. This way it is clear what part of our conclusions derives from our modeling assumptions and what part is concluded from the data.

1.2 Likelihood-Based Econometric Inference

Given the context of a probabilistic model, the goal of econometric inference is to learn from the data what parameter values are plausible. It is therefore important to think about what information is actually contained in the observed data. For the purpose of the present work, we can say that all information present in the data is described by the *likelihood function*: Our probabilistic model defines a distribution for the dependent variable y, conditional on the explanatory variables x and the unknown parameters θ, denoted by $p(y|x;\theta)$. If we fix the arguments x and y to their values observed in our data, we are left with a function of θ only, often denoted as $\mathrm{L}(\theta)$, called the likelihood function. Loosely speaking, the likelihood function tells us how well the data is explained by the model for

any given value of θ.

Although not uncontroversial, it is often claimed that the likelihood function contains all information present in the data, a statement called the *likelihood principle*. In practice, this principle states that two different data sets with the same likelihood function should lead to the same conclusions (under the same model). Formal arguments in favor of the likelihood principle were developed by several authors in 1962 (Barnard et al., 1962; Birnbaum, 1962; Savage, 1962), although the idea is much older, going back to the work of R.A. Fisher in the 1920s. The likelihood principle is based on Fisher's more primitive *conditionality principle*, which states that our conclusions should only be based on the experiments that were actually performed in generating our data, not on those that might have been performed but were not.

Considering that all (or at least most) information in our data is contained in the likelihood function, it only seems reasonable to make it the central quantity in learning from the data and in estimating the unknown model parameters θ. Indeed, likelihood-based inference procedures are currently among the dominant estimation methods used in econometrics. Although various ways of using the likelihood have been proposed, the two most popular likelihood-based estimation methods are the *method of maximum likelihood* and *Bayesian analysis*.

Using the method of maximum likelihood, the estimated values of the parameters θ are those that maximize the likelihood:

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \mathrm{L}(\theta). \tag{1.1}$$

In other words, we select the model (as described by θ) for which the observed data is most likely. The method of maximum likelihood provides a unified approach to estimation which can be applied to (almost) any data set and probabilistic model. Moreover, the maximum likelihood estimator has some desirable statistical properties such as *consistency*, *asymptotic normality*, *efficiency*, and *invariance to changes in parameterization*. This has made it the most popular method of estimation in economics.

Bayesian analysis provides us not just with a point estimate of the parameters, but with a whole *posterior distribution*, representing how probable we believe each parameter

value to be after having seen the data. In order to perform a Bayesian analysis of a problem, we first need to formulate how probable we believe each parameter value to be before having seen the data. This is done in the form of a *prior distribution*, denoted by $p(\theta)$. The posterior distribution is then obtained as

$$p(\theta|y) \propto p(y|x;\theta)p(\theta). \tag{1.2}$$

Here the likelihood plays the role of a weighting function over the hypotheses encoded by θ. The prior distribution $p(\theta)$ represents the weight we assign to each parameter value before having seen the data. The likelihood then updates these weights multiplicatively to give the weights we should assign to these parameter values after having seen the data. This multiplicative weighting by the likelihood may seem arbitrary, but it is the only coherent way to update the initial weights $p(\theta)$; any other method of updating is self-contradictory (see e.g. Jaynes and Bretthorst, 2003). In cases were the likelihood is much more informative than the prior, for example because there is a lot of data, the inference provided by Bayesian analysis is often similar to that given by the method of Maximum Likelihood.

1.3 Computational Challenges

Bayesian analysis and the method of Maximum Likelihood are both conceptually simple; their essence can be described by a single equation (equation (1.1) and (1.2) respectively). However, doing the computations described by these equations can be quite difficult. In order to compute the likelihood function $\mathrm{L}(\theta)$ we need to integrate over the error term ϵ and any latent random variables s:

$$\mathrm{L}(\theta) = \int\int p(y, \epsilon, s|x;\theta) d\epsilon ds. \tag{1.3}$$

When using Maximum Likelihood we have to maximize this likelihood function, which can be difficult. Using Bayesian analysis, we have to integrate over it in order to characterize the posterior distribution. For example, to calculate the posterior expectation of a

given function of the parameters $f(\theta)$ we need to solve the following integral:

$$\mathbb{E}[f(\theta)|y] = \int f(\theta)p(\theta|y)d\theta. \tag{1.4}$$

Both integrals (1.3) and (1.4) can often not be solved analytically.

Numerically calculating these integrals is difficult if θ, ϵ and s are of high dimension. In fact, inference in general probabilistic models has the NP-hard computational complexity (Cooper, 1990; Bacchus et al., 2003; Dagum and Luby, 1993), which means that the required amount of computational work grows exponentially with the dimension of θ, ϵ and s. In practice this makes exact inference infeasible very quickly in general probabilistic models. This means that in order to be able to perform econometric inference we either have to use convenient *conjugate* specifications, or we have to use approximations.

The most common type of approximation used in econometrics is the class of *Monte Carlo methods*, but various *deterministic approximations* also exist. More recently these two types of approximations have also been successfully combined. Over the last years, the developments in Monte Carlo methods and approximate Bayesian inference have opened up many problems to Bayesian inference and Maximum Likelihood estimation, but the associated computational difficulties are still far from solved.

A result by Dagum and Luby (1993) states that even approximating integrals (1.3) and (1.4) to a given degree of accuracy is NP-hard in the general case, which means that in fact no single inference algorithm can exist that will efficiently provide inference for all probabilistic models. Monte Carlo methods may seem to escape this trap, since the accuracy of sampling based approximations does not directly dependent on the dimensionality, but unfortunately sampling from general distributions is NP-hard in itself. Cooper (1990) states that this 'suggests that research should be directed away from the search for a general, efficient probabilistic inference algorithm, and toward the design of efficient special-case, average-case, and approximation algorithms.' This is exactly the aim of the present work: To develop approximate inference algorithms for interesting problems in economics were likelihood-based econometric inference was previously difficult or infeasible.

1.4 Outline

We pursue the goals set forth above in five separate chapters, which are all self-contained and can be read independently. The first two chapters deal with maximum likelihood estimation: They describe two different settings in which the likelihood is difficult to evaluate, and they develop deterministic methods to solve this problem. The remaining chapters take a Bayesian perspective: They represent cases where inference is difficult either because an elaborate non-conjugate model is used, or simply because there is a very large amount of data. Here we use a mixture of Monte Carlo methods and deterministic approximations in order to facilitate inference.

Chapter 2 is based on Abbring and Salimans (2013). In this chapter, we present a method for efficiently computing the likelihood of a mixed hitting-time model that specifies durations as the first time a latent Lévy process crosses a heterogeneous threshold. This likelihood is not generally known in closed form, but its Laplace transform is. Our approach to its computation relies on numerical methods for inverting Laplace transforms that exploit special properties of the first passage times of Lévy processes. We use our method to implement a maximum likelihood estimator of the mixed hitting-time model in MATLAB. We illustrate the application of this estimator with an analysis of Kennan's (1985) strike data.

Chapter 3 is based on Salimans and Fok (2013). Here we propose an efficient algorithm to perform approximate maximum likelihood estimation for non-Gaussian state-space models of very high dimension. Our approach is based on the Expectation Maximization [EM] algorithm, where we approximate the expectation step using the Expectation Propagation [EP] approach introduced by Minka (2001). Using simulation we show that this methods performs very well in terms of parameter recovery. We also successfully apply the algorithm to two empirical cases: (i) the problem of forecasting newspaper sales across individual outlets; and (ii) forecasting the outcomes of chess matches. Both applications require the use of a very large non-Gaussian dynamic model.

Chapter 4 is based on Salimans (2012). This chapter concerns regression analyses of cross-country economic growth data, which are complicated by two main forms of model uncertainty: the uncertainty in selecting explanatory variables and the uncertainty in spec-

ifying the functional form of the regression function. Most discussions in the literature address these problems independently, yet a joint treatment is essential. We present a new framework that makes such a joint treatment possible, using flexible nonlinear models specified by Gaussian process priors and addressing the variable selection problem by means of Bayesian model averaging. Using this framework, we extend the linear model to allow for parameter heterogeneity of the type suggested by new growth theory, while taking into account the uncertainty in selecting explanatory variables. Controlling for variable selection uncertainty, we confirm the evidence in favor of parameter heterogeneity presented in several earlier studies. However, controlling for functional form uncertainty, we find that the effects of many of the explanatory variables identified in the literature are not robust across countries and variable selections.

Chapter 5 is based on Salimans and Knowles (2013). Here we propose a general algorithm for approximating nonstandard Bayesian posterior distributions. The algorithm minimizes the Kullback-Leibler divergence of an approximating distribution to the intractable posterior distribution. Our method can be used to approximate any posterior distribution, provided that it is given in closed form up to the proportionality constant. The approximation can be any distribution in the exponential family or any mixture of such distributions, which means that it can be made arbitrarily precise. Several examples illustrate the speed and accuracy of our approximation method in practice.

Chapter 6 is based on Salimans et al. (2012). In this chapter, we propose a model for learning preference rankings for the purpose of making product recommendations. The model allows us to learn from pairwise preference statements or from (incomplete) rankings over more than two items. We present two algorithms for performing inference in this model, both with excellent scaling in the number of users and items. The superior predictive performance of the new method is demonstrated on the well-known sushi preference data set. In addition, we show how the model can be used effectively in an active learning setting where we select only a small number of informative items for learning.

Finally, chapter 7 contains a summary of this work in Dutch.

Chapter 2

The likelihood of mixed hitting times

Joined work with Jaap Abbring

2.1 Introduction

Mixed hitting-time (MHT) models are mixture duration models that specify durations as the first time a latent stochastic process crosses a heterogeneous threshold. They are of substantial interest because they can be applied to the analysis of optimal stopping decisions by heterogeneous agents. In particular, they can be applied to problems that do not lead to the mixed proportional hazards model, Lancaster's (1979) and Vaupel et al.'s (1979) popular extension of the Cox (1972) proportional hazards model. Examples include models of job durations, marriage durations, and the entry and exit of firms that are driven by Brownian motions and more general persistent processes. First hitting time duration models are also increasingly popular in statistics for their structural and descriptive appeal (Lee and Whitmore, 2006).

This chapter considers likelihood-based empirical methods for an MHT model in which the latent process is a spectrally-negative Lévy process, a continuous-time process with stationary and independent increments and no positive jumps, and the threshold is proportional in the effects of observed regressors and unobserved heterogeneity. Spectrally-negative Lévy processes include Brownian motions with linear drifts and Poisson processes compounded with negative shocks as well-known special cases. Following empirical practice with mixture duration models such as the mixed proportional hazards

model, we focus on parametric MHT models, and propose flexible parameterizations that can approximate arbitrary functional forms by increasing the number of parameters. The main obstacle in applying standard parametric likelihood methods is that, in general, we have no explicit expression for the MHT model's likelihood. However, an explicit expression for its Laplace transform is generally available. Our approach to likelihood computation exploits this.

We adapt numerical methods for the inversion of the Laplace transforms of the first hitting times of Lévy processes to compute the conditional density and survival function implied by the MHT model. In turn, these are used to construct a likelihood for independently censored duration data. In the special case that the latent process is a Brownian motion, the likelihood can be explicitly expressed as a mixture of inverse Gaussian densities and survival functions. Therefore, we can use this special case as a benchmark for evaluating the quality of our procedure for computing the likelihood. We show that the numerical inversion that is required in the general case is sufficiently fast and precise to make maximum likelihood estimation feasible even if no explicit expression of the likelihood is available.

We implement a maximum likelihood estimator that uses this computational strategy in MATLAB, and illustrate its application with a reconsideration of Kennan's (1985) empirical analysis of US contract strike durations. Our strategy for computing the MHT model's likelihood can also be used to implement other likelihood-based empirical methods. For example, it can be combined with data augmentation and Markov chain Monte Carlo techniques to implement Bayesian estimators of the MHT model.

Abbring (2012) presented the MHT model studied in this chapter, analyzed its empirical content, and highlighted its close relation to optimal stopping problems in economics. This chapter operationalizes this model by providing and analyzing feasible methods for computing its likelihood and its maximum likelihood estimator.

The remainder of this chapter is organized as follows. Section 2.2 reviews the MHT model and the corresponding characterization of the data presented in Abbring (2012). Section 2.3 present a method for the computation of this model's likelihood and its derivatives. Section 2.4 presents flexible model parameterizations and discusses the implementation of a maximum likelihood estimator. Section 2.5 applies this estimator to strike data.

Section 2.6 concludes.

2.2 Mixed Hitting-Time Model

2.2.1 Specification

We model the distribution of a random duration T conditional on observed covariates X by specifying T as the first time a real-valued Lévy process $\{Y\} \equiv \{Y(t); t \geq 0\}$ crosses a threshold that depends on X and some unobservables V.

A Lévy process is the continuous-time equivalent of a random walk: It has stationary and independent increments. Bertoin (1996) provides a comprehensive analysis of Lévy processes. Formally, we have

Definition 1. *A* Lévy process *is a stochastic process* $\{Y\}$ *such that the increment* $Y(t + \Delta) - Y(t)$ *is independent of* $\{Y(\tau); 0 \leq \tau \leq t\}$ *and has the same distribution as* $Y(\Delta)$*, for every* $t, \Delta \geq 0$*.*

We take $\{Y\}$ to have right-continuous sample paths with left limits. Note that Definition 1 implies that $Y(0) = 0$ almost surely.

An important example of a Lévy process is the scalar Brownian motion with drift, in which case $Y(\Delta)$ is normally distributed with mean $\mu\Delta$ and variance $\sigma^2\Delta$, for some scalar parameters $\mu \in \mathbb{R}$ and $\sigma \in [0, \infty)$. The Brownian motion is the single Lévy process with continuous sample paths. In general, Lévy processes may have jumps. Examples are compound Poisson processes, which have independently and identically distributed jumps at Poisson times. More generally, the jump process $\{\Delta Y\}$ of a Lévy process $\{Y\}$ is a Poisson point process with characteristic measure Υ such that $\int \min\{1, x^2\}\Upsilon(dx) < \infty$, and any Lévy process $\{Y\}$ can be written as the sum of a Brownian motion with drift and an independent pure-jump process with jumps governed by such a point process (Bertoin, 1996, Chapter I. Theorem 1). The characteristic measure of $\{Y\}$'s jump process is called its *Lévy measure* and, together with the drift and variance parameters of its Brownian motion component, fully characterizes $\{Y\}$'s distributional properties.

Throughout the chapter, we will focus on spectrally-negative Lévy processes. These are Lévy processes of which the characteristic measure Υ has negative support, *i.e.* Lévy

processes without positive jumps. Let $\{Y\}$ be such a process. Then, the (proportional) mixed hitting-time (MHT) model specifies that T is the first time that $Y(t)$ crosses $\phi(X)V$, or

$$T = \inf\{t \geq 0 : Y(t) > \phi(X)V\}, \tag{2.1}$$

for some observed covariates X with support $\mathcal{X} \in \mathbb{R}^K$, measurable function $\phi : \mathcal{X} \mapsto (0, \infty)$, and nonnegative random variable V, with (X, V) independent of $\{Y\}$. We use the convention that $\inf \emptyset \equiv \infty$; that is, we set $T = \infty$ if $\{Y\}$ never crosses $\phi(X)V$. The assumption that there are no positive jumps greatly facilitates the analysis of hitting times, because it excludes that the process jumps across the threshold.

The factor V is interpreted as an unobserved individual effect and is assumed to be distributed independently of X with distribution G on $[0, \infty]$. This explicitly allows for an unobserved subpopulation $\{V = \infty\}$ of *stayers*, on which $T = \infty$. In addition, there may be *defecting movers*: For some specifications of $\{Y\}$, $T = \infty$ with positive probability on $\{V < \infty\}$. The distinction between stayers and defective movers can be of substantial interest (see Abbring, 2002, for discussion). We exclude the two trivial cases in which $T = \infty$ almost surely, the case in which the population consists of only stayers ($\Pr(V < \infty) = 0$) and the case in which all movers defect ($\{Y\}$ is nonpositive). For expositional convenience only, we also assume that $\Pr(V = 0) = 0$. Abbring (2012) provides further discussion.

2.2.2 Characterization

The distribution of T conditional on (X, V) is fully determined, up to almost-sure equivalence, by its Laplace transform,

$$\mathcal{L}_T(s|X, V) \equiv \mathbb{E}\left[\exp\left(-sT\right) I(T < \infty)|X, V\right], \quad s \in [0, \infty),$$

with $I(\cdot) = 1$ if $\cdot$ is true, and 0 otherwise. The factor $I(T < \infty)$ makes explicit the possibility that the distribution of $T|X, V$ is defective. Note that the defect has mass $1 - \Pr(T < \infty|X, V) = 1 - \mathcal{L}_T(0|X, V)$.

Unlike the distribution of $T|(X, V)$, the Laplace transform $\mathcal{L}_T(\cdot|X, V)$ can be ex-

plicitly given for any specification of the latent process $\{Y\}$. This first requires a common probabilistic characterization of $\{Y\}$, in terms of its characteristic function. Bertoin (1996, Section VII.1) shows that

$$\mathbb{E}\left[\exp\left(sY(t)\right)\right] = \exp\left[\psi(s)t\right],$$

for all $s \in \mathcal{C}$ with nonnegative real parts, with the *Laplace exponent* ψ given by the Lévy-Khintchine formula,

$$\psi(s) = \tilde{\mu}s + \frac{\sigma^2}{2}s^2 + \int_{(-\infty,0)} \left\{\mathrm{e}^{sx} - 1 - sxI(x > -1)\right\} \Upsilon(dx). \tag{2.2}$$

Here, $\tilde{\mu} \in \mathbb{R}$ absorbs any linear drift of $\{Y\}$, $\sigma \geq 0$ is the dispersion parameter of its Brownian motion component; and Υ is the Lévy measure of its jump component, where Υ satisfies $\int \min\{1, x^2\}\Upsilon(dx) < \infty$ and has negative support. The Laplace exponent ψ of $\{Y\}$ fully characterizes its distributions, through its characteristic function $\mathbb{E}\left[\exp\left(iuY(t)\right)\right] = \exp\left[\psi(iu)t\right]$ for all $u \in \mathbb{R}$.

Equation (2.2) gives the most common parameterization of ψ. It corresponds to the Lévy-Itô decomposition of $\{Y\}$ in a Brownian motion with linear drift $\tilde{\mu}t$, a compound Poisson process with jumps in $(-\infty, -1]$, and a pure-jump martingale with jumps in $(-1, 0)$ (Bertoin, 1996, Section I.1). Alternative parameterizations arise if we decompose the jumps of $\{Y\}$ in small and large shocks in other ways. These parameterizations all have the same dispersion parameter σ and Lévy measure Υ, but have different drift parameters. For example, in the special case that $\int_0^1 x\Upsilon(dx) < \infty$, the *compensator* term for the small shocks in (2.2),

$$\int_{(-\infty,0)} sxI(x > -1)\Upsilon(dx) = \int_{(-1,0)} x\Upsilon(dx)s,$$

is a well-defined linear function of s. Therefore, in this case, we can alternatively parameterize ψ as

$$\psi(s) = \mu s + \frac{\sigma^2}{2}s^2 + \int_{(-\infty,0)} \left(\mathrm{e}^{sx} - 1\right) \Upsilon(dx), \tag{2.3}$$

where $\mu \equiv \tilde{\mu} + \int_{(-1,0)} x\Upsilon(dx)$. This includes the important special case that $\int_{(-\infty,0)} \Upsilon(dx) <$

∞, in which $\{Y\}$ is the sum of a Brownian motion with drift parameter μ and a compound Poisson process with jumps *of all sizes* in $(-\infty, 0)$. In general, any of the equivalent parameterizations of ψ can be used in the MHT model's specification, but some are numerically and statistically more convenient than others; we return to this in Section 2.4.

With ψ determined, we are ready to analyze the Laplace transform $\mathcal{L}_T(\cdot|X, V)$. The Laplace exponent, as a function on $[0, \infty)$, is continuous and convex, and satisfies $\psi(0) = 0$ and $\lim_{s\to\infty} \psi(s) = \infty$. Therefore, there exists a largest solution $\Lambda(0) \geq 0$ to $\psi(\Lambda(0)) = 0$ and an inverse $\Lambda : [0, \infty) \to [\Lambda(0), \infty)$ of the restriction of ψ to $[\Lambda(0), \infty)$. Theorem 1 of Bertoin (1996, Chapter VII) implies that (see Abbring, 2012)

$$\mathcal{L}_T(s|X, V) = \exp\left[-\Lambda(s)\phi(X)V\right].$$

The Laplace transform of the distribution of $T|X$ therefore is

$$\mathcal{L}_T(s|X) = \mathcal{L}\left[\Lambda(s)\phi(X)\right], \tag{2.4}$$

with $\mathcal{L}$ again the Laplace transform of the unobservable's distribution G.

2.2.3 A Gaussian Example

Suppose that $\{Y\}$ is a Brownian motion with general drift coefficient $\mu \in \mathbb{R}$ and dispersion coefficient $\sigma \in (0, \infty)$. Then, we have that $\psi(s) = \mu s + \sigma^2 s^2/2$, so that $\Lambda(0)$ equals $\Lambda_{\mathrm{BM}}(0) \equiv \min\{0, -2\mu/\sigma^2\}$ and $\Lambda(s)$ equals

$$\Lambda_{\mathrm{BM}}(s) \equiv \frac{\sqrt{\mu^2 + 2\sigma^2 s} - \mu}{\sigma^2}. \tag{2.5}$$

Because there are no jumps, there is no ambiguity in the treatment of small and large jumps, and this parameterization of ψ is unique. In particular, the Lévy-Khintchine representations (2.2) and (2.3) of ψ coincide, and $\mu = \tilde{\mu}$.

In this special case, for positive $\phi(X)V$, the distribution of $T|X, V$ is inverse Gaussian

(Cox and Miller, 1965, Section 5.4), with Lebesgue density

$$f_{\mathrm{BM}}(t|X,V) = \frac{\phi(X)V}{\sigma\sqrt{2\pi t^3}} \exp\left(-\frac{(\phi(X)V-\mu t)^2}{2\sigma^2 t}\right) \tag{2.6}$$

and survival function

$$\begin{aligned}\overline{F}_{\mathrm{BM}}(t|X,V) &\equiv \Pr\left(T>t|X,V\right)\\ &= \Phi\left(\frac{\phi(X)V-\mu t}{\sigma\sqrt{t}}\right) - \exp\left(\frac{2\mu\phi(X)V}{\sigma^2}\right)\Phi\left(-\frac{\phi(X)V+\mu t}{\sigma\sqrt{t}}\right).\end{aligned} \tag{2.7}$$

Here, Φ is the cumulative standard normal distribution function. If $\mu \geq 0$, then $\Lambda_{\mathrm{BM}}(0) = 0$ and the distribution of $T|X,V$ is nondefective for positive $\phi(X)V$. If $\mu < 0$, however, $\Lambda_{\mathrm{BM}}(0) = -2\mu/\sigma^2 > 0$ and the distribution of $T|X,V$ has a defect of size $1 - \exp(2\phi(X)V\mu/\sigma^2)$. Note that in this case, $\sigma = 0$ is excluded to avoid the trivial outcome that $T = \infty$ almost surely.

Either way, the MHT model (2.1) specifies a mixed inverse Gaussian distribution for $T|X$ in this special case. Mixed inverse Gaussian distributions have been used to model duration data in the statistical literature. For example, Aalen and Gjessing (2001) propose such a model with parametric mixing over the Brownian motion's drift coefficient μ. This chapter extends and adapts this literature with estimators that allow for more general latent processes and mixing distributions.

2.3 Likelihood Computation

2.3.1 Parameterization

Let ψ, ϕ and $\mathcal{L}$ be specified up to a finite vector of unknown parameters $\alpha \in \mathcal{A}$. Assume that this parameterization is one-to-one, so that α is uniquely determined by (ψ, ϕ, G). In the case that $\ln\phi(X) = \delta + X'\beta$ for some scalar intercept δ and $K \times 1$ vector of slope parameters β, for example, this requires the "rank condition" that the support $\mathcal{X}$ of X contains a nonempty open set in $\mathbb{R}^K$.

With such a parameterization, under mild additional conditions, Abbring's (2012) results imply that α is uniquely determined ("identified") from the distribution of $T|X$. In

particular, it is sufficient that

1. the scales of $\{Y\}$, $\phi(X)$, and V are appropriately normalized;

2. $\phi(X)$ is nondegenerate; and

3. either V has a finite mean or the latent process $\{Y\}$ is such that $0 < |\psi'(0+)| < \infty$.

Throughout, we assume that the first two conditions hold, and explicitly note the assumptions on $\mathcal{L}$ and ψ required to ensure that the third condition holds as well.

The first condition's scale normalizations are innocuous, but need to be carefully implemented in any estimation procedure. They are needed because the durations T implied by the first hitting-time specification (2.1) are not affected by rescaling both the latent process $\{Y(t)\}$ and the threshold $\phi(X)V$ by the same factor, nor by rescaling the threshold factors $\phi(X)$ and V without changing the threshold itself. Specifically, any two specifications $(\psi, \phi, \mathcal{L})$ and $(\tilde{\psi}, \tilde{\phi}, \tilde{\mathcal{L}})$; with $\tilde{\psi}(s) = \psi(cs)$, $\tilde{\phi} = (c/d)\phi$, and $\tilde{\mathcal{L}}(v) = \mathcal{L}(dv)$ for some $c, d > 0$; are observationally equivalent. Stated differently, if $(\psi, \phi, \mathcal{L})$ corresponds to a latent process $\{Y\}$ and threshold $\phi(X)V$; and $(\tilde{\psi}, \tilde{\phi}, \tilde{\mathcal{L}})$ corresponds to a latent process $\{cY\}$, an observed threshold factor $c\phi(X)/d$, and an unobserved threshold factor dV; then the corresponding first hitting times are the same:

$$\inf\{t \geq 0 : Y(t) > \phi(X)V\} = \inf\{t \geq 0 : cY(t) > (c/d)\phi(X)dV\}$$

Identification therefore requires that the scale of two of $\{Y\}$, $\phi(X)$ and V are normalized. The most convenient way of implementing these two normalizations depends on the chosen parameterization, and will be discussed as we go.

The second condition ensures that the threshold varies with the regressors on their support. Such variation is key to the separate identification of the latent process and hetereogeneity. Abbring (2012) provides the following simple example of two MHT models without covariates ($\phi(X) \equiv 1$) that induce the same distribution of T. Both a model in which $\{Y\}$ is a Brownian motion with drift and V is degenerate at a single threshold value (that is, without heterogeneity) and a model in which $\{Y\}$ is degenerate linear drift ($\sigma = \Upsilon = 0$) and V has an inverse Gaussian distribution lead to an inverse Gaussian distribution of T.

The third condition is reminiscent of the conditions for identifiability of the mixed proportional hazards model. Abbring (2012) provides extensive discussion.

We also require that the parameterization of $(\psi, \phi, \mathcal{L})$ is sufficiently smooth to allow for the application of standard asymptotic theory. The choice of an appropriate parameterization of ψ is particularly important. We further discuss this in the context of specific parameterizations in Section 2.4.

2.3.2 Sampling

We explicitly deal with censoring, which is a common problem in applied duration analysis. Let $\{(T_1^*, X_1), \ldots, (T_N^*, X_N)\}$ be a (complete) random sample from the distribution of (T, X) induced by the MHT model at the "true" parameter vector $\alpha_0 \in \mathcal{A}$ and some marginal distribution of X. We do not directly observe this complete sample, but only a censored version of it: $\{(T_1, D_1, X_1), \ldots, (T_N, D_N, X_N)\}$. Here, $T_i \equiv \min\{T_i^*, C_i\}$ is the observed duration and $D_i \equiv I(T_i^* \leq C_i)$ a censoring indicator, for some random censoring time C_i; $i = 1, \ldots, N$.

For expositional convenience, we focus on a simple type of independent right-censoring (Andersen et al., 1993). Assume that the complete observations (T_i^*, C_i, X_i) are independent across i and that, conditional on X_i, C_i is independent of T_i^*. That is, censoring times are not informative on the durations of interest. For example, if data are only collected for a deterministic time C_i, then C_i is trivially independent of T_i^*. The independent censoring assumption ensures that the likelihood of the observed durations T_i conditional on (C_i, X_i) only depends on the parameters α of the MHT model. We take the marginal distributions of the (C_i, X_i) to be ancillary, and focus on estimation of α_0 by maximizing this conditional likelihood.

With more general independent right censoring schemes, the resulting estimator remains a valid (but often, partial) likelihood estimator (Andersen et al., 1993). Moreover, the likelihood, and the corresponding estimator, can easily be adapted to other practically relevant sampling schemes, such as those involving interval censoring.

In the next section, we first consider the Gaussian special case. This allows us to discuss some practical details concerning normalizations in a well-understood framework

in which the likelihood can be explicitly given. Section 2.3.4 then discusses likelihood computation in the general case.

2.3.3 Gaussian Special Case

Suppose that $\{Y\}$ is a Brownian motion with drift, so that, by the analysis in Section 2.2.3, $T|X$ has a mixed inverse Gaussian distribution. Because $|\psi'(0+)| = |\mu|$ in this case, identification of α_0 can be guaranteed by either assuming that G has finite mean or that $\mu \neq 0$ (Abbring, 2012).

In this special case, the log likelihood $\ell_N(\alpha)$ of α for $(T_1, \ldots, T_N)|\{(D_1, X_1), \ldots\}$ can be constructed using the explicit expression for the density and survival functions of $T|X, V$ in (2.6) and (2.7):

$$\ell_N(\alpha) = \sum_{i=1}^{N} \ln \int \theta_{\mathrm{BM}}(T_i|X_i, v)^{D_i} \overline{F}_{\mathrm{BM}}(T_i|X_i, v) dG(v), \tag{2.8}$$

with $\theta_{\mathrm{BM}} \equiv f_{\mathrm{BM}}/\overline{F}_{\mathrm{BM}}$ the hazard rate corresponding to f_{BM}. Here, the dependence of θ_{BM} and $\overline{F}_{\mathrm{BM}}$ (through μ, σ, and ϕ) and G on the parameter vector α is kept implicit. Under standard regularity conditions, the maximizer $\hat{\alpha}_N$ of $\ell_N(\alpha)$ is a consistent and asymptotically normal estimator of α_0. The estimator's asymptotic covariance matrix can be estimated in the standard way using either the score or Hessian characterization of the Fisher information matrix. It is asymptotically efficient under the assumption that the marginal distribution of X and the censoring times carry no information on α_0.

A typical parameterization would specify $\ln \phi(X) = \delta + X'\beta$, and a mixing distribution G that has finite support $\{v_1, \ldots, v_L\}$, for some fixed $L \in \mathbb{N}$, with parameters

$$\pi_l \equiv \Pr(V = v_l) = G(v_l) - G(v_l-); \quad l = 1, \ldots, L. \tag{2.9}$$

A finite discrete specification of G is popular because of its versatility and computational convenience; it also appears naturally in Heckman and Singer's (1984) influential work on semiparametric estimation of the MPH model. With it, the log likelihood in (2.8) reduces

to

$$\ell_N(\alpha) = \sum_{i=1}^{N} \ln \sum_{l=1}^{L} \pi_l \theta_{\mathrm{BM}}(T_i|X_i, v_l)^{D_i} \overline{F}_{\mathrm{BM}}(T_i|X_i, v_l),$$

which is easy to compute using (2.6) and (2.7). In this parameterization, the two normalizations required can be implemented by setting $\delta = 0$, and setting $v_1 = 1$ with $\pi_1 > 0$. In the case that $\mu \neq 0$ is assumed, one of these normalizations can be replaced by a normalization of μ, such as $|\mu| = 1$.

The maximum likelihood estimator for the Gaussian special case of the MHT model and its asymptotic distribution are as easy to compute as, say, the maximum likelihood estimator of the mixed proportional hazards model. In particular, with a computationally convenient specification of G like the discrete example above, explicit expressions for the likelihood and its derivatives are available; and computation can proceed directly by a search for a likelihood maximizer using standard numerical methods. The Gaussian special case shares this feature with many of the models studied in the statistics literature (Lee and Whitmore, 2006). In the general Lévy case or with general heterogeneity distributions, however, such explicit expressions are not available, and maximum likelihood cannot be implemented directly. The next section develops methods for computing the maximum likelihood estimator and its asymptotic distribution in this general case.

2.3.4 General Case

In general, the density and survival function of $T|X$ are not explicitly known, but can be computed by numerically inverting their Laplace transforms. We will develop fast and effective methods for computing the likelihood; its maximizer, the ML estimator; and its derivatives by adapting existing results for inverting the Laplace transform of the first hitting time of a Lévy process. We focus on the case with a nontrivial Gaussian component: $\sigma > 0$.

Our approach is based on the work of Rogers (2000), who applies a variant of Abate and Whitt's (1992) inversion method to the problem of calculating the first-passage-time distribution of a spectrally one-sided Lévy process. This approach builds on the fact that the Laplace transform $\mathcal{L}_T(\cdot|X) = \mathcal{L}\left[\Lambda(s)\phi(X)\right]$ of $T|X$ in (2.4) represents a one-to-one

transformation of the probability density function $f(\cdot|X)$ of $T|X$,

$$\mathcal{L}\left[\Lambda(s)\phi(X)\right] = \int_0^\infty \exp(-st) f(t|X) dt. \tag{2.10}$$

The probability density function $f(\cdot|X)$ can be obtained by inverting this transformation using *Mellin's inverse formula* (see Davies, 2002),

$$f(t|X) = \frac{1}{2\pi \mathrm{i}} \lim_{N\to\infty} \int_{\gamma_N} \exp(st)\mathcal{L}\left[\Lambda(s)\phi(X)\right] ds. \tag{2.11}$$

Here, the integration is along the path $\gamma_N : u \in [-1, 1] \mapsto \gamma + \mathrm{i}Nu$, which traces out a straight line in $\mathbb{C}$, parallel to the imaginary axis from $\gamma - \mathrm{i}N$ to $\gamma + \mathrm{i}N$. Its parameter $\gamma \in \mathbb{R}$ should, in general, be chosen such that it is larger than the real part of any singularity in the Laplace transform $\mathcal{L}_T(\cdot|X)$. Because $\mathcal{L}_T(\cdot|X)$ is analytic for any s with nonnegative real part, we can choose any $\gamma \geq 0$.

The integral in (2.11) does not generally have an explicit solution, but can be efficiently approximated using numerical methods. A key complication is that our specification of $\mathcal{L}_T(\cdot|X)$ involves the inverse function Λ, which cannot generally be expressed in closed form. To circumvent this problem, we follow Rogers (2000) and integrate along the transformed path $\tilde{\gamma}_N = \psi \circ \Lambda_{\mathrm{BM}} \circ \gamma_N$ instead, which traces out a curve in $\mathbb{C}$ from $\psi\left[\Lambda_{\mathrm{BM}}\left(\gamma - iN\right)\right]$ to $\psi\left[\Lambda_{\mathrm{BM}}\left(\gamma + iN\right)\right]$ (where $\circ$ denotes function composition). Here, ψ is again the Laplace exponent of the latent process $\{Y\}$ and Λ_{BM} the inverse of the Laplace exponent of its Brownian motion component, for which (2.5) gives an explicit expression. Note that Λ_{BM} necessarily has the same dispersion parameter σ as ψ, but that its drift parameter is not uniquely pinned down (because the drift parameter of ψ depends on the way we deal with small shocks; see Section 2.2.2). Fortunately, the exact value of the drift parameter of Λ_{BM} plays no role in the argument that follows. It can generally be set to the drift parameter in the specific parameterization of ψ used; for example, $\tilde{\mu}$ in (2.2) or μ in (2.3). The MATLAB code accompanying this chapter applies to specifications of ψ with compound Poisson jumps and sets the drift parameter of Λ_{BM} equal to μ in (2.3) (see Section 2.4).

Rogers (2000) shows that the transformed path $\tilde{\gamma}_N$ is close enough to γ_N, so that we can integrate along $\tilde{\gamma}_N$ in (2.11) instead. This gives an expression for $f(\cdot|X)$ that does not

involve Λ:

$$\begin{aligned} f(t|X) &= \frac{1}{2\pi \mathrm{i}} \lim_{N\to\infty} \int_{\tilde{\gamma}_N} \exp(st)\mathcal{L}\left[\Lambda(s)\phi(X)\right] ds \\ &= \frac{1}{2\pi \mathrm{i}} \lim_{N\to\infty} \int_{\gamma_N} \exp\left[\psi\left\{\Lambda_{\mathrm{BM}}(s)\right\} t\right] \mathcal{L}\left[\Lambda_{\mathrm{BM}}(s)\phi(X)\right] d\psi[\Lambda_{\mathrm{BM}}(s)]. \end{aligned} \tag{2.12}$$

This convenient change of integration path is valid because the differences of the end points of the curves mapped out by γ_N and $\tilde{\gamma}_N$ converge to zero as N grows large. In particular, because $\sigma > 0$,

$$\left|\frac{\gamma_N(1) - \tilde{\gamma}_N(1)}{\gamma_N(1)}\right| = \left|\frac{\gamma + \mathrm{i}N - \psi\left[\Lambda_{\mathrm{BM}}\left(\gamma + \mathrm{i}N\right)\right]}{\gamma + \mathrm{i}N}\right| = \left|\frac{\psi_{\mathrm{BM}}(z_N) - \psi(z_N)}{\psi_{\mathrm{BM}}(z_N)}\right|,$$

with $z_N \equiv \Lambda_{\mathrm{BM}}\left(\gamma + \mathrm{i}N\right)$, converges to zero as $N \to \infty$, since the Laplace exponent defined in (2.2) is then dominated by the Gaussian drift term. The same result can be obtained for the other end point $\gamma_N(-1)$ if we instead take $z_N \equiv \Lambda_{\mathrm{BM}}\left(\gamma - \mathrm{i}N\right)$.

Following Abate and Whitt (1992), we can apply the trapezoidal rule to approximate (2.12) with the infinite sum

$$\begin{aligned} S_\infty(t|X) &\equiv \frac{h}{2\pi \mathrm{i}} \sum_{r=-\infty}^{\infty} g(t, r|X), \text{ where} \\ g(t, r|X) &\equiv \exp\left\{\psi\left[\Lambda_{\mathrm{BM}}\left(\gamma + \mathrm{i}rh\right)\right]t)\right\} \mathcal{L}\left[\Lambda_{\mathrm{BM}}(\gamma + \mathrm{i}rh)\phi(X)\right] \frac{d}{ds}\psi\left[\Lambda_{\mathrm{BM}}(s)\right]\Big|_{\gamma+\mathrm{i}rh}, \end{aligned} \tag{2.13}$$

where h is the step-size used with the trapezoidal rule. Although simple, the trapezoidal rule is an effective approximation rule for the current integration problem since the integrant $g(t, r|X)$ is a nearly periodic function in r. Abate and Whitt (1992) discuss the error introduced by this integral approximation and they give error bounds for the inversion of Laplace transforms of general CDF's. Importantly, the approximation can be made arbitrarily precise by reducing the step size h.

To work with this integral approximation in practice, we truncate the infinite sum in (2.13) to

$$S_R(t|X) \equiv \frac{h}{2\pi \mathrm{i}} \sum_{r=-R}^{R} g(t, r|X),$$

and then use extrapolation to approximate the case where $R \to \infty$. Because $S_R(t|X)$ is a nearly periodic function in R, the limit $\lim_{R\to\infty} S_R(t|X)$ can be efficiently approximated using Euler summation:

$$f(t|X) \approx E(R, M, t|X) \equiv \sum_{m=0}^{M} 2^{-M} \binom{M}{m} S_{R+m}(t|X), \tag{2.14}$$

for some $M, R \in \mathbb{N}$. Abate and Whitt (1992) find that for most probability densities the error introduced by approximating the limit $R \to \infty$ by an Euler summation is well estimated by $E(R, M+1, t|X) - E(R, M, t|X)$. In our case, this estimated error quickly tends to zero as M is increased, suggesting the approximation is accurate.

The log likelihood function of a sample of complete durations and covariates from an MHT model with parameters α can be computed by combining the individual approximate probabilities from (2.14) into the sum of their logarithms,

$$\ell_N(\alpha) = \sum_{i=1}^{N} \ln f(T_i|X_i) \approx \sum_{i=1}^{N} \ln(E(R, M, T_i|X_i)) \tag{2.15}$$

It is straightforward to extend this approach to independently censored data. The computation of the log likelihood contribution of a censored observation requires the computation of the survival function $\overline{F}(\cdot|X)$ at the censoring time and the corresponding covariate value. This survival function can be approximated along the lines above, using that the Laplace transform of $\overline{F}(\cdot|X)$ can be explicitly expressed in terms of the known transform $\mathcal{L}_T(\cdot|X)$ of $f(\cdot|X)$. In particular, using integration by parts, it is easy to show that

$$\frac{1 - \mathcal{L}\left[\Lambda(s)\phi(X)\right]}{s} = \int_0^\infty \exp(-st)\overline{F}(t|X)dt = \frac{1 - \mathcal{L}_T\left(s|X\right)}{s}.$$

With (2.4), this allows us to express a known function of the model's parameters as the Laplace transform of the survival function $\overline{F}(\cdot|X)$, analogously to the expression for the density in (2.10). This transformation can be numerically inverted to compute the survival function, and the likelihood contribution of each censored observation, using the strategy developed for the density. One minor difference is that the Laplace transform of the survival function may have a singularity at 0 if the durations do not have a (finite) mean; then, it is necessary to set $\gamma > 0$.

We approximate the score and Hessian of the log likelihood with the analytical first and second derivatives of the approximate log likelihood function. These exist and are well behaved because our approximation of the log likelihood function in (2.15) is smooth in the parameters.

The implementation of this method for computing the likelihood and its derivatives requires that we set the parameters that control the approximation in (2.14): γ, h, R, and M. Rogers (2000) provides guidance. We find that his suggestions for γ and h, $\gamma = 11/t$ and $h = 1/t$, for duration t, yield good numerical performance in our case. We will adopt these as our default settings, together with $R = 6$ and $M = 15$, which Rogers claims provide a good accuracy to speed trade-off. As discussed below, additional accuracy can be obtained when needed by setting M higher.

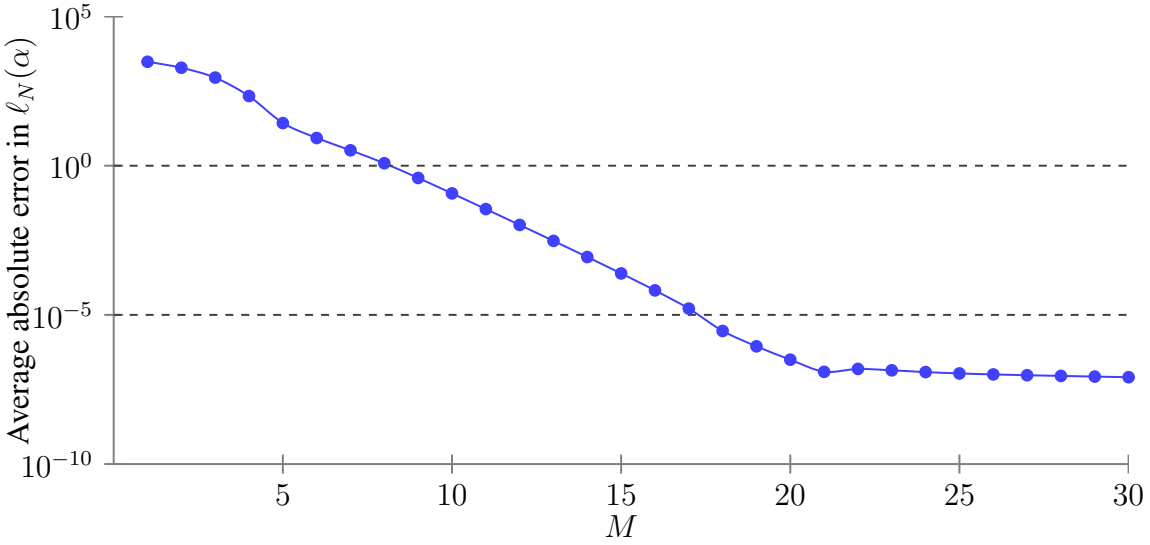
2.3.5 Numerical Experiments

We have investigated the accuracy of the proposed likelihood approximation by conducting a range of numerical experiments. We discuss the results of two of these experiments here. Both experiments use the default settings for the parameters that control the approximation, unless explicitly stated otherwise.

The first experiment compares direct computations of the log likelihood function of the mixed inverse Gaussian model using the explicit expression for the density in (2.6) to its numerical approximations as we vary M. The log likelihood is calculated on the data set that we use in Section 2.5. This ensures that this experiment provides both a real life test case and a check on the results we present in that section. The data contain 566 complete strike durations. Because the approximation errors are close to unbiased, the error in the log likelihood scales with the root of the sample size.

Figure 2.1 plots the average of the absolute approximation error of the log likelihood, for different values of M, over a large set of model parameters randomly generated at the scale of their maximum likelihood estimates. We find that this average absolute error decreases exponentially with M; this result is robust across the various parameter values over which the plotted results are averaged. Consistently with Rogers (2000), we see that $M = 15$ already provides a decent approximation for most practical purposes. However,

Figure 2.1: Approximation Error of the Log Likelihood for Various M

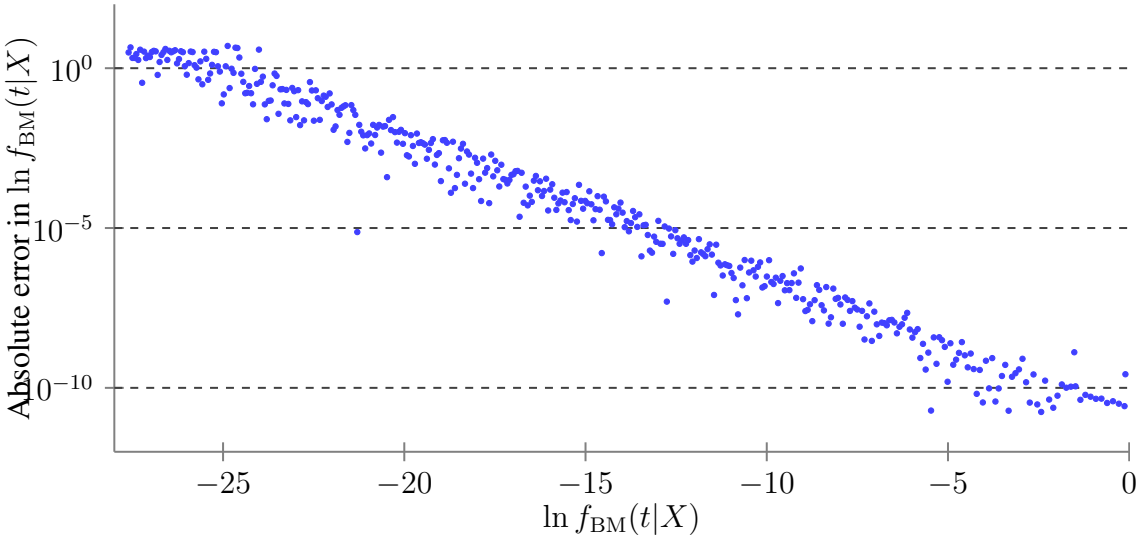


Note: This figure is based on the log likelihood $\ell_N(\alpha)$ of an MHT model with a Brownian motion latent process and discrete unobserved heterogeneity with three support points for Kennan's (1985) complete strike duration data. It plots the average absolute difference between $\ell_N(\alpha)$ and its numerical approximation over 100 randomly drawn parameter values α, for a range of values of M. The errors are plotted on a logarithmic scale. The parameters are generated using our method of setting starting values for maximum likelihood estimation. This method sets the drift and variance parameters equal to their maximum likelihood estimates for a simple inverse Gaussian model with $\phi(X)V = 1$, which are known in closed form. Starting values for the support points v_l of the heterogeneity distribution are generated by exponentiating draws from a standard normal distribution. This ensures that the v_l vary in level, but are all approximately of the right scale. All three support points v_l receive probability mass $1/3$. The parameter β multiplying the covariates is set to zero. For the current experiment, we found that setting the parameters to their final maximum likelihood estimates instead produced almost identical results.

because the time required for the calculations grows only linearly in M, an extra thousandfold increase in precision can be obtained at a very low computational cost by setting $M = 20$ instead. Once $M > 20$, other factors, such as rounding errors, become important, and the approximation error levels off. We also find that, with $M = 20$, increasing R or decreasing the step size h adds very little to the precision of the inversion. The numerical approximation of the log likelihood takes about 15–20 times as long to calculate as the analytical expression. However, in absolute terms this is still very manageable: A maximization of the log likelihood function can be performed in under a minute on a regular computer for most model specifications.

The second experiment takes a closer look at the numerical approximation of the density f_{BM} of a basic inverse Gaussian model with parameters such that $\mu = \sigma^2 =$

Figure 2.2: Approximation Error of the Log Inverse Gaussian Density Function



Note: This figure plots the absolute difference between the log inverse Gaussian density $\ln f_{\mathrm{BM}}(t|X)$ with parameters $\mu = \sigma^2 = \phi(X)V = 1$ and its numerical approximation, on a logarithmic scale, against $\ln f_{\mathrm{BM}}(t|X)$, for a range of times t.

$\phi(X)V = 1$. We only present results for $M = 25$, but found very similar results for any $M \geq 15$. For the purpose of maximum likelihood estimation, we care most about the errors in the approximation of the *log* density, $\ln f_{\mathrm{BM}}$. Figure 2.2 plots the absolute error of this approximation against the log density itself, on a logarithmic scale. The (log-)linear relation displayed by the graph implies that the absolute error in the approximation of $\ln f_{\mathrm{BM}}(t|X)$ roughly equals $10^{-11}/f_{\mathrm{BM}}(t|X)$. Consequently, the approximation error is generally small, but the approximation breaks down when the density gets very small (say, $f_{\mathrm{BM}}(t|X) < 10^{-10}$, or $\ln f_{\mathrm{BM}}(t|X) < -23$). When estimating the model with maximum likelihood, we can easily avoid this by setting reasonable starting values for the parameters. This ensures that the approximation is sufficiently precise for numerically robust maximum likelihood estimation.

2.4 Maximum Likelihood Estimation

This chapter is accompanied with MATLAB code that implements a maximum likelihood estimator based on the previous section's approximate likelihood. We maximize this likelihood by means of a quasi-Newton algorithm with BFGS updates for the Hessian (see Nocedal and Wright, 2006). We use the analytical derivatives of the approximate

likelihood to ensure quick and stable maximization, and to construct asymptotic standard errors.

We have implemented a range of computationally feasible, flexible parameterizations of the model. This section's remainder discusses these parameterizations.

2.4.1 Latent Process

We consider two parameterizations of the latent process $\{Y\}$. Both include a Brownian motion component with $\sigma > 0$.

The main specification specifies that $\{Y\}$ is a convolution of a nondegenerate Brownian motion with drift and a compound Poisson process with a finitely discrete shock distribution. Because $\int_{(-1,0)} x\Upsilon(dx) < \infty$ in this case, the Lévy-Khintchine formula (2.3) now offers the simplest way to parameterize ψ:

$$\psi(s) = \mu s + \frac{\sigma^2}{2}s^2 + \sum_{q=1}^{Q} \lambda_q \left(\mathrm{e}^{s\nu_q} - 1\right),$$

where μ and $\sigma^2 \geq 0$ are the Brownian drift and variance per time unit, and λ_q is the Poisson rate at which shocks of size $\nu_q < 0$ arrive; $q = 1, \ldots, Q$. Equivalently, in this specification, shocks arrive at a rate $\lambda \equiv \sum_{q=1}^{Q} \lambda_q$ and are drawn independently from a distribution with Q points of support $(\nu_1, \ldots, \nu_Q)$ with probabilities $(\lambda_1/\lambda, \ldots, \lambda_Q/\lambda)$.

An alternative is to specify $\{Y\}$ as a convolution of a nondegenerate Brownian motion with drift and a compound Poisson process with a gamma shock distribution. In this specification, shocks arrive at a Poisson rate λ, with their absolute sizes distributed according to a two-parameter gamma distribution $\Gamma_{\nu,\rho}$, with corresponding density

$$\frac{\nu^\rho}{\Gamma(\rho)}\, x^{\rho-1} \exp(-\nu x); \qquad \nu, \rho > 0;$$

and Laplace transform

$$\mathcal{L}_{\Gamma_{\nu,\rho}}(s) = \frac{1}{(s/\nu + 1)^\rho}. \tag{2.16}$$

We can again use (2.3), which now gives

$$\psi(s) = \mu s + \frac{\sigma^2}{2} s^2 + \lambda \left\{ \frac{1}{(s/\nu + 1)^\rho} - 1 \right\}.$$

2.4.2 Effect of the Observed Covariates

The threshold is naturally specified to be loglinear in the covariates:

$$\phi(X) = \exp(\delta + X\beta).$$

We assume that the $N \times (K + 1)$ matrix with sampled observations of $(1\ X')$ in each row has full column rank.

2.4.3 Unobserved Heterogeneity

Finally, our procedure for computing the likelihood only depends on the unobserved heterogeneity distribution G through its Laplace transform $\mathcal{L}$. Therefore, any distribution with nonnegative support that admits an explicit expression for its Laplace transform is a convenient candidate for G. We consider two such specifications.

The main specification is Section 2.3's finite discrete distribution. The corresponding Laplace transform is

$$\mathcal{L}(s) = \sum_{l=1}^{L} \pi_l \exp(-s v_l).$$

A simple and low-dimensional alternative is to specify a gamma distribution $\Gamma_{\omega,\tau}$ for G. Analogously to (2.16), this gives

$$\mathcal{L}(s) = \frac{1}{(s/\omega + 1)^\tau}.$$

2.4.4 Scale Normalizations

Recall from Section 2.3.1 that we need to normalize the scales of two out of ψ, ϕ, and $\mathcal{L}$. The MATLAB code currently normalizes the covariate effects $\phi(X)$ by setting $\delta = 0$, and ψ by setting $\mu = 1$. Note that this implicitly assumes that $\mu > 0$. It would be straightforward to adapt the code to allow more generally for $|\mu| = 1$.

One of these normalizations can be replaced by a normalization on $\mathcal{L}$. A discrete unobserved heterogeneity distribution can, for example, be normalized by requiring $v_1 = 1$ and $\pi_1 > 0$. A gamma distribution can be normalized by setting its scale parameter $\omega = 1$.

2.5 Strike Durations

The mere existence of nontrivial delays in labor agreements has puzzled economists; duration patterns in their resolution have been studied to learn more about underlying bargaining games and information structures.

Lancaster (1972) analyzes strike durations using a Gaussian MHT model with regressors, but without unobserved heterogeneity. He interprets the gap between the Brownian motion and the threshold as the level of disagreement, and concludes that this model fits his data for the United Kingdom well. Others have used proportional hazards models to study strike durations. Kennan (1985), in particular, shows that the US strike duration hazard is U-shaped and takes this as evidence against Lancaster's (homogeneous) MHT model. He notes that this aspect of the data can be interpreted in terms of heterogeneity in the conflicts underlying the strikes, but does not subsequently pursue this in his empirical analysis.

Here, we will investigate whether Kennan's strike data can be matched well by a more general MHT model that explicitly takes into account unobserved heterogeneity in strikes. Such a model comes with Lancaster's attractive interpretation in terms of a level of disagreement that may both vary over time and may initially be heterogeneous between strikes. We will explicitly discuss our estimation results in terms of this interpretation, with an implicit understanding that it is our modest objective to illustrate our methods and the descriptive and potential structural appeal of the MHT model, without providing a fully structural analysis of strike durations.

Kennan's (1985)'s data cover all contract strikes in US manufacturing in the period 1968–1976 that involved at least a thousand workers, and that were classified to be primarily about "general wage changes". They include the durations in days of 566 strikes and, for each strike, a measure of the state of the business cycle in the month it started:

The residuals of a regression of log industrial production in US manufacturing on linear and quadratic trend terms and seasonal dummies. We obtained the data in a fixed format text file `strkdur.asc` from Cameron and Trivedi's (2005) web page. We divided all strike durations by seven, so that they are measured in weeks.

Table 2.1 reports maximum likelihood estimates for a range of Section 2.4's flexible parameterizations. All reported estimates are computed using Section 2.3.4's numerical methods, with $M = 25$. To further check these methods and their MATLAB implementation, we have also computed the same estimates for lower values of $M \geq 15$ (not reported), and estimates for the first five specifications using the explicit expressions for the log likelihood that are available in these cases (not reported). These results are virtually identical to those reported in Table 2.1.

In all cases, we specify $\phi(X) = \exp(X\beta)$, with X the scalar business cycle indicator. Columns I–V presents estimates of models with Brownian motion latent processes and discrete unobserved heterogeneity. Throughout, the drift is normalized to 1 per week ($\mu = 1$), so that $\mathbb{E}[T|X, V] = -\mathcal{L}_T'(0 + |X, V) = \exp(X\beta)V$. By its construction as a regression residual, X varies around zero and is close to zero on average in the sample. Consequently, V can be interpreted as the unobserved initial level of disagreement, measured as the mean number of strike weeks it commands.

The log likelihood substantially improves when adding a second, third and fourth support point to the distribution of V, between Columns I and IV, but a fifth support point (Column V) hardly changes the fit and the other parameters' estimates. The estimates indicate that there is both substantial heterogeneity in the strikes' initial levels of disagreement and uncertainty in their evolution over time. The numbers in Column IV imply that there are four unobserved types of labor conflict, on average commanding respectively 1.10, 3.21, 7.17, and 18.56 strike weeks. Each type's level of disagreement evolves with a standard deviation per week just above the unit drift towards agreement.

It is instructive to note that the variance of the latent process drops substantially, from close to 20 to just over 1, when more heterogeneity is added between Columns I and IV. Clearly, Column I's specification falsely attributes heterogeneity in the strikes' initial levels of disagreement to uncertainty in their evolution over time.

The estimates of the coefficient β reflect the effect of the business cycle on strike

Table 2.1: Maximum Likelihood Estimates for Kennan's (1985) Strike Duration Data

	I	II	III	IV	V	VI	VII
μ	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
σ^2	19.6592 (3.1752)	6.2185 (0.8702)	2.0675 (0.4433)	1.2272 (0.2423)	1.1966 (0.2224)	0.5423 (0.2808)	5.1469 (0.9768)
λ						0.0186 (0.0183)	
ν						−5.1321 (2.3211)	
β	−0.9306 (0.6010)	−1.7722 (0.6855)	−1.0846 (0.6572)	−0.8669 (0.6514)	−0.8623 (0.6338)	−0.5788 (0.6148)	−2.1198 (0.7881)
ω							0.4446 (0.0730)
τ							2.7911 (0.4373)
v_1	6.2603 (0.4688)	2.5431 (0.1993)	1.5369 (0.1508)	1.1045 (0.1213)	1.0312 (0.1644)	0.7546 (0.1602)	
v_2		8.7509 (0.5194)	5.8883 (0.3999)	3.2094 (0.4531)	1.7564 (1.0282)	2.0832 (0.5127)	
v_3			18.1612 (1.0108)	7.1654 (0.5598)	3.5180 (0.7618)	4.1380 (0.8364)	
v_4				18.5572 (0.7028)	7.3032 (0.6467)	7.4121 (0.5533)	
v_5					18.5749 (0.6945)	17.0035 (1.1016)	
π_1	1 (0)	0.3991 (0.0439)	0.3534 (0.0335)	0.2519 (0.0380)	0.1986 (0.1160)	0.1978 (0.0398)	
π_2		0.6009 (0.0439)	0.4923 (0.0347)	0.2826 (0.0507)	0.0981 (0.1300)	0.2009 (0.0688)	
π_3			0.1543 (0.0231)	0.3146 (0.0541)	0.2561 (0.0825)	0.2230 (0.0617)	
π_4				0.1508 (0.0191)	0.2969 (0.0646)	0.2379 (0.0609)	
π_5					0.1503 (0.0190)	0.1403 (0.0200)	
ℓ_N	−1658.9	−1588.7	−1583.0	−1576.3	−1576.1	−1575.4	−1594.2

Note: The drift is normalized to 1 per week. All specifications include a single covariate, Kennan's (1985) deseasonalized and detrended log industrial production. Asymptotic standard errors are in parentheses.

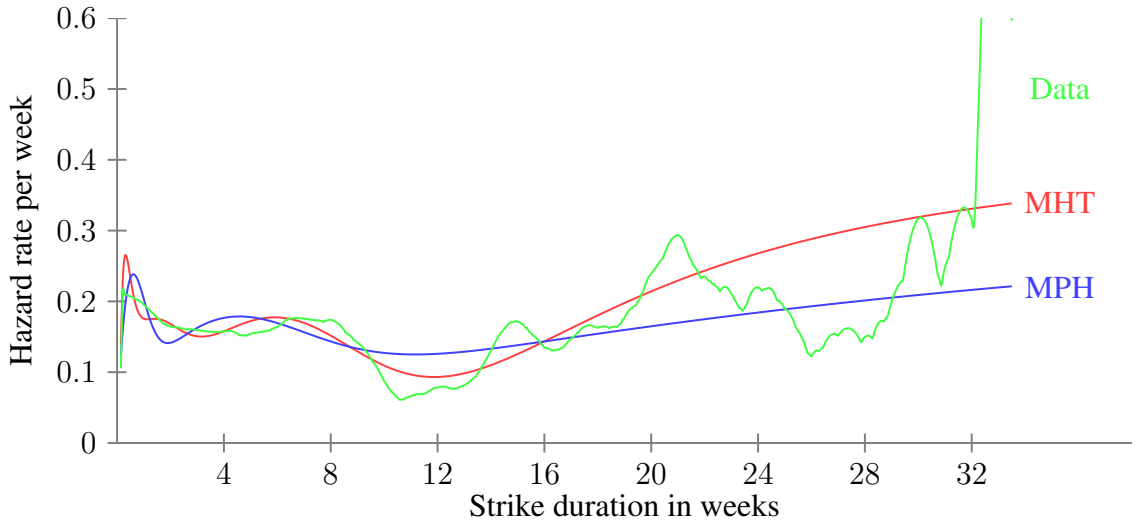
durations. In line with Kennan's (1985) results, strikes that begin in months with low production last longer. In the MHT model, this is captured by a countercyclical threshold: In times with low production, in expectation, conflicts command more strike days. One interpretation is that strike days are less costly in times with low production. The precision of the estimates of β is low. This is consistent with Kennan's results. He obtains more precise results with a binary cyclical indicator constructed from the indicator used here. For simplicity, we do not follow this lead here.

Column VI reports an estimate of a specification that includes discrete shocks of size ν at Poisson times. The estimates point to an infrequent shock that sets back just over five weeks of drift towards agreement. The shock only somewhat improves the likelihood; a specification without shock, such as those in Columns IV and V, seems to be sufficient.

A very similar result is found with a gamma shock at a Poisson time (not reported). With this specification, virtually the same estimate of the arrival rate of the shocks is obtained. Moreover, the estimated gamma shock distribution is close to degenerate at Column VI's estimate of the shock size (ν). Specifically, the estimates of the shape (ρ) and scale (ν) parameters of the gamma distribution are both very large, and their ratio equals Column VI's estimated shock size. As expected, the same log likelihood is found.

Finally, Column VII reports estimates of a specification with gamma heterogeneity. This specification is clearly inferior to that with any amount of discrete heterogeneity.

Figure 2.3: Aggregate Strike End Hazard Rates



Note: This graph plots the empirical strike end hazard rate (Data), computed with Epanechnikov kernel smoothing from Kennan's (1985) data, and the corresponding hazards implied by estimated MHT and MPH models. For the MHT model, the ML estimates in Table 2.1 for a specification with a latent Brownian motion and a discrete unobserved heterogeneity distribution with four support points are used. For the MPH model, we use ML estimates of a model with the same discrete heterogeneity distribution and a Weibull baseline. Estimated hazard rates of the unconditional distribution of T are plotted, based on the estimated distributions of $T|X$ implied by the models and the empirical distribution of the covariate X.

Figure 2.3 plots the aggregate hazard implied by the MHT model's estimates in Column IV of Table 2.1. It also plots the hazard implied by estimates a MPH hazard model with a Weibull baseline and a discrete heterogeneity distribution with four support points. Note that this MPH specification has exactly the same number of parameters as Column IV's MHT specification.[1] In both cases, we computed the distribution of $T|X$ implied by these estimates, integrated over the empirical distribution of X, and computed and plotted the hazard rate of the resulting distribution. Figure 2.3 also plots the empirical hazard rate, computed by kernel smoothing the raw data.

The MHT model fits the empirical hazard well. The MPH model's fit seems to be slightly worse. This is confirmed by the MPH model's log likelihood, which, at -1583.4, is more than seven points lower. Because the Weibull baseline is monotonic, the Weibull MPH model can only fit the nonmonotonic strike hazard by compensating an increasing baseline hazard with negative duration dependence due to unobserved heterogeneity. Of

[1] However, estimates of two of the support points of the heterogeneity distribution converged to the same value.

course, usually MPH models with richer specifications of the baseline hazard are estimated and a sufficiently rich specification can fit the empirical hazard arbitrarily well.

2.6 Conclusion

The results in this chapter enable applied researchers to analyze duration data with mixed hitting-time (MHT) models using standard likelihood-based estimation and inference methods. The MATLAB code for maximum likelihood estimation that accompanies this chapter can directly be applied to either complete or independently right-censored duration data, and is easy to adapt to more general censoring schemes. Alternatively, the procedures for likelihood computation provided with this code can be used to implement other likelihood-based methods. For example, they can be combined with data augmentation and Markov chain Monte Carlo methods to implement a Bayesian estimator that can flexibly deal with unobserved heterogeneity.

Two types of empirical application of the MHT framework can be distinguished. First, it can be used as a descriptive framework, much like Cox's (1972) proportional hazards model and Lancaster's (1979) mixed proportional hazards model. Section 2.5's analysis of Kennan's (1985) strike data shows that estimates of the MHT model have descriptive appeal, with natural interpretations that nicely complement those that could be obtained from a proportional hazards analysis. Indeed, in statistics, there is increasing interest in the descriptive analysis of duration data with first hitting time models (Singpurwalla, 1995; Yashin and Manton, 1997; Aalen and Gjessing, 2001; Lee and Whitmore, 2006).

Second, it can be applied to the structural empirical analysis of heterogeneous agents' optimal stopping decisions. Abbring (2012) presents a range of examples, based on the type of optimal stopping models that are reviewed and analyzed in Dixit and Pindyck (1994); Stokey (2009); Kyprianou (2006); Boyarchenko and Levendorskiĭ (2007). These include McDonald and Siegel's (1986) model for the optimal timing of an irreversible investment; a model of unemployment durations based on Dixit's (1989) model of entry and exit, complemented with heterogeneity in transition costs; and a model of job separations with heterogeneous search. The identification results in Abbring (2012, 2010) show that data on durations and covariates are informative on the economic primitives

of such models. The methods developed in this chapter can be applied to measure those primitives.

Chapter 3

Approximate expectation-maximization for large non-Gaussian state space models

Joined work with Dennis Fok

3.1 Introduction

The state-space model has proven to be an extremely flexible model to capture dynamic patterns in many different situations. In the typical application of the state-space model, the dependent variable and all state variables are assumed to have a conditional Gaussian distribution. Computationally efficient implementations of the Kalman filter and smoother make classical and Bayesian inference in such models relatively straightforward, see e.g. Durbin and Koopman (2001). Even if the state space is reasonably large and/or the model is multivariate, computationally efficient inference is still feasible. Things change however if the distributions are not Gaussian. In this chapter we consider the case where the dependent variable does not have a conditional Gaussian distribution. For example, the dependent variable may be a binary or a count variable. When the Gaussian distribution no longer holds, standard algorithms break down. Alternative approaches to inference based on Monte Carlo methods are commonly applied in this situation, but in very high dimensions these fail due to the curse of dimensionality as is discussed in Section 3.4.2.

We propose an algorithm to perform maximum likelihood estimation of such multivariate non-Gaussian dynamic models and we explicitly consider the case where the state space is so large, say in the other of tens of thousands, that sampling-based algorithms are no longer feasible.

Our approach is based on the Expectation Maximization [EM] algorithm, where we treat all states as latent variables. The expectation step of the EM algorithm requires calculating the expectation of the complete data log likelihood over these latent variables, conditional on the data. In the case of non-Gaussian dependent variables, this step is not straightforward. We propose to perform the expectation step using the Expectation Propagation [EP] approach introduced by Minka (2001a). In general, EP yields a good approximation of the necessary conditional expectation. If the dependent variable has a Gaussian conditional distribution, EP is exact. Overall, our approach can be seen as a generalization of EM. In this chapter, we argue and show that the combination of EM and EP performs very well, even if the dependent variable has a non-Gaussian distribution.

The remainder of this chapter is structured as follows. In Section 3.2 we formalize the setup of the model. In Section 3.3 we present our approximate Maximum Likelihood algorithm based on a combination of Expectation Maximization and Expectation Propagation [EM-EP]. We discuss alternative estimation methods for non-Gaussian dynamic models in Section 3.4. Here we explain the main advantage of our method: it aims to directly approximate the maximizer of the log-likelihood, while existing approaches approximate the log-likelihood itself. Our method thereby avoids many of the complications that arise in approximating the log-likelihood of a truly high dimensional dynamic model. The quality of our approximate Maximum Likelihood estimator is demonstrated using simulations in Section 3.5. We empirically illustrate the performance and added value of our technique by applying the EM-EP algorithm to two practical cases. In Section 3.6, we consider a model for time series of newspaper sales over a large number of individual outlets. Here, conditional on state variables, the sales figures are assumed to follow a censored Poisson distribution. The censoring is caused by the possibility that the newspaper is sold out. In Section 3.7 we present a dynamic model for forecasting the outcomes of individual chess matches. In this case the dependent variable is trinomial (white wins, black wins, or draw). This dependent variable is explained by latent dynamic player-specific

skill variables. We conclude the chapter in Section 3.8 with a summary and some further discussion.

3.2 Non-Gaussian dynamic models

In this section we describe our model specification in detail, where we present the setup of the model in a very general form. We consider a time series of a J-dimensional multivariate dependent variable, y_t, where $t = 1, \ldots, T$. We denote the elements of the vector y_t by $y_{j,t}$, $j = 1, \ldots, J$. The distribution function of $y_{j,t}$ is parametrized by the random variable $\mu_{j,t}$, and denoted as $p(y_{j,t}|\mu_{j,t})$, where $p()$ denotes a general (conditional) distribution function. Conditional on the $\mu_{j,t}$ parameters, all observations are independent. For the estimation approach we present below, we only need to assume that $p(y_{j,t}|\mu_{j,t})$ can easily be evaluated and that all parameters are contained in $\mu_{j,t}$. Note that for most cases the dimension of $\mu_{j,t}$ will be small, in practice it often has dimension one or two. For example, if $y_{j,t}$ is a count variable, $p(y_{j,t}|\mu_{j,t})$ could be the Negative Binomial distribution. In this case $\mu_{j,t}$ would contain (transformations of) the two parameters of this distribution. We treat $\mu_{j,t}$ as latent state variables. All these latent variables are collected in the vector $\mu_t = (\mu'_{1,t}, \ldots, \mu'_{J,t})'$. Next, there may be other state variables that are not directly related to the distribution of $y_{j,t}$, we denote the vector of these state variables by β_t. These state variables can be used to model a particular dependence structure among the $\mu_{j,t}$, and therefore $y_{j,t}$, variables. The development of the complete state vector $s_t = (\mu'_t, \beta'_t)'$ is specified as

$$\begin{aligned} s_t &= a_t + B_t s_{t-1} + C_t e_t, \\ e_t &\sim N(0, \Sigma_t), \end{aligned} \tag{3.1}$$

where a_t, B_t, C_t, and Σ_t are deterministic, possibly time varying vectors or matrices. Conditional on these vectors and matrices, the state follows a linear evolution over time with multivariate Gaussian noise. The model is completed with an assumed Gaussian distribution for the initial state

$$s_1 \sim N(m, D), \tag{3.2}$$

where m is often set to zero and D to $\kappa\mathbf{I}$ with κ a large number. Note that the dependence among the elements of y_t and between y_t and its past is completely captured through the unobserved dynamic state variables denoted by the vector s_t, that is,

$$p(y_{j,t}|s_1,\ldots,s_T,\{y_{s,l}\}_{(s,l)\neq(j,t)}) = p(y_{j,t}|s_1,\ldots,s_J) = p(y_{j,t}|\mu_{j,t}). \tag{3.3}$$

In our first empirical application on newspaper sales, we will use a censored Poisson distribution for $p(y_{j,t}|\mu_{j,t})$, where $\mu_{j,t}$ is one dimensional and $\exp(\mu_{j,t})$ equals the Poisson rate. For the chess match application, $p(y_{j,t}|\mu_{j,t})$ is a multinomial distribution, where the probabilities depend on the one dimensional state variable $\mu_{j,t}$. In both cases the dimension of s_t is very large, that is, in the order of tens of thousands or more.

As is well known in the state space literature, (3.1) allows for all kinds of dynamic patterns. Many of the parameters are usually pre-specified to obtain a particular structure. It is for example possible to obtain deterministic state variables, a dynamic factor specification, an ARMA specification for $\mu_{j,t}$, or a specification with exogenous regressors (see e.g. Durbin and Koopman (2001)). To show how exogenous variables may be included, suppose that $J = 1$ and that the density of $y_{1,t}$ is parametrized by only one parameter. For the case of K exogenous variables, the state vector would become

$$s_t = \begin{pmatrix} \mu_{1,t} \\ \beta_t \end{pmatrix}, \tag{3.4}$$

with β_t a $K \times 1$ dimensional vector. A setting with time varying effects of exogenous variables could be obtained by specifying

$$\begin{pmatrix} \mu_{1,t} \\ \beta_t \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix} + \begin{pmatrix} 0' & x_t' \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mu_{1,t-1} \\ \beta_{t-1} \end{pmatrix} + \begin{pmatrix} 0' \\ \mathbf{I} \end{pmatrix} e_t. \tag{3.5}$$

If we set the variance of e_t to zero we obtain a specification with constant coefficients.

Inference in the model given by (3.1), (3.2), and (3.3) usually consists of three parts. First of all the interest is in obtaining estimates of the parameters contained in a_t, B_t, C_t, and Σ_t given observed data $y = (y_1,\ldots,y_T)$. Next the interest is in inference on the underlying state variables. In Gaussian dynamic models the Kalman filter and smoother

can be used to calculate filtered estimates, that is, $\mathbb{E}[s_{j,t}|y_1, \ldots, y_{t-1}]$ or smoothed estimates, that is, $\mathbb{E}[s_{j,t}|y]$. In fact in Gaussian models the filtered and smoothed *densities* are obtained. The EM-EP algorithm that we present in the next section yields (approximate) Maximum Likelihood estimates of the parameters and as a natural by-product the approximate smoothed estimates for the state variables. Filtered estimates can also be obtained.

3.3 Approximate Expectation Maximization

Let θ denote the vector of all unknowns in (a_t, B_t, C_t, Σ_t) defined above. A popular way of obtaining a maximum likelihood estimate of θ in models with latent variables is by using the *Expectation Maximization* [EM] algorithm. Intuitively, the standard EM algorithm is based on the so-called complete data likelihood, $p(y, s; \theta)$, where $s = (s_1, \ldots, s_T)$. Next, given some initial value for θ, in the expectation step of the algorithm $\mathbb{E}[\log p(y, s; \theta)|y]$ is calculated, where the expectation is taken over s. The maximization step yields an updated estimate of θ by maximizing the expected log complete data likelihood over θ. Alternating both steps yields the ML estimator of θ.

For our purpose it is important to discuss the formal background of the EM algorithm. In general, the expectation step can be seen as constructing a lower bound to the marginal likelihood $p(y; \theta)$, see Minka (1998) for this perspective. Next, this lower bound is maximized in the maximization step. At convergence, the bound is tight and the EM algorithm therefore yields a (local) maximum of the marginal likelihood function. In the expectation step, this lower bound on the marginal likelihood is constructed by making use of Jensen's inequality, that is,

$$\begin{aligned} \ell(\theta) = \log p(y;\theta) &= \log \int p(y, s; \theta) ds \\ &= \log \int Q(s) \frac{p(y, s; \theta)}{Q(s)} ds \\ &\geq \int Q(s) \log \frac{p(y, s; \theta)}{Q(s)} ds \\ &= \int Q(s) \log[p(y, s; \theta)] ds - \int Q(s) \log[Q(s)] ds, \end{aligned} \tag{3.6}$$

where $Q(s)$ can be any probability distribution on the state variables s, as long as it is positive everywhere on the support of $p(y, s; \theta)$. This shows that the log likelihood equals the conditional expectation of the log complete data likelihood under the distribution $Q(s)$ plus the *entropy* of this distribution. The expectation step of the EM algorithm can be seen as choosing a particular $Q(s)$

The maximization step of the EM algorithm then proceeds by maximizing this lower bound over θ, while keeping the $Q(s)$ distribution fixed. If we keep $Q(s)$ fixed, the entropy term of the bound does not depend on θ. The maximization step therefore amounts to maximizing the expectation of the log complete data likelihood under the distribution $Q(s)$.

Finally, we need to choose how to set $Q(s)$ in the expectation step. The optimal choice for $Q(s)$ is the distribution that would make the bound tight. It is straightforward to show that this happens if, for a given θ, we set $Q(s)$ equal to the conditional distribution $p(s|y; \theta)$, that is,

$$\begin{aligned} \int Q(s) \log \frac{p(y, s; \theta)}{Q(s)} ds &= \int p(s|y; \theta) \log \frac{p(y, s; \theta)}{p(s|y; \theta)} ds \\ &= \int p(s|y; \theta) \log p(y; \theta) ds = \log p(y; \theta). \end{aligned} \tag{3.7}$$

However, this choice makes $Q(s)$ depend on θ and the entropy part could not be ignored in the maximization step. To approximate this situation as closely as possible, in the i-th iteration of the EM algorithm we set $Q(s)$ equal to $p(s|y; \hat{\theta}_i)$ where $\hat{\theta}_i$ is the (current) best estimate of θ. In this case the bound is only tight if the log likelihood is evaluated at $\hat{\theta}_i$. Note that as the EM algorithm proceeds, $\hat{\theta}_i$ converges to the ML estimate and the bound becomes tight when the evaluation is done at the ML estimate.

For the model we consider, the complete data log-likelihood is given by

$$\begin{aligned}\log[p(y,s;\theta)] &= \log[p(s_1;\theta)] + \sum_{t=2}^{T} \log[p(s_t|s_{t-1};\theta)] + \sum_{t=1}^{T} \log[p(y_t|s_t)] \\ &= \text{constant} - \frac{1}{2}\log(|D|) - \frac{1}{2}(s_0-m)'D^{-1}(s_0-m) \\ &+ \sum_{t=2}^{T} -\frac{1}{2}\log(|C_t\Sigma_t C_t'|) - \frac{1}{2}(s_t-a_t-B_t s_{t-1})'(C_t\Sigma_t C_t')^{-1}(s_t-a_t-B_t s_{t-1}),\end{aligned} \tag{3.8}$$

where we make use of the fact that all terms that are independent of θ in this function can be ignored for the purpose of maximum likelihood estimation. Note that this includes $p(y_t|s_t)$. The i-th iteration of the EM algorithm takes the expectation of (3.8) with respect to the distribution $Q_i(s)$ and maximizes this expectation over θ. In the i-th iteration we obtain the updated estimate of θ ($\hat{\theta}_i$) from

$$\hat{\theta}_i = \arg\max_{\theta} \mathbb{E}_{Q_i(s)}[\log p(y,s;\theta)]. \tag{3.9}$$

Next the distribution $Q_i(s)$ is updated to $Q_{i+1}(s)$ based on the value of $\hat{\theta}_i$, which is then used to find a new estimate $\hat{\theta}_{i+1}$. If $Q_{i+1}(s)$ is updated to $p(s|y;\hat{\theta}_i)$ this process is guaranteed to converge to a local maximum of the likelihood function $\mathcal{L}(\theta)$. This result continues to hold when we replace the maximization step (3.9) by a step that increases (but does not maximize) the expected complete data log likelihood. In particular, instead of performing a joint maximization it is often more convenient to maximize over several components of θ in turn, which is called Expectation Conditional Maximization or ECM (Meng and Rubin, 1993).

The above is only feasible if $p(y_{j,t}|\mu_{j,t})$ is a Gaussian distribution function, since $p(s|y;\theta)$ is then also Gaussian. For a general distribution $p(y_{j,t}|\mu_{j,t})$, the conditional distribution $p(s|y;\theta)$ will not have a standard form, and exact EM is intractable. However, using (3.8) we see that the update equation (3.9) only depends on the first two moments of $Q(s)$, hence any distribution $Q(s)$ with its first two moments similar to those of $p(s|y;\hat{\theta}_i)$ yields a sensible update of the estimate $\hat{\theta}_{i+1}$. Based on this finding we propose to update θ using a distribution $Q_{i+1}(s)$ that approximates the first two moments of $p(s|y;\hat{\theta}_i)$ as well

as possible.

We have multiple options for approximating the first two moments of $p(s|y;\hat{\theta}_i)$. One is to draw samples from this distribution using MCMC methods and to set the moments equal to their sample realizations. This method is known as Monte Carlo Expectation Maximization (Wei and Tanner, 1990) and has the attractive feature that it can be made arbitrarily accurate by increasing the number of samples. However, we explicitly consider large scale non-Gaussian models where the dimension of s_t can be in the order of tens of thousands. Moreover, the state may have an intricate dependency structure which makes sampling from $p(s|y;\hat{\theta}_i)$ too computationally expensive. Another possibility is to directly minimize the gap between the lower bound and the likelihood in (3.6) with respect to $Q(s)$ over some class of tractable distributions. This is the approach taken by Ormerod and Wand (2011) for quasi maximum likelihood estimation in the case of generalized linear mixed models. This technique is known as Gaussian variational approximation. Although this is an elegant method that can be hundreds to thousands times faster than using MCMC methods, the approach outlined by Ormerod and Wand (2011) is not optimal for application to high-dimensional dynamic models since it does not use the Kalman filter. In addition, the Gaussian variational approximation for $Q(s)$ does not necessarily do a good job at approximating the moments of $p(s|y;\hat{\theta}_i)$. The same can be said about the methods of Lee and Nelder (2006) and Rue et al. (2009) which are based on a Laplace approximation to $p(s|y;\hat{\theta})$.

Since the moments of $Q(s)$ determine the final estimate $\hat{\theta}$, we more directly approximate the moments of $p(s|y;\hat{\theta}_i)$ using a method called *Expectation Propagation* (Minka, 2001a). Minka (2001a), Minka and Lafferty (2002) and Kuss and Rasmussen (2005), among others, find that this method often does a better job at approximating the moments of a distribution than do variational approximations or Laplace approximations. The use of Expectation Propagation in an EM algorithm is known as EM-EP and was used earlier by Minka and Lafferty (2002), Qi et al. (2004) and Kim and Ghahramani (2006) in different contexts.

3.3.1 Expectation propagation

In our EM implementation, we construct $Q_{i+1}(s)$ by approximating the conditional distribution $p(s|y;\hat{\theta}_i)$ by a multivariate Gaussian using the Expectation Propagation method of Minka (2001a). The goal is to choose the Gaussian distribution that matches the first two moments of $p(s|y;\hat{\theta}_i)$ as closely as possible. For the models we consider, the conditional distribution is given by

$$p(s|y;\hat{\theta}_i) \propto p(s;\hat{\theta}_i)p(y|s) = p(s_1;\hat{\theta}_i)\prod_{t=2}^{T} p(s_t|s_{t-1};\hat{\theta}_i)\prod_{t=1}^{T} p(y_t|s_t), \tag{3.10}$$

where

$$p(y_t|s_t) = \prod_{j=1}^{k} p(y_{j,t}|\mu_{j,t}). \tag{3.11}$$

Recall that $\mu_{j,t}$ is a particular element of the state vector s_t, see (3.3). The unconditional distribution $p(s;\hat{\theta}_i)$ in this expression is Gaussian, but the distribution of the dependent variable $p(y_{j,t}|\mu_{j,t})$ is not, which means that $p(s|y;\hat{\theta}_i)$ is non-Gaussian. Approximating the conditional distribution $p(s|y;\hat{\theta}_i)$ by a Gaussian thus comes down to replacing the likelihood terms $p(y_{j,t}|\mu_{j,t})$ by Gaussian distribution functions in $\mu_{j,t}$. Our approximation will thus be of the following form

$$Q_i(s) \propto p(s_1;\hat{\theta}_{i-1})\prod_{t=2}^{T} p(s_t|s_{t-1};\hat{\theta}_{i-1})\prod_{t=1}^{T}\prod_{j=1}^{k} \phi(\mu_{j,t};m_{j,t},v_{j,t}), \tag{3.12}$$

where the $\phi(\mu_{j,t};m_{j,t},v_{j,t})$ denotes a Gaussian probability density function in $\mu_{j,t}$ with mean vector $m_{j,t}$ and covariance matrix $v_{j,t}$.

The only remaining part of the approximation is to set the parameters $m_{j,t}$ and $v_{j,t}$ in such a way that the approximating distribution (3.12) closely matches the first two moments of the exact conditional distribution (3.10). The approach taken in Minka's (2001a) Expectation Propagation method is to consider this problem one factor at a time: Suppose that we have already approximated all but the last likelihood contribution $p(y_{k,T}|\mu_{k,T})$. Leaving in the exact likelihood contribution would then gives us the following approxi-

mating distribution

$$\tilde{Q}_i(s) \propto p(s_1; \hat{\theta}_{i-1}) \prod_{t=2}^{T} p(s_t|s_{t-1}; \hat{\theta}_{i-1}) \frac{\prod_{t=1}^{T} \prod_{j=1}^{k} \phi(\mu_{j,t}; m_{j,t}, v_{j,t})}{\phi(y_{k,T}; m_{k,T}, v_{k,T}))} p(y_{k,T}|\mu_{k,T}) \quad (3.13)$$

Since the dimension of $\mu_{k,T}$ is low, we can determine the first two moments of this distribution efficiently by using the Kalman filter and smoother combined with numerical integration over $\mu_{k,T}$. We may then approximate $p(y_{k,T}|\mu_{k,T})$ by that Gaussian distribution $\phi(\mu_{k,T}; m_{k,T}, v_{k,T})$ which produces an approximation $Q_i(s)$ with the exact same first two moments. Such a Gaussian probability density function always exists and can be uniquely determined. Moreover, since $p(y_{k,T}|\mu_{k,T})$ only directly influences the moments of $\mu_{k,T}$, the approximation of this factor will be a function of the only the low-dimensional vector $\mu_{k,T}$ even though it conserves the first two moments of the entire state s. After approximating this likelihood contribution, we can then move on to one of the other $p(y_{j,t}|\mu_{j,t})$ that we had already approximated and repeat the procedure as if this was the final factor. In this fashion, we iteratively set all the approximate likelihood terms, giving the following algorithm

Algorithm 1 Gaussian Expectation Propagation for State Space Models

Initialize all $m_{j,t}$ to zero and all $v_{j,t}$ to large values
Define $Q_i(s) \propto p(s; \hat{\theta}_{i-1}) \prod_{t=1}^{T} \prod_{j=1}^{k} \phi(\mu_{j,t}; m_{j,t}, v_{j,t})$
while not converged **do**
  for all j, t **do**
    Find the moments of $Q_i^{\setminus j,t}(s) = Q_i(s)/\phi(\mu_{j,t}; m_{j,t}, v_{j,t})$ using the Kalman smoother, denote the mean of $\mu_{j,t}$ by $m_{j,t}^{\setminus j,t}$ and its covariance by $v_{j,t}^{\setminus j,t}$
    Find the moments of $\tilde{Q}_i(s) = Q_i^{\setminus j,t}(s) p(y_{j,t}|\mu_{j,t})$ using Gauss-Hermite quadrature, denote the mean of $\mu_{j,t}$ by $\tilde{m}_{j,t}$ and its covariance by $\tilde{v}_{j,t}$
    Set $v_{j,t} = [\tilde{v}_{j,t}^{-1} - (v_{j,t}^{\setminus j,t})^{-1}]^{-1}$
    Set $m_{j,t} = v_{j,t}[\tilde{v}_{j,t}^{-1} \tilde{m}_{j,t} - (v_{j,t}^{\setminus j,t})^{-1} m_{j,t}^{\setminus j,t}]$
  end for
end while

Given the final combined approximating distribution $Q_i(s)$, the lower bound on the marginal likelihood (3.6) can then be evaluated explicitly and we can perform the maximization step of the EM algorithm. This maximization is straightforward since we are using a linear and Gaussian specification for the state variables. Given a new estimate

$\hat{\theta}_i$ we can then again refine $Q_i(s)$ using Expectation Propagation and this is iterated until convergence.

In an application of the above procedure there may be some practical issues. First, there is no theoretical guarantee that the procedure for constructing $Q_i(s)$ will always converge (see Minka, 2001a). However, in practice this rarely is a problem, and we have not encountered any divergence issues within the class of models considered here. If the likelihood is very far from Gaussian, the convergence of the approximate expectation step may be improved by using the EP algorithms by Seeger and Nickisch (2011) or Qi and Guo (2012). Also, although all individual terms $\phi(\mu_{j,t}; m_{j,t}, v_{j,t})$ are chosen to match the moment contributions of the exact likelihood terms $p(y_{j,t}|\mu_{j,t})$, this does not mean that the combined approximate distribution $Q_i(s)$ exactly matches the first two moments of $p(s|y; \hat{\theta}_i)$. For this reason, the maximization step of the EM algorithm is not guaranteed to improve the likelihood $\mathcal{L}(\theta)$ as it is with the exact implementation. In order to ensure convergence, we found that in some cases it can be necessary to replace the update (3.9) by a *damped* version

$$\hat{\theta}_i = \alpha\hat{\theta}_{i-1} + (1-\alpha)\arg\max_{\theta} \mathbb{E}_{Q_i(s)}[\log p(y, s; \theta)], \tag{3.14}$$

where $\alpha \in (0, 1)$ is a damping constant. For most problems, the EM-EP algorithm or its damped implementation is an efficient method for finding quasi-maximum likelihood estimates. However, as with the regular EM algorithm, there also exist problems for which convergence can be slow. For these problems, it may be more efficient to use gradient based methods. Note that the gradient of the approximate likelihood at $\hat{\theta}_i$ can be obtained analytically as $\mathbb{E}_{Q_i(s)}[d/d\theta \log p(y, s; \theta)]$, evaluated at $\hat{\theta}_i$.

The quality of the EP approximation will depend on the particular specification for $p(y_{j,t}|\mu_{j,t})$, with the approximation being exact if this is a Gaussian distribution. Our experience and that of others suggests that the EP approximation is very accurate for unimodal distributions (e.g. Kuss and Rasmussen, 2005), but that it can break down when $p(y_{j,t}|\mu_{j,t})$ has multiple modes (see Minka, 2001a). An empirical investigation of the approximation quality for a unimodal case is performed in Section 3.5.

3.3.2 Additional approximations

Although the methods outlined above already provide a large reduction in computation time from exact implementations, it still requires the use of the Kalman Filter to calculate the moments of $Q_i(s)$. For every time t this requires us to store and invert a covariance matrix $K_t = \mathrm{Cov}(s_t|y_1, \ldots, y_t)$ with dimensionality matching that of s_t. This may still become prohibitive if that dimension is very large and if the covariance matrix is not easily factorizable. An additional approximation can be made here by deleting some of the off-diagonal elements of K_t before performing a Kalman filtering or smoothing step. Technically, this comes down to replacing the multivariate normal distribution $p(s_t|s_{t-1}; \hat{\theta}_i)$ in (3.12) by a normal distribution with a simpler covariance structure.

In our applications in Sections 3.6 and 3.7 we use an extreme version of this, that is, we delete all off-diagonal elements of K_t in order to accommodate a state vector with a dimensionality in the tens of thousands. The loss of accuracy caused by this additional approximation will depend on the magnitude of the deleted covariance elements. In our case, these covariances are expected to be fairly small, and the loss of accuracy is acceptable.

3.3.3 Standard errors

The Expectation Propagation procedure explained in Section 3.3.1 provides us with the following approximation to the likelihood

$$p(y;\theta) \approx \text{constant} \times \int p(\mu;\theta) \prod_{t=1}^{T} \prod_{j=1}^{k} \phi(\mu_{j,t}; m_{j,t}, v_{j,t}) \mathrm{d}\mu, \tag{3.15}$$

where μ denotes those elements of the state s that directly influence the distribution of the observed y, and the $\phi(\mu_{j,t}; m_{j,t}, v_{j,t})$ are the Gaussian approximations to the true likelihood terms $p(y_{j,t}|\mu_{j,t})$. Since this likelihood approximation is Gaussian, we can perform the integration efficiently using the Kalman filter. The gradient and Hessian of the approximate log likelihood can also be obtained using standard state space methods. The negative inverse of the Hessian can then be used as an approximate estimate of the asymptotic covariance matrix of our EM-EP estimator.

Alternatively, the approximate Hessian of the log likelihood can be obtained by approximating its derivatives directly. The derivative of the true log likelihood $\log p(y;\theta)$ with respect to a parameter θ_i is given by $\mathbb{E}_{p(s|y;\theta)}[d/d\theta_i \log p(y,s;\theta)]$, where $\log p(y,s;\theta)$ is the complete data log likelihood (see McLachlan and Krishnan, 1997). In our case, this derivative can be approximated by $\mathbb{E}_{Q(s;\theta)}[d/d\theta_i \log p(y,s;\theta)]$ where $Q(s;\theta)$ denotes the approximation to $p(s|y;\theta)$ that is obtained by running Expectation Propagation to convergence. Note that this derivative is zero at the final EM-EP estimate. Since the derivative of the complete data log likelihood depends on $p(s|y;\theta)$ only through its first two moments, this approximate derivative will be a close approximation to the derivative of $\log p(y;\theta)$. By taking a derivative once more we can also get a good approximation to the Hessian of $\log p(y;\theta)$. Let $\chi(\theta) = (\mathbb{E}_{Q(s;\theta)}s', \text{vech}[\mathbb{E}_{Q(s;\theta)}ss']')'$ denote the $p \times 1$ vector of sufficient statistics of $Q(s;\theta)$, and denote the derivative of our lower bound with respect to θ_i by $g_i(\theta,\chi(\theta)) = \mathbb{E}_\chi[d/d\theta_i \log p(y,s;\theta)]$, where $\mathbb{E}_\chi$ denotes taking an expectation with respect to the multivariate normal distribution with the moments $\chi(\theta)$. The element (i,j) of the Hessian of $\log p(y;\theta)$ can then be approximated as

$$\frac{d^2}{d\theta_i d\theta_j}\log p(y;\theta) \approx \frac{d}{d\theta_j}g_i(\theta,\chi(\theta)) = \frac{\partial}{\partial\theta_j}g_i(\theta,\chi) + \sum_{l=1}^{p}\frac{\partial}{\partial\chi_l}g_i(\theta,\chi)\frac{d\chi_l}{d\theta_j} \qquad (3.16)$$

where $d\chi_l/d\theta_j$ can be calculated numerically by recalculating the EP approximation $Q(s;\theta)$ for values slightly below and above θ_j. The other quantities can be obtained analytically.

We find that the two different methods of approximating the Hessian give almost identical results in practice and that both give very close approximations to the Hessian of the exact log likelihood. In Section 3.5 we investigate the quality of our EM-EP estimator and its approximate standard errors obtained in this way. In the next section we first discuss some alternative methods.

3.4 Other likelihood approximation methods

Direct maximum likelihood estimation is not possible for the non-Gaussian dynamic models we consider (see section 3.2) as the latent effects cannot be integrated analytically from the data likelihood. The method proposed here is one way of performing this in-

tegration approximately, but other methods have been proposed in the literature before. Here we review these alternative methods, and we argue why our method compares positively with the existing approaches. In Section 3.4.1 we discuss alternative deterministic approximations to the intractable likelihood, while Section 3.4.2 reviews Monte Carlo approximations.

3.4.1 Deterministic solutions: Laplace and Variational approximations

One of the most often used deterministic approximations for dealing with intractable likelihoods is the Laplace approximation. This method replaces the exact complete data log-likelihood $\log p(y, s; \theta)$ by a second order Taylor approximation in s around its mode $s^*(\theta)$, that is, the approximated log likelihood equals

$$\log \tilde{p}(y, s; \theta) = \log p(y, s^*(\theta); \theta) + \frac{1}{2}[s - s^*(\theta)]' H(\theta)[s - s^*(\theta)], \tag{3.17}$$

$$\text{with} \qquad H(\theta) = \left. \frac{\partial^2 \log p(y, s; \theta)}{\partial s \partial s'} \right|_{s=s^*(\theta)}. \tag{3.18}$$

Since the exponent of this Taylor approximation, $\tilde{p}(y, s; \theta)$, is now a Gaussian in s, we can easily use it to obtain an approximate likelihood with s integrated out.

$$\tilde{l}(\theta) = \log \int \tilde{p}(y, s; \theta) ds = \log p(y, s^*(\theta); \theta) + \frac{n \log 2\pi}{2} - \frac{1}{2} \log \left| -H(\theta) \right|, \tag{3.19}$$

where the last term is the log determinant of minus the Hessian matrix of $\log p(y, s; \theta)$ at its mode. Many popular techniques for approximate maximum likelihood and approximate Bayesian inference (Tierney and Kadane, 1986; Breslow and Clayton, 1993; Lee and Nelder, 2006; Rue et al., 2009) all use $\tilde{l}(\theta)$, or further approximations of this expression, as their quasi log-likelihood function. More advanced likelihood approximations based on higher order Taylor expansions have been proposed for different kinds of models (e.g. linear mixed models: Raudenbush et al., 2000), but these are not applicable to the dynamic models considered here.

Maximization of (3.19) leads to a QML estimator that has a close relationship to those

obtained using approximate expectation maximization. Specifically, note that

$$\frac{d\tilde{l}(\theta, s^*(\theta))}{d\theta} = \frac{\partial\tilde{l}(\theta, s^*(\theta))}{\partial\theta} + \mathcal{O}(3), \tag{3.20}$$

since $\partial \log p(y, s^*; \theta)/\partial s^* = 0$ and the remainder only depends on s^* through the third derivative of $\log p(y, s; \theta)$ in s. If $p(y, s; \theta)$ is reasonably well approximated by a Gaussian we expect this third derivative to be small, i.e. the local curvature of $\log p(y, s; \theta)$ around its mode should be roughly constant. For this reason, the third order term in (3.20) is often ignored when fitting models using a Laplace approximation (e.g. Breslow and Clayton, 1993). If the Gaussian approximation in s is reasonable, the maximizer of (3.19) will thus satisfy the following first order conditions

$$\frac{d\tilde{l}(\theta)}{d\theta} \approx \frac{\partial\tilde{l}(\theta)}{\partial\theta} = \mathbb{E}_{Q(s)}\frac{\partial p(s; \theta)}{\partial\theta} = 0, \tag{3.21}$$

with $Q(s) = N[s^*(\theta), -H^{-1}(\theta)]$, and where we have made use of the fact that

$$\log \tilde{p}(y, s; \theta) = \log \tilde{p}(y; s) + \log p(s; \theta).$$

These first order conditions are almost the same as those for exact ML estimation (see section 3.3), but now the expectation is taken with respect to a second order Taylor approximation of $p(s|y; \theta)$, rather than the exact distribution.

In a recent paper, Ormerod and Wand (2011) suggest approximating the ML estimator by maximizing the lower bound on the likelihood (3.6) directly. To do this they use a *Gaussian variational approximation*, i.e. they specify $Q(s)$ as a general multivariate normal distribution, with free parameters μ and Σ, and then maximize the lower bound jointly with respect to θ, μ and Σ. As shown by Opper and Archambeau (2009), the Gaussian variational approximation is closely related to the Laplace approximation, but it is nonlocal in the sense that it does not only depend on the shape of $p(s|y; \theta)$ at a single point like the Laplace approximation. Although Ormerod and Wand (2011) do not take the perspective of expectation maximization, their estimator does obey the first order conditions given in (3.21), with a Gaussian variational approximation for $Q(s)$, so it too can be seen as an approximate EM procedure.

By viewing quasi ML estimation using Laplace approximations and Gaussian variational approximations in the context of expectation maximization, we can easily compare them to the method we propose here. As discussed in section 3.3, the first order conditions in (3.21) tell us that the quality of the resulting estimator depends only on how well $Q(s)$ approximates the first two moments of $p(s|y;\theta)$, with the estimator being equal to the exact ML estimator if the first two moments are matched exactly. Therefore, it seems natural to construct $Q(s)$ by trying to match these moments directly using EP as proposed here, rather than by defining $Q(s)$ through an objective function that might not capture these moments at all. Several studies (Minka, 2001a; Minka and Lafferty, 2002; Kuss and Rasmussen, 2005; Cseke and Heskes, 2011) show that in practice EP indeed does a better job of approximating the first two moments of $p(s|y;\theta)$ compared to Laplace approximations or Gaussian variational approximations. For this reason, EM-EP will also provide a closer approximation to the exact ML estimator. The computational costs of the deterministic approximations discussed here are comparable.

The difference in quality between EM-EP and other deterministic approximations becomes especially large when we make additional approximations in the expectation step, as discussed in section 3.3.2. For EM-EP, these additional approximations amount to deleting certain off-diagonal elements of the covariance matrix of $p(s|y;\theta)$ in each Kalman filtering step. As long as the solution of (3.21) does not depend on these particular elements of the covariance (as is the case in our examples), the quality of the estimator is not affected directly. A similar approximation could be made with other deterministic methods, for example by ignoring off-diagonal elements of the Hessian $H(\theta)$ in the Laplace approximation, but the result would be similar to deleting these elements in the inverse covariance matrix, or precision matrix, rather than the covariance itself. Deleting these elements in the precision matrix will affect all elements of the corresponding covariance matrix, which will then deteriorate the estimator. The analysis in Opper and Archambeau (2009) shows that a similar effect would occur if we would use a restricted (diagonal) covariance matrix Σ in the Gaussian variational approximation.

3.4.2 Monte Carlo integration

In the literature a number of Monte Carlo based methods have been proposed to estimate the parameters of (non-linear) non-Gaussian state space models. In case of a linear model with Gaussian variables, very efficient methods are available to calculate the likelihood function using the Kalman filter, see, for example, Schweppe (1965), Harvey (1989), and Durbin and Koopman (2001). Very efficient implementations of the Kalman filter are available (Koopman, 1993; Koopman et al., 1999), such that large scale models can straightforwardly be analyzed. In case a Bayesian analysis is preferred, efficient Markov Chain Monte Carlo algorithms are also available, see Carter and Kohn (1994) and Frühwirth-Schnatter (2004).

For the case where the measurement variable is non-Gaussian the standard Kalman filter is not adequate. Kitagawa (1987) proposes to use numerical integration to replace some of the steps of the Kalman filter. However, this is only feasible for very low-dimensional problems. Two types of alternative estimation methods have become popular, one based on Bayesian statistics and one based on classical statistics. Both methods rely on an approximation of the density $p(\mu|y,\theta)$. The method popular in classical statistics uses Importance Sampling (Kloek and van Dijk, 1978) to calculate the log likelihood function, which is next maximized over the parameters of the model. In importance sampling the density $p(\mu|y,\theta)$ is approximated by the (Gaussian) density $f(\mu|y,\theta)$. For the evaluation of the log likelihood the identity

$$p(y|\theta) = \int \frac{p(y|\mu,\theta)p(\mu|\theta)}{f(\mu|y,\theta)} f(\mu|y,\theta)ds, \tag{3.22}$$

is used. Next one obtains draws $\mu^{(l)}$, $l = 1,\ldots,L$ from $f(\mu|y,\theta)$ and calculates

$$\log \hat{p}(y|\theta) = \log \frac{1}{L} \sum_{l} \frac{p(y|\mu^{(l)},\theta)p(\mu^{(l)}|\theta)}{f(\mu^{(l)}|y,\theta)}. \tag{3.23}$$

The suggested methods in the literature differ in the approximation made. Jungbacker and Koopman (2007) generalize the methods proposed by Shephard and Pitt (1997) and Durbin and Koopman (1997) by obtaining the approximating density using a Laplace transformation of the smoothing density $p(\mu|y,\theta)$. Note that μ is the part of the state

vector directly influencing y. This approximation requires finding the mode of μ conditional on y. Finding this mode is complex and may require large scale numerical optimization. The dimension of this optimization problem is equal to the number of observations times the dimension of μ_t. Jungbacker and Koopman (2007) propose an algorithm to find the mode which is based on repeatedly applying the Kalman filter to a particular linear model. This linear model is constructed using the first and second order derivatives of $p(\mu|y,\theta)$, with respect to μ, which are relatively easy to obtain given that $\log p(\mu|y,\theta) = \log p(y|\mu,\theta) + \log p(\mu,\theta) - \log p(y|\theta)$. The density $p(\mu|\theta)$ is Gaussian, $p(y|\mu,\theta)$ is analytically available, and $p(y|\theta)$ does not depend on μ. Conditional on this mode, the proposal density is set to be a multivariate normal with expected value equal to the calculate mode and variance equal to the negative of the inverse of the associated Hessian matrix.

Alternatively, Richard and Zhang (2007) propose to choose the mean and variance of the Gaussian approximating distribution $f(\mu|y,\theta)$ such that the Monte Carlo sampling variance in (3.23) is approximately minimized. To obtain this mean and variance, Richard and Zhang (2007) propose to recursively solve a set of auxiliary least squares optimization problems. They call this procedure Efficient Importance Sampling [EIS]. One can see EIS as using a global approximation to $p(\mu|y,\theta)$ instead of a local approximation at the mode as discussed above. In the context of state space models, Koopman and Nguyen (2012) further improve the computational efficiency of this method by using Kalman filter and smoothing methods to facilitate the optimization. Koopman et al. (2011) propose to use repeated (one-dimensional) numerical integration to improve the performance in obtaining the optimal mean and variance for the proposal.

The analysis in Salimans and Knowles (2013) shows that the typical implementation of EIS (setting its importance weights to one) corresponds to fitting a variational approximation similar to the type described in section 3.4.1. It is well known (e.g. Minka, 2005) that the tails of such an approximation tend to be thinner than the those of the target distribution. This is not a problem as long as the type of approximating distribution is extremely close to the target, but it will otherwise lead to infinite variance problems in the Importance Sampling step. For the truly high dimensional problems we considere here, the approximation cannot possibly be made accurate enough to avoid these problems,

making this approach infeasible.

Given a Gaussian approximation $f(\mu|y,\theta)$ to the true smoothed distribution $p(\mu|y,\theta)$, all elements in (3.23) are now relatively easy to calculate, and simulation from $f(\mu|y,\theta)$ can be done efficiently using the simulation smoother of Jong and Shephard (1995). The approximated log likelihood now gives the basis for a Maximum Likelihood estimator. However, for each evaluation of the log likelihood this entire procedure needs to be repeated. In large scale models this procedure can be very time consuming. Moreover, the estimator of the log likelihood is biased for finite L. This leads to a bias in the parameter estimates. In fact, in order for the bias to vanish asymptotically L needs to increase with the sample size. This makes this procedure computationally even more intensive. There are ways to correct for this bias, however in many cases such bias correction substantially increases the variance of the estimator. As a result the added value of this bias correction is unclear. In some cases the performance of the importance sampler can be improved by using antithetic and control variables, see Durbin and Koopman (1997).

In the Bayesian tradition a similar approximation to the density $p(\mu|y,\theta)$ is used. However, now this density is used as a proposal in a Metropolis Hastings sampler. In this approach the states are sampled together with the parameters, see for example Shephard and Pitt (1997).

Both approaches crucially rely on the quality of the approximation of $p(\mu|y,\theta)$ by $f(\mu|y,\theta)$. If the approximation is poor, one will need many draws in (3.23) to obtain a good estimate of the log likelihood or one will have very low acceptance rates in the Metropolis Hastings sampler. In the latter case, one can choose to apply blocking and sample μ in blocks (Shephard and Pitt, 1997) this increases the acceptance rate at the cost of poorer mixing. In the extreme case one may sample one state at a time (Carlin et al., 1992). However this procedure tends to have very poor mixing in many settings, see also Carter and Kohn (1994) for evidence on the performance of blocking in Gaussian state space models. Furthermore, the calculation of the approximating distribution and sampling from this distribution may still be computationally very demanding. In general one needs some replications of the (linear) Kalman filter and the calculation of the inverse of large matrices. Especially in the large dimensional cases that we consider, the computational costs may be excessive: with our examples in Sections 3.6 and 3.7 the state space

is so large than we cannot even keep the full covariance matrix of $p(\mu|y,\theta)$ in memory, which makes forming an truly accurate approximation $f(\mu|y,\theta)$ impossible. Since sampling methods in high dimensions fail without a very good proposal distribution $f(\mu|y,\theta)$ such approaches are not applicable on the scale of the problems that we are interested in. In practice, importance sampling tends to fail even with comparatively good proposal densities if the dimension gets high enough, as analyzed by Bickel et al. (2008); Snyder et al. (2008); Bengtsson et al. (2008); Lee et al. (2011); Beskos et al. (2011) among others. In general the number of samples required to reach a given accuracy scales exponentially in the dimension of the state vector, which becomes prohibitive extremely quickly. Note that our approach only relies on the approximation of the first two moments of multiple small subsets of μ, rather than having to faithfully approximate the whole distribution. This makes our approach fundamentally more robust in truly high dimensions.

In general, the importance sampling methods presented above provide a way to approximate the log likelihood function. However, to obtain this approximation many computations need to be done. The added value of all these computations is not clear. Most computations are aimed at adequately approximating the log likelihood. However, in the end we are interested in the parameter estimates and inference on the state variables. So, we should instead aim to approximate the maximizer of the log likelihood function. It is not clear how approximation error in the log likelihood function translates to errors in the maximizer.

In our approach we solve this issue by using the EM-EP algorithm. Our approximations are such that the approximation error will have a minimal impact on the final solution. The performance of the EM-EP algorithm depends on the quality of the approximation of the posterior mean and posterior variance of low dimensional subsets of the state vector. The EP part of the algorithm provides these approximations. In case the model is Gaussian, the EP algorithm exactly yields the posterior mean and variance, and our EM-EP algorithm exactly equals the maximum likelihood estimator. If the model is not Gaussian, EP yields an accurate approximation of the posterior mean and posterior variance. In other words, our approach is aimed at accurately approximating the solution instead of the problem.

Another advantage of our approach is that the maximization step of EM-EP is usually

trivial and does not require the use of the Kalman filter or smoother. This part can be done very efficiently. In the importance sampling approach one will need to use a numerical maximization algorithm to optimize the approximated log likelihood function over a (potentially) large set of parameters. At each function evaluation the entire approximation problem needs to be solved again. If the model is large in terms of parameters, number of states, or number of observations the IS approach becomes infeasible. Another added value of our approach is that it is deterministic and the properties of EM are well known. In the (E)IS approach a problem may be that the Monte Carlo variance in (3.23) may not exist, that is, it is infinite. In practice such a situation may be very difficult to detect.

3.5 Quality of approximation

In order to assess the quality of our approximate maximum likelihood estimator and its approximate standard errors, we performed a Monte Carlo study with a simplified version of the model used in Section 3.6. We consider a dynamic model for a non-normally distributed dependent variable. The dependent variable has a censored Poisson distribution, and the dynamics is specified as a random walk on the log Poisson rate. The model reads

$$\begin{aligned} y_t &\sim \text{Cens. Pois}(\lambda_t, \xi) && (3.24)\\ \log \lambda_t &\sim N(\log \lambda_{t-1}, \sigma_\lambda^2) && (3.25)\\ p(\lambda_1) &= 1/\lambda_1 \text{(i.e. diffuse on } \log \lambda_t), && (3.26) \end{aligned}$$

where Cens. Pois(λ_t, ξ) denotes a Poisson distribution with rate λ_t, censored from above at quantile ξ.

We simulate ten thousand artificial data sets $y_1, \ldots, y_T$ from this model, using $T = 500$, $\log \lambda_1 = 2$, $\xi = 0.9$ and $\sigma_\lambda^2 = 0.5T^{-1}$. These settings where chosen to match the characteristics of the data set used in Section 3.6. Since the Poisson is a discrete distribution, these settings correspond to an average censoring percentage of about 12%, rather than $1 - \xi$ exactly. For each simulated data set we obtained the EM-EP estimate of $\log(\sigma_\lambda^2)$ and its approximate standard error. The reason for working with $\log(\hat{\sigma_\lambda^2})$ instead of $\hat{\sigma_\lambda^2}$ is that the sampling distribution of the former converges faster to the corresponding

asymptotic normal distribution. The sampling distribution of the EM-EP estimator is shown in Figure 3.1. The sampling distribution of the approximate standard error is shown in Figure 3.2.

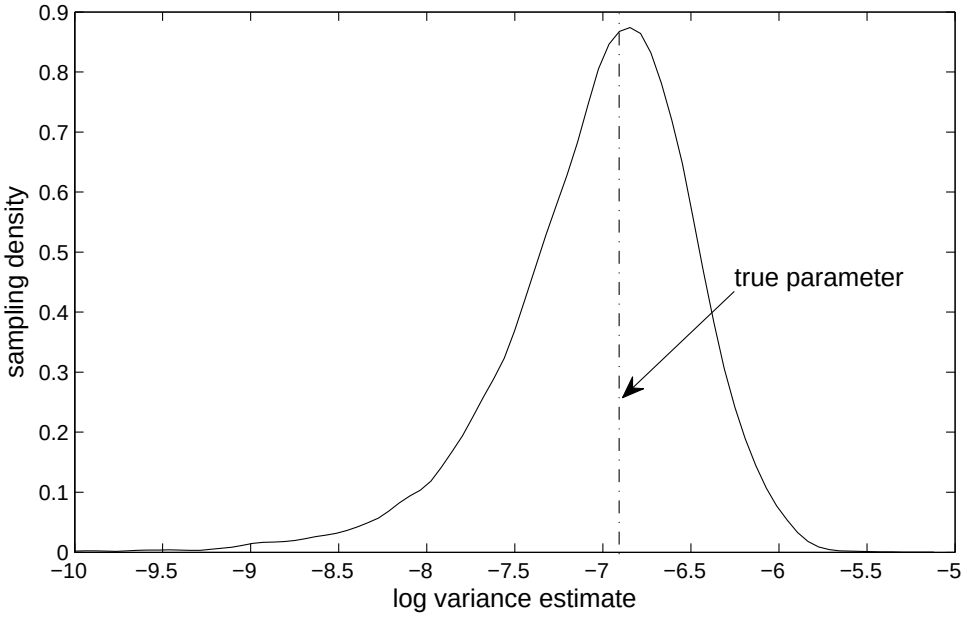


Figure 3.1: Sampling distribution of the QMLE for the censored Poisson model

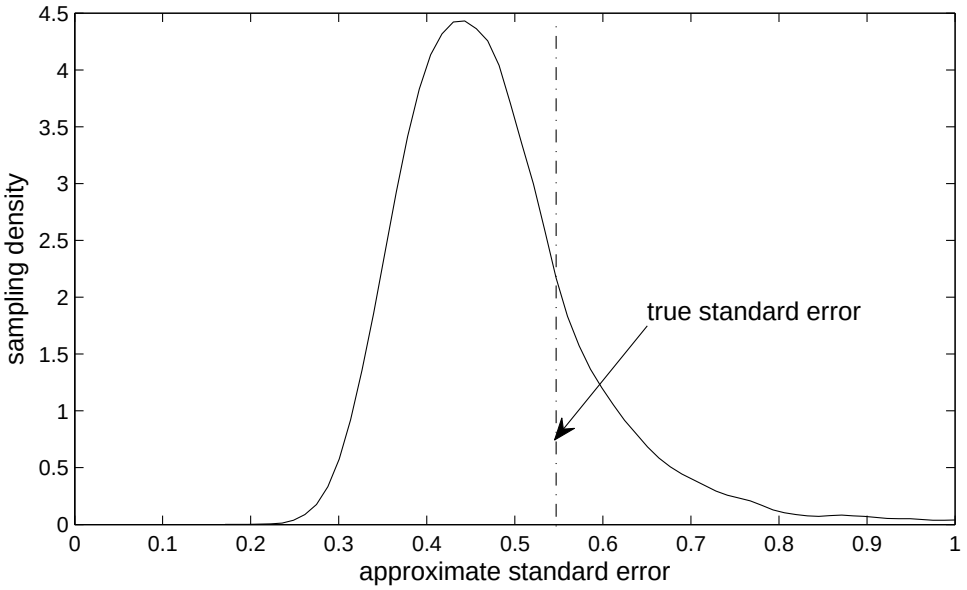


Figure 3.2: Sampling distribution of approximate standard error for the censored Poisson model

As can be seen in Figure 3.1, the sampling distribution of the EM-EP estimate is quite

close to a normal distribution, but with a heavier left tail. This non-normality is most likely caused by the small sample size, rather than the EM-EP estimation method. Most importantly, the EM-EP estimator is nearly unbiased in this application. Figure 3.2 shows that the standard error is slightly underestimated on average, which is again due to the thick left tail of the sampling distribution. The realized coverage of the asymptotic 95% confidence intervals based on these approximate standard errors is 91%, which means that the asymptotic confidence intervals are slightly too small. Note that such overconfidence is common to almost all maximum-likelihood based methods used in small samples. Overall, the asymptotic approximation fits the true sampling distribution quite well. This also means that the EM-EP estimator is quite efficient, otherwise it would have had a larger standard error than the asymptotically minimal standard error.

3.6 Forecasting newspaper sales

In this section we show the results of our first real-world application of the EM-EP method. The central problem in this section is demand forecasting for a national newspaper. We will argue why it is absolutely necessary to apply the EM-EP method, and we will show that applying the method can lead to a substantial increase in profit for the newspaper publisher.

Newspaper publishers usually distribute their paper to individual subscribers on a daily basis. Next to these subscriptions they sell individual newspapers at different outlets, such as supermarkets, gas stations, kiosks, etc. The sales at each outlet tend to be relatively small, but there are many outlets such that the total sales of the individual newspapers tends to be substantial. For the publisher it is important to supply each outlet with the "correct" number of newspapers. Supplying less than the demand will lead to a stock-out, supplying too much is also not efficient. To be able to calculate the number of papers to supply, the publisher needs a model of the demand for newspapers at each outlet. The demand is influenced by seasonality, special editions of the newspaper, and many other factors. A dynamic model is likely to fit the demand well, as past sales are very informative to predict future demand.

For each outlet j and each time point t, the publisher knows how many newspapers

were supplied, denoted by $W_{j,t}$, and how much papers were *not* sold. This means that the actual sales, denoted by, $S_{j,t}$, can be calculated which of course is related to the demand ($D_{j,t}$) as

$$S_{j,t} = \begin{cases} D_{j,t} & \text{if } D_{j,t} < W_{j,t} \\ W_{j,t} & \text{if } D_{j,t} \geq W_{j,t}. \end{cases} \tag{3.27}$$

As mentioned above, the demand per outlet tends to be relatively small so that a count model is necessary. We specify a Poisson distribution for $D_{j,t}$ such that the sales are distributed as a censored Poisson with exogenous censoring $W_{j,t}$. The log Poisson rate for store j at time t is given by $\log \lambda_{j,t}$ and is assumed to follow a Gaussian linear dynamic model. This dynamic model may also specify a dependence structure among the log Poisson rates of outlets.

Given the demand model, the costs of supplying newspapers, the price of a newspaper, and the opportunity costs of stock-out it is relatively straightforward to work out the supply that maximizes the expected profit for each outlet.

3.6.1 Data

We analyze a data set containing sales data on a major newspaper in the Netherlands, spanning a time period of over two years. During (part of) this time 9000 different outlets were active selling this newspaper. Most stores sold between 0 and 10 newspapers daily, with a few larger stores selling an average of up to 50 newspapers daily. At a typical store, a stock-out occured between 15% and 20% of the time, more frequently for the smaller outlets than for the larger ones.

In addition to the sales of the newspapers, the data set contains a list of all promotions and discounts that occured throughout the time period. Also, a list was compiled with all special news events that led the publisher to increase the supply of newspapers. Finally, the location of each outlet is available, together with a variable indicating whether this location was seasonal or not. An example of a seasonal location would be the beach.

3.6.2 Model

For the purpose of forecasting the newspaper demand we construct a dynamic model with latent Gaussian variables and a non-Gaussian dependent variable. The most important characteristics of the distribution of the newspaper sales are that the sales numbers are discrete and that they are regularly censored from above, i.e. when the newspaper is sold out. To capture these characteristics we model the newspaper sales with a censored Poisson distribution, that is,

$$S_{j,t} \sim \text{Cens. Pois}(\lambda_{j,t}, W_{j,t}), \tag{3.28}$$

where the sales at store j and time t are censored from above by the supply of newspapers $W_{j,t}$. In order to determine the optimal supply of newspapers, we need to forecast the demand, or the potential sales, which we denote by $D_{j,t}$. The relationship between sales and demand was given in (3.27), which implies the following model for the demand:

$$D_{j,t} \sim \text{Pois}(\lambda_{j,t}). \tag{3.29}$$

The sales data show persistent shifts over time for the different outlets in our sample, suggesting a dynamic process for the Poisson-rate of the demand $\lambda_{j,t}$. In addition, the sales data are positively correlated across stores: on some days the outlets sell many newspapers on average, for example due to some event in the news, and on other days they sell very little, for example because it is a very rainy day with few people passing by the stores. The different stores also share seasonal effects, e.g. they sell more in summer than in winter and more on Saturday than on Wednesday. These shared effects seem to influence sales proportionally rather than additively, with larger stores selling more additional newspapers on Saturday than smaller stores. In order to capture these correlation patterns, we model the log Poisson rate $\log \lambda_{j,t}$ with a dynamic additive factor model. Multiple specifications were examined, but all models have the following stylized form:

$$\log \lambda_{j,t} = a_{j,t} + b_t + \sum_{i=1}^{6} \mathbb{I}[t = \text{day of week } i] c^i_{j,t} + \sum_{i=1}^{n} \mathbb{I}[t \in \mathcal{D}^i] d^i_t + \sum_{i=1}^{p} \mathbb{I}[(t,j) \in \mathcal{E}^i] e^i_t. \tag{3.30}$$

We distinguish five different kinds of factors in the above model:

1. A single factor $a_{j,t}$ for each separate store that applies on all days in the sample. This factor captures effects such as the shop-size and the trend in the mean sales level for a particular store.

2. A single factor b_t applying to all stores in the sample. This variable captures the global trend in newspaper sales over stores. We found that this factor seems to gradually trend down over time, most likely reflecting the loss of newspaper sales due to online alternatives.

3. Store-specific weekly seasonality factors $c^i_{j,t}$ for each day of the week. These factors allow us to model how the sales pattern for a given store departs from the weekly seasonality of other stores. For example, most stores sell more newspapers on Saturday than on Monday, but this is not true for all stores in the sample. The factors $c^i_{j,t}$ are specified to be dynamic in order to capture the fact that the weekly sales pattern of a store sometimes shift over time.

4. Time specific global factors d^i_t that apply to all stores, but often only on given days (as indicated by the indicator function $\mathbb{I}[t \in \mathcal{D}^i]$). This type of factor is used to model the effects of major news events, country-wide promotions, and possible country-wide unobserved effects. In addition, these factors capture the seasonality patterns that are common to all stores: for example, almost all stores sell more newspapers in summer than in winter.

5. Dynamic factors e^i_t that apply to more than one store, but not all (as indicated by $\mathbb{I}[(t,j) \in \mathcal{E}^i]$). For example, a given factor may only apply to stores at specific seasonal locations, such as beach, since such stores obviously share some of the same seasonal effects. Another example is a factor that applies to all outlets in a particular city, which may sell more as a result of a specific local news event such as the local football club winning a match.

The individual factors in $(a_{j,t}, b_t, c^i_{j,t}, d^i_t, e^i_t)$ are generally modeled using a random walk

$$f^i_{j,t} \sim N(f^i_{j,t-1}, \sigma^2_i), \tag{3.31}$$

where $f_{j,t}^i$ denotes one of the individual factors. More general autoregressive specifications such as those described in Section 3.2 are also possible in our framework, but for our particular application we found that all persistent effects seem to be modeled well by a simple one-dimensional random walk. An exception are a small number of factors that describe unique one-time events as identified by the publisher. Such factors are modeled as a single independent normal random variables $f^i \sim N(m_i, v_i)$. The parameters of the distribution of such one-time factors are elicited from the publisher, while the parameters of all dynamic factors are estimated from the data.

The initial values of the dynamic factors are modeled as

$$f_{j,0}^i \sim N(\mu_i, \eta_i) \tag{3.32}$$

By using such an 'informed' initialization of the Kalman filter, rather than a diffuse initialization, we are able to shrink all store-specific variables to common values. Doing so allows us to more quickly provide reasonable forecasts for new stores that are opened. Since these initial distributions generally apply to multiple factors (e.g. multiple shops) we can estimate their parameters from the data.

Using this standardized model specification allowed us to write very general computer code which we used to test many different combinations of factors. Our most succesful model specifications contained 10-20 different factors. The most important factors turned out to be related to the size of the newspaper outlet, seasonalities (both yearly and weekly), and breaking news events.

3.6.3 Inference

The model specified above describes the evolution of a state vector with dimensionality in the tens of thousands, which is challenging for any inference algorithm. The observation equation (3.28) is non-Gaussian which means that the model is intractible for exact likelihood based methods. The conditional distribution of the persistent elements of the state vector displays strong correlation over time, which makes Monte Carlo inference very difficult. Efficient simulation smoothers exist, but these can not easily be applied in this problem, because the likelihood is non-Gaussian and because the state variables are

also correlated between shops.

The EM-EP algorithm proposed here is able to very quickly provide a good approximation to the conditional distribution of the state vector using the EP step. This approximation can then be used to estimate the mean and variance parameters of the model using the EM step. However, the correlation between shops, induced by their shared factors, also gives computational problems for our algorithm. The full covariance matrix is too large to be used in an exact step of the Kalman filter. However, our algorithm is still able to provide a good approximation to the conditional distribution by deleting these covariance elements from the covariance matrix before each step of the Kalman filter and smoother. The shared factors have a support of thousands of shops and are thus quite well determined from the data compared to the shop-specific factors. This means that the deleted covariance elements are very small and can safely be ignored. The resulting algorithm has a runtime in the order of hours, which is fast enough to be used for practical purposes. For additional speed we can take the approximation one step further and also delete the other (shop-specific) off-diagonal elements of the covariance matrix before each step of the Kalman filter. Our experiments indicate that this does not hurt the predictive ability of the model very much, while further decreasing the computation time to the order of minutes. The final result is an algorithm that allows us to quickly make good forecasts using an advanced high dimensional non-Gaussian dynamic factor model that is imposible to estimate using conventional methods.

3.6.4 Results

The EM-EP algorithm allows us to quickly and efficiently apply the model presented above. Besides quasi maximum likelihood estimates of the parameters, the algorithm also provides smoothed and filtered estimates of the latent factors. These latent factor estimates allow us to visualize the developments for a shop over time and to forecast future newspaper demand.

Figure 3.3 shows the development of the smoothed λ_{jt} for a typical store. Apparent are the periodic fluctuations that are due to the weekly seasonality: Like most stores this particular store sells most of its newspapers on Saturday. In addition, the smoothed λ_{jt}

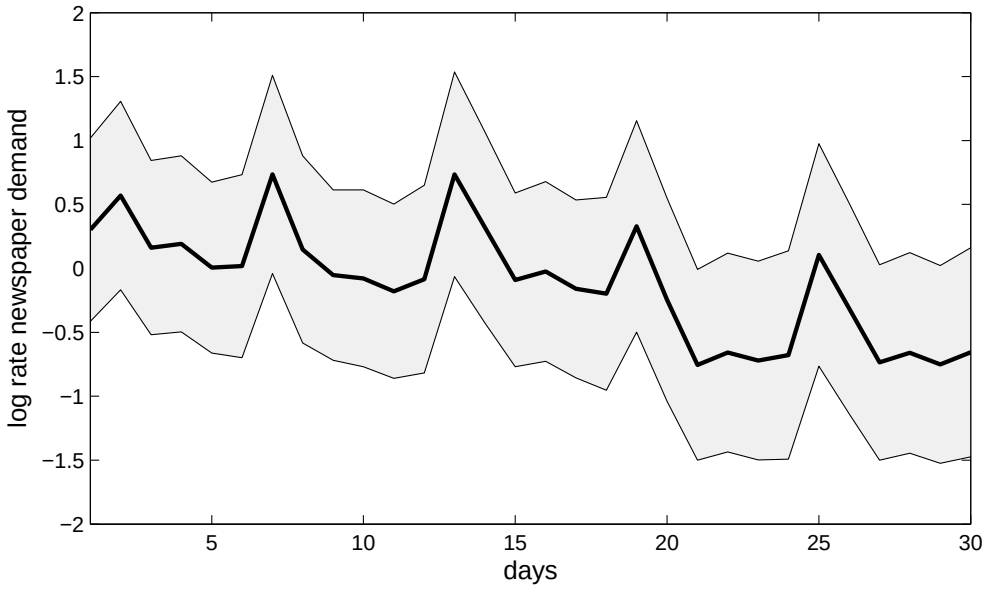


Figure 3.3: Posterior mean and 95% posterior confidence region for the log rate of the newspaper demand at a particular store

show a gradual downward trend in the number of sales for this store. This trend is shown more clearly in Figure 3.4 which shows the development of the base level of sales for this store.

Figures 3.3 and 3.4 show that the uncertainty in the newspaper demand for this store is quite large. This observation is typical for the newspaper outlets in our dataset: most outlets in the sample have only been selling the newspaper for a short period of time, and the observations are not very informative since the number of newspapers sold is often quite low (0-3) and censored from above.

By comparing Figures 3.3 and 3.4 we see that the uncertainty in the log rate for the newspaper demand is largely due to the uncertainty in the 'shop size' factor for this store. The outlets in the sample show frequent persistent shifts in the demand for newspapers at the individual stores, which is why our EM-EP estimate of the variance parameter for this factor is large. On the other hand, the weekly seasonality in sales seems much more stable, and is also largely equal across stores, which is why this factor is much more certain.

Using the filtered forecast distribution for the λ_{jt}, we can calculate predictive distributions for the demand $D_{j,t}$ at each store. By taking into account the number of newspapers $W_{j,t}$ that are provided to each store, we can also predict the number of newspaper sold. The predicted sales have a correlation of 0.78 with the true sales. This correlation is

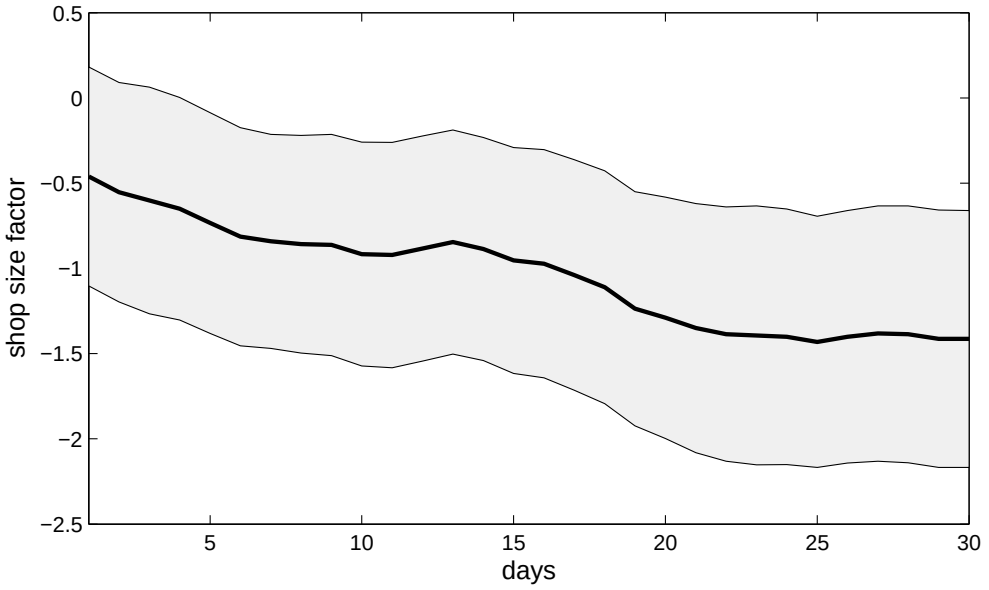


Figure 3.4: Posterior mean and 95% posterior confidence region for the 'shop size' factor of the store

higher than that for the predictive model that the publisher was using, and indicates that the model has good predictive power. The predictions were approximately unbiased.

Given the predictive distribution for the demand $D_{j,t}$, the optimal number of newspapers to deliver ($W_{j,t}$) can be determined by maximizing the expected profit. We assume that profits are linear in both the number of newspapers that are delivered and the number that are sold. The profit function becomes

$$\pi(D_{j,t}, W_{j,t}) = \alpha \min(D_{j,t}, W_{j,t}) - \beta D_{j,t}, \tag{3.33}$$

where α captures both the direct profit from selling a newspaper as well as the expected long term profit from potentially acquiring a new customer, β consists of the costs made in printing and delivering the newspaper. Both parameters were elicited in discussions with the newspaper publisher and by looking at the preferences implied by their current distribution decisions. The indirect part of the profit in α is substantial. The newspaper publisher had precise figures for the β parameter.

Given the linear profit function and the log-normal predictive distribution for λ_{jt}, the expected profit will be a unimodal concave function of $W_{j,t}$ that is easily optimized. Figure 3.5 shows the expected profit curve together with the predictive distribution for $D_{j,t}$.

We used this optimization procedure to generate out-of-sample delivery decisions $W_{j,t}$. The results show that the proposed model and estimation technique could increase the publisher's profits by 10-15 % in comparison with their current predictive system. The increase in profit comes primarily from delivering fewer newspapers to those stores that are unlikely to sell all of them.

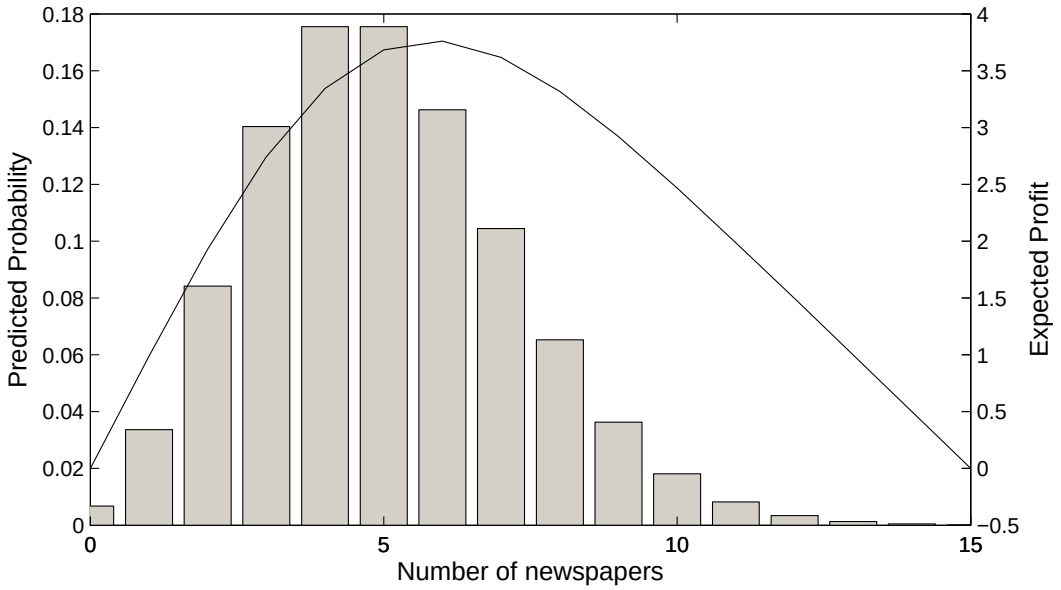


Figure 3.5: Expected profit as a function of delivered newspapers

3.7 Forecasting chess matches

In this section we apply our EM-EP methodology to a model to predict outcomes of chess matches. Herbrich et al. (2007) introduce a dynamic latent Gaussian model for assigning ratings to players in the Xbox Live game Halo 2. Dangauthier et al. (2008) apply this model to the problem of rating chess players. Their model assumes that the outcome of a game of chess is determined by the skills of the players which can be described by a single number. They infer these skills using factored approximate Bayesian inference using Expectation Propagation as discussed in Section 3.3. However, they do not mention how to infer the parameters of their model. Here we propose to use approximate Expectation Maximization to infer these parameters.

3.7.1 Data: the Deloitte/FIDE chess rating challenge

From February 7 to May 4, 2011 the *Deloitte/FIDE chess rating challenge* was held at Kaggle.com[1], a platform for hosting statistical prediction competitions. The Deloitte/FIDE chess rating challenge was a prediction contest sponsored by services company Deloitte Australia and by the world chess federation FIDE. The aim of the contest was to develop a statistical system to forecast the results of chess matches. The training data for this competition contains over 2 million matches played by over 50 thousand players over a period of 132 months. The objective was to use this data to forecast the results of 100,000 matches that were played in the three months following this period.

The competition attracted 189 teams from all over the world. The competition was won by the first author, using a combination of techniques including the EM-EP method discussed in this chapter. Below we present the basic model underlying the winning entry. The model is largely based on the TrueSkill model of Herbrich et al. (2007). Details on the exact forecasting approach that was used can be found on the homepage of the first author.

3.7.2 Model

Let us index observed chess matches by i, with $i \in [1, \ldots, n]$ and n the number of matches, index time with $t \in [1, \ldots, T]$ and T the number of months in the sample, and index players by j with $j \in [1, \ldots, k]$ and k the number of players. We will then describe player j's skill by a single number $s_{j,t}$. This skill is assigned a Gaussian prior and is assumed to evolve according to a random walk.

$$\begin{aligned} s_{j,1} &\sim N(0, \sigma_s^2) \\ s_{j,t+1} &= s_{j,t} + \nu_{j,t}, \quad \nu_{j,t} \sim N(0, \sigma_\nu^2) \end{aligned} \tag{3.34}$$

For a given match i, let w_i denote the index of the player playing white, let b_i denote the index of the player playing black, and let t_i denote the time at which this match is played.

[1] http://www.kaggle.com/c/ChessRatings2

The outcome of match i (m_i) is then modeled as follows

$$
\begin{aligned}
d_i &= s_{w_i,t_i} - s_{b_i,t_i} + \gamma + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \\
m_i &= \begin{cases} \text{player } w_i \text{ wins} & \text{if } d_i > 1 \\ \text{player } b_i \text{ wins} & \text{if } d_i \leq -1 \\ \text{players } w_i \text{ and } b_i \text{ draw} & \text{if } -1 < d_i \leq 1 \end{cases},
\end{aligned}
\tag{3.35}
$$

where d_i can be seen as the performance difference between players w_i and b_i in this match, γ is the advantage of playing white, and ϵ_i is an error term.

After observing a set of match results m, a factored approximate posterior distribution for the skills s, the skill innovations ν, and the errors ϵ can be obtained using Expectation Propagation (see Dangauthier et al. (2008) for details). This factored approximate posterior corresponds to deleting the off diagonal elements in the Kalman covariance matrix as explained in Section 3.3.2. The first and second moments of this approximate posterior can then be used to infer the advantage of playing white (γ), the prior skill variance σ_s^2, the variance of the skill innovations σ_ν^2 and the variance of the noise term σ_ϵ^2 by approximate Expectation Maximization as described in Section 3.3. The maximization updates are given as follows

$$
\begin{aligned}
\gamma_{\text{new}} &= \gamma_{\text{old}} + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\epsilon_i | m) \\
\sigma^2_{\epsilon,\text{new}} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\epsilon_i^2 | m) - \mathbb{E}(\epsilon_i | m)^2 \\
\sigma^2_{s,\text{new}} &= \frac{1}{k} \sum_{j=1}^{k} \mathbb{E}(s_{j,1}^2 | m) \\
\sigma^2_{\nu,\text{new}} &= \frac{1}{kT} \sum_{t=1}^{T} \sum_{j=1}^{k} \mathbb{E}(\nu_{j,t}^2 | m).
\end{aligned}
\tag{3.36}
$$

3.7.3 Results

The EM-EP algorithm allowed us to estimate the model parameters in minutes, while exact maximum likelihood estimation is intractable because of the lack of conjugacy and because of the very high dimension of the state space (over 50,000 players per time pe-

riod). The predictive performance of the model is state-of-the-art as evidenced by our victory in the Deloitte/FIDE chess rating challenge. As an example of the obtained inference we present the development of the latent skill of a randomly selected chess player in Figure 3.6. The figure shows large changes in the inferred skill of this player in months 4 and 9. These events correspond to unexpected wins or draws by this player in matches in these months.

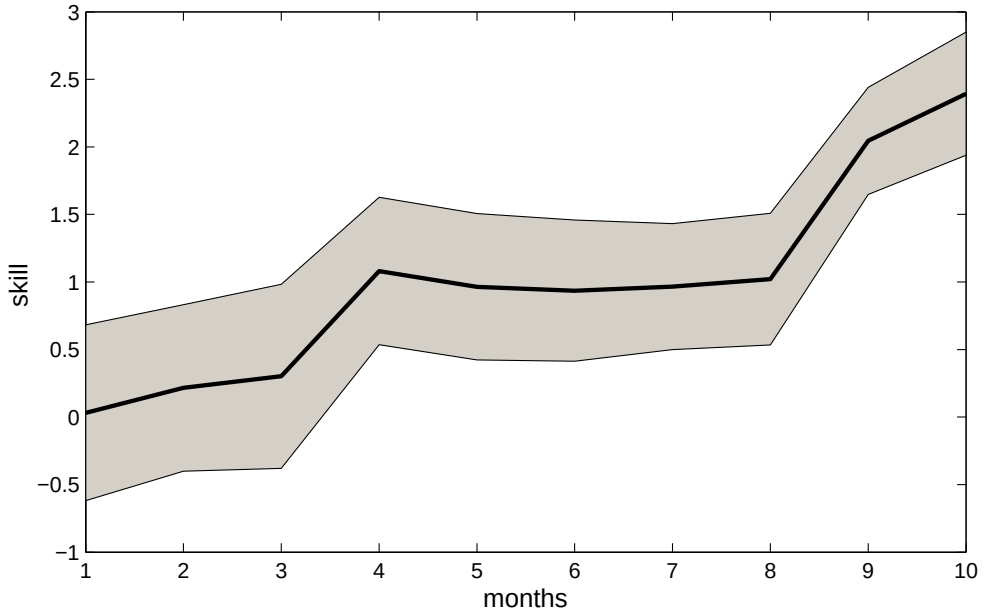


Figure 3.6: Filtered expectation and 95% confidence region for the skill of an anonymous chess player

3.8 Conclusion

State-space models have proven to be very useful in many situations. Standard techniques are available for the fully Gaussian case. These techniques even perform well when the state space is very large. However, these methods break down when the dependent variable does not have a conditional Gaussian distribution. In this chapter we consider exactly this case: large non-Gaussian dynamic models.

Our estimation procedure combines the well-known Expectation Maximization [EM] algorithm with the Expectation Propagation [EP] approach. In the non-Gaussian case, the Expectation step of the EM algorithm is usually not tractable. The EP approach allows us to efficiently obtain a good approximation of the needed conditional expectation. We

show that combination of the two methods, labeled as EM-EP, gives a very flexible and useful estimation methodology for non Gaussian dynamic models.

Although the EM-EP method relies on approximations, and therefore should be seen as a quasi Maximum Likelihood procedure, we have shown using simulations that the method is approximately unbiased and the resulting (asymptotic) standard errors are useful.

Finally, our method also shows to be useful in practice. We have applied the methodology to two problems. In the first, we model the time series of sales of newspapers across 9,000 outlets. The combination of the model and the estimation method result in a potential profit increase of about 10 to 15% for the publisher. In the second application, we model the results of chess matches player by 50 thousand players. In this case the model and the EM-EP methodology proved to deliver outstanding predictive performance.

We believe that the EM-EP method is a promising estimation method for a wide variety of large scale, non-Gaussian dynamic models. We hope that our discussion in this chapter and the examples lead others to apply and further improve this methodology.

Chapter 4

Variable selection and functional form uncertainty in cross-country growth regressions

4.1 Introduction

Many economic studies aim to determine the driving factors of economic growth. Following the seminal work of Kormendi and Meguire (1985) and Barro (1991), an important tool in this endeavor has been the cross-country growth regression, i.e. the use of regression analysis to determine what variables are correlated with economic growth in a cross-section of countries. The literature has identified two major problems with this technique. The first is that there is only a limited number of countries and a potentially very large number of variables to explain economic growth. The decision of which variables to include in the regression therefore has a strong influence on the conclusions that are drawn from the analysis. Since this decision is often guided by nothing but the whim of the researcher, there is no guarantee that these conclusions are not the product of data mining and selective presentation of data (see Leamer (1983) and Geweke (2005, sections 8.4 and 8.5)).

The second objection raised against cross-country growth regression is that most studies unreasonably restrict attention to the set of linear regression models. The linear re-

gression model complies with the classical Solow model (see Mankiw et al., 1992) which specifies that log output is an additive linear function of technology, capital and labor. However, a range of new growth models, collectively known as *new growth theory*, pose the existence of multiple steady states in economic growth (see Aghion et al., 1999; Azariadis and Drazen, 1990; Durlauf, 1993). Although these models typically specify the growth path of each country to be linear in its variables, the slopes of the growth path depend on which steady state the country is in. Since the steady state of a country depends on its initial conditions, such as its level of economic development and human capital, the process determining economic growth in these models is nonlinear in the regressors. In addition to new growth theory, some versions of neoclassical growth theories also give rise to parameter heterogeneity (Binder and Pesaran, 1999; Barro and Sala-i-Martin, 2003).

Both issues have received much discussion in the literature, however only rarely in the same paper. Yet, a joint treatment of these two sources of model uncertainty is absolutely essential: Variable selection methods do not necessarily select the same variables under different model specifications, and evidence of nonlinearity may not hold up under different variable selections. To examine these issues, this chapter presents an integrated analysis of variable selection and functional form specification in cross-country growth regressions. We perform this joint treatment by extending the linear growth regression model to explicitly allow for parameter heterogeneity as suggested by new growth theory, while simultaneously addressing the variable selection problem by performing Bayesian model averaging. Estimating the new models on the data sets of Sala-i-Martin et al. (2004) and Fernandez et al. (2001b) provides evidence of multiple-regime parameter heterogeneity of the type predicted by new growth theory and empirically documented by Durlauf and Johnson (1995) and Liu and Stengos (1999). In addition, we find that many of the explanatory variables indicated by the literature do not have robust marginal effects across countries when allowing for a more flexible model specification, contradicting the results of Minier (2007). Our results offer some new insights into the form of the parameter heterogeneity in the growth data and we discuss its connection to phenomena like the natural resources curse.

The outline of the remainder of this chapter is as follows. Section 4.2 provides a short review of the existing literature on growth regression. The statistical methodology of the

chapter is explained in Section 4.3, where we introduce a new set of models that allows for multiple-regime nonlinearity. In Sections 4.4 and 4.5 we present the estimation results for these models and compare them with the linear model specification. Finally, Section 4.6 concludes.

4.2 Robustness in growth regressions

The large body of literature on cross-country growth regression started with the work of Kormendi and Meguire (1985), Grier and Tullock (1989) and Barro (1991). Since then the literature has identified a large number of variables correlated with economic growth. However, these variables were not discovered by the analysis of an ever greater amount of data, but rather by the specification of an ever greater amount of different models, casting doubt onto the statistical validity of these findings. Attention to this problem was first raised by the influential paper of Levine and Renelt (1992), who investigated the robustness of earlier findings to different model specifications by employing a variant of the extreme bounds analysis of Leamer (1983). This analysis proceeds by estimating the coefficient of a regressor in many different linear regression models, each controlling for a different subset of regressors, and analyzing the different results. If the regressor of interest is found to be significantly different from zero in each regression, with the same sign, the influence of the regressor is called *robust*. Otherwise it is called *fragile*. Levine and Renelt (1992) (henceforth LR) found that many of the relationships uncovered in earlier work on growth were in fact fragile.

The paper by LR was followed by many responses from the growth community. Although widely appreciated for bringing attention to the explosion of different model specifications in the growth literature, its methodology received criticism from several authors. An influential response was the one by Sala-i-Martin (1997), who argued that the extreme bounds approach was overly stringent. An important property of the approach is that a negative result from a single model specification can potentially negate the positive results from a much larger number of models, even if those other models fit the data much better. Sala-i-Martin argued that this property, combined with the large number of different model specifications, was almost guaranteed to produce the negative results reported by

LR. As an alternative method of providing robust inference, he proposed to instead look at the average result of the regressions, with each model receiving a weight proportional to its data likelihood. His follow-up paper (Sala-i-Martin et al., 2004) further developed this approach by deriving a new weighting method based on Bayesian model averaging using a large sample approximation for the model weights. Another contribution using the concept of model averaging is Fernandez et al. (2001b) who presented a formal Bayesian analysis without such an approximation.

A second criticism the LR study received is that it unreasonably restricted attention to the set of linear models, while new growth theory predicts a nonlinear relationship between growth and the explanatory variables. This criticism was backed up by empirical evidence provided by Durlauf and Johnson (1995), among many others, who documented the existence of multiple-regime parameter heterogeneity. By performing a tree regression, they allocated the countries in their data set to multiple regimes based on initial conditions related to the levels of economic development and human capital of the country. Liu and Stengos (1999) confirmed these results by estimating a classical semi-parametric model on the LR dataset, modeling the same kind of multiple-regime nonlinearity. They found that their nonlinear model improved upon the linear specification. Additional evidence in support of heterogeneity was found by Paap et al. (2005) and Basturk et al. (2012) who modeled economic growth using mixtures of linear regression models. While these studies took into account the uncertainty in the functional form of the growth equation, they considered only small fixed sets of explanatory variables, ignoring the uncertainty in the variable selection process.

To the author's knowledge, Minier (2007) and Cuaresma and Doppelhofer (2007) are the only attempts to date at combining both sources of uncertainty in an analysis of cross-country growth data. Minier (2007) investigated the influence of nonlinearities in the fragile variables of LR by repeating their analysis, but subsequently introducing quadratic and interaction terms into the specification as well as allowing for different growth regimes by splitting the sample according to initial conditions. By doing so, several more variables relating to fiscal policy appeared robust in her specification compared to the original LR analysis. However, in determining the robustness of the variables she only looked at the parameters of the linear terms, ignoring the coefficients of the higher order regressors.

This ignores the fact that, through the quadratic and interaction terms, the 'robust' regressors may very well have marginal effects with different signs in different specifications, which makes her conclusions difficult to compare with the original LR results.

Cuaresma and Doppelhofer (2007) extended the approach of Sala-i-Martin et al. (2004) to allow for threshold effects in the model specification. Similar to the analyis of Durlauf and Johnson (1995), they effectively split the sample based on explanatory variables, but instead of defining the splitting thresholds a priori they estimated them together with the regression coefficients. Using this approach, they performed Bayesian model averaging over the subset of variables found to be robustly correlated with growth by Sala-i-Martin et al. (2004). Contrary to the results of Durlauf and Johnson (1995), Liu and Stengos (1999) and others they did not find much evidence for nonlinearity. It is worth investigating whether this is due to their particular model specification or their (limited) consideration of variable selection uncertainty. In addition, a further investigation into the influence of possible nonlinearities on the variable selection problem is needed.

For a different application, Hoeting et al. (2002) proposed dealing with the functional form uncertainty by expanding the model space to include a number of different transformations of the explanatory variables. The variable selection problem can then be solved by performing Bayesian model averaging in this larger model space. This is a simple and elegant solution, but Hoeting et al. (2002) do not consider interaction terms between the variables. In principle, allowing for such interactions is possible using their framework, but this would lead to an impractically large model space for our current application.

4.3 Statistical Methodology

Following the tradition of the growth literature, as exemplified by Barro (1991), Levine and Renelt (1992), Sala-i-Martin et al. (2004) and many others, we will consider linear regression models of the form

$$y_i = \alpha + \beta_{i,1}x_{i,1} + \beta_{i,2}x_{i,2} + ... + \beta_{i,p}x_{i,p} + \epsilon_i, \qquad \epsilon_i \sim NID(0, \sigma^2), \tag{4.1}$$

where y_i is the i-th country's long term growth rate, with i ranging from 1 to n, $\{x_{i,1}, ..., x_{i,p}\}$ is a collection of p explanatory variables, α is an intercept, $\{\beta_{i,1}, ..., \beta_{i,p}\}$ is a collection of (possibly country-specific) regression coefficients and ϵ_i is an error term. Most recent work on growth regressions has focused on the selection of explanatory variables to include in this model (Fernandez et al., 2001b; Sala-i-Martin et al., 2004) or on the specification of country-specific parameters (Kalaitzidakis et al., 2000; Minier, 2007; Paap et al., 2005; Basturk et al., 2012; Maasoumi et al., 2007). A rigorous regression analysis of cross-country growth data should simultaneously take into account both of these sources of uncertainty, which is the main contribution of the present work. Note that we do maintain the assumption of homoskedasticity made in these earlier studies in order to allow for comparison and to single out the influence of model uncertainty on the inference. Relaxing this assumption would be a logical next step as discussed in section 4.6.

Bayesian analysis is ideally suited for an analysis of model uncertainty as it offers a systematic method of quantifying this uncertainty that is not offered by classical statistics. Indeed, the majority of the literature dealing with these issues in growth regressions builds on the Bayesian framework (e.g. Levine and Renelt (1992), Sala-i-Martin (1997), Fernandez et al. (2001b), Sala-i-Martin et al. (2004), Minier (2007) and many others) as does this work.

4.3.1 Variable selection

The first problem to be addressed is the selection of explanatory variables to include in the model. If our data set contains k potential explanatory variables, we have 2^k different possible subsets of regressors to include in our model. We denote the variable selection by s, a $k \times 1$ binary vector, with $s_j = 1$ if the j-th explanatory variable is included in the model and $s_j = 0$ if it is not. All 2^k possible variable selections usually represent reasonable models for economic growth and we cannot be sure a priori which subset of variables we should use. We therefore proceed by assigning a prior distribution $\mathrm{P}(s)$ to the variable selection indicator vector s. Given the data $y = (y_1, ..., y_n)'$, we can then obtain posterior variable selection probabilities $\mathrm{P}(s|y)$ by applying Bayes' rule. Our final conclusions about economic growth can then be obtained by averaging over all variable selections and

weighing each possible selection by its posterior probability. This procedure is known as *Bayesian model averaging* (see Mitchell and Beauchamp, 1988; Raftery et al., 1997) and was also used by Sala-i-Martin (1997), Fernandez et al. (2001b) and Sala-i-Martin et al. (2004) to study economic growth. Following these earlier studies, we define our prior over models by assigning each potential explanatory variable an independent prior inclusion probability of θ. This gives the following prior distribution for the variable selection vector s:

$$\mathrm{P}(s|\theta) = \prod_{j=1}^{k} \theta^{s_j}(1-\theta)^{1-s_j}. \tag{4.2}$$

As can be seen from (4.2), the prior probability of a particular variable selection only depends on its number of included regressors $p_s = \sum_{j=1}^{k} s_j$. Assigning independent prior inclusion probabilities leads to a binomial prior on p_s:

$$\mathrm{P}(p_s|\theta) = \binom{k}{p_s}\theta^{p_s}(1-\theta)^{k-p_s}, \quad \text{for } p_s = 1,\ldots,k. \tag{4.3}$$

Common choices for θ are 1/2, as used by Fernandez et al. (2001b), or 7/k as used by Sala-i-Martin et al. (2004). It is not obvious a priori what a good value for θ should be and the analysis can be quite sensitive to this value. Ley and Steel (2009) show that a more robust choice is to specify a Beta-hyperprior on θ. If the hyperprior on θ is Beta(a, b), this corresponds to a Binomial-Beta distribution on the model size p_s:

$$\mathrm{P}(p_s) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+k)}\binom{k}{p_s}\Gamma(a+p_s)\Gamma(b+k-p_s), \quad \text{for } p_s = 1,\ldots,k. \tag{4.4}$$

If our prior on θ is uninformative (i.e. a and b are low), the analysis of Ley and Steel (2009) suggests that the results of the model averaging should be relatively insensitive to the exact choice of a and b, which we can confirm for our application. We choose to express our prior ignorance about θ by choosing $a = b = 1$ which gives uniform prior distributions on θ and p_s. The resulting prior on s is then given by

$$\mathrm{P}(s) = \int \mathrm{P}(s|\theta)p(\theta)d\theta = \frac{1}{k+1}\binom{k}{p_s}^{-1}. \tag{4.5}$$

The prior probability of a variable selection is now inversely proportional to the number of models with the same number of variables. This prior corresponds to the *Bayesian multiplicity adjustment* proposed by Scott and Berger (2010).

4.3.2 Prior specification for the regression parameters

The uncertainty due to the unknown β parameters can also be quantified in terms of a prior distribution. Except for the empty model $s_j = 0 \ \forall j$, all models contain more parameters than we have observations (see equation (4.1)). The specification of this prior distribution will thus have an important effect on our analysis. The earlier studies by Sala-i-Martin (1997), Fernandez et al. (2001b) and Sala-i-Martin et al. (2004) all assume that the β parameters are the same across countries. Subject to this constraint, they use the popular conjugate g-prior introduced by Zellner (1986) to complete the specification. The advantage of this prior is its analytical tractability and its invariance to the scale of the regressors. Stacking the $\{\beta_{i,1}, ..., \beta_{i,p_s}\}$ parameters into a vector β_i, this prior is given by

$$\beta_i | s, \sigma^2 \sim N[0, \sigma^2 (g X_s' X_s)^{-1}]. \tag{4.6}$$

where X_s is the $n \times p_s$ matrix of explanatory variables included in the model and g is a parameter to be set by the researcher. Note that β_i only denotes the coefficients for the regressors that are included in the model. The coefficients that are not included in the model have a prior unit point mass at zero. Sala-i-Martin et al. (2004) set $g = 1/n$ and then take an approximation based on the sample size n becoming large. Alternatively, Fernandez et al. (2001b) set the g parameter to $1/k^2$, where k is the number of candidate regressors. They choose this constant because they find it generally leads to good estimation results, as they show in Fernandez et al. (2001a). We choose $g = 1/n$ to facilitate comparison with Sala-i-Martin et al. (2004) and because this choice has the intuitively attractive property of keeping the scale of the prior variance constant when the sample size changes. The choice $g = 1/n$ corresponds closely to the *unit information prior* proposed by Kass and Wasserman (1995): the information content of the prior is approximately equal to the information contained in one observation. Another possibility would be to also specify a hyperprior for g, as was done by Liang et al. (2008) among others. We have chosen to

avoid this extra layer of complexity here.

The assumption that all countries have the same parameters is consistent with the Solow model, as discussed in the introduction, but not with new growth theory and some versions of neoclassical growth theory. According to new growth theory, countries end up in different growth regimes when they are subject to different initial conditions. These different regimes create nonlinearities in the growth data, as documented by Durlauf and Johnson (1995), Liu and Stengos (1999), Kalaitzidakis et al. (2000), Minier (2007) and others. A prior specification that puts 100% probability on the parameters being equal across countries seems unreasonably dogmatic as it completely rules out such nonlinearities a priori. Here we present one way of allowing for heterogeneity in the parameters. To facilitate comparison we start out by adopting (4.6) as the marginal prior distribution for the country specific parameters $\beta_i = (\beta_{i,1}, ..., \beta_{i,p_s})'$ as in the earlier studies. However, instead of making the classical assumption that the parameters are equal across countries, we follow new growth theory in allowing them to vary according to initial conditions. This is the intuition behind the work of Durlauf and Johnson (1995) and Minier (2007), who split the sample according to initial output and human capital and find evidence of parameter heterogeneity between the different groups. A problem with this approach is that it is not clear a priori at what level the sample should be split or how many of these splits should be made. This makes it hard to do derive statistical conclusions from such a procedure, as discussed by Durlauf and Johnson (1995). Instead, we formalize the intuition of new growth theory by introducing prior covariances between the country-specific parameter vectors

$$\text{Cov}(\beta_i, \beta_j | s, \sigma^2) = \rho_{i,j}\sigma^2(gX_s'X_s)^{-1}, \tag{4.7}$$

where β_i and β_j are the vectors of regressor coefficients for countries i and j, and $\rho_{i,j}$ is the prior correlation between the parameters for these two countries. In other words, we specify a joint multivariate normal prior for all β parameters:

$$\begin{aligned} \beta &= (\beta_1', \ldots, \beta_n')' \sim N(0, B_0), \\ \text{with } B_0 &= \rho \otimes \sigma^2(gX_s'X_s)^{-1}, \end{aligned} \tag{4.8}$$

where ρ is the $n \times n$ matrix of all cross-country correlations $\rho_{i,j}$. New growth theory

suggests that countries with similar initial conditions are likely to have similar parameters. We formalize this idea by making $\rho_{i,j}$ dependent on a distance measure $d_{i,j}$ that captures how dissimilar the initial conditions of countries i and j are:

$$\rho_{i,j} = \exp(-\gamma d_{i,j}^2), \tag{4.9}$$

where γ is an unknown nonnegative parameter that determines the degree of nonlinearity in the model. This exponential specification makes sure that all cross-country correlations $\rho_{i,j}$ lie between 0 and 1, with higher correlations for similar countries (low $d_{i,j}$) than for dissimilar countries (high $d_{i,j}$). For the remainder of this chapter we will assume that $d_{i,j}$ is the Euclidean distance between the vectors of initial conditions for country i and country j:

$$d_{i,j} = ||c_i - c_j||, \tag{4.10}$$

although other distance measures are certainly also possible. Rasmussen and Williams (2006) offers some guidance on choosing distance measures which will guarantee a positive-definite correlation matrix.

The vectors of initial conditions c_i and c_j in equation (4.10) should contain those variables that determine the growth regime of a country. Of these initial conditions, output and human capital are believed to be the most important (e.g. Durlauf and Johnson, 1995; Liu and Stengos, 1999; Minier, 2007). In Section 4.5 we therefore explore different specifications with c_i and c_j containing representations of these conditions.

It may seem strange that we do not also allow the intercept α to vary across countries. However, since the data described in Section 4.3.5 only contain a single data point for each country, the noise term ϵ_i can be interpreted as containing a country-dependent intercept:

$$\epsilon_i = a_i + \eta_i, \qquad a_i \sim NID(0, \xi), \qquad \eta_i \sim NID(0, \sigma^2 - \xi), \tag{4.11}$$

where a_i can be seen as the intercept and η_i as the error. We could also introduce a prior correlation between the country-dependent intercepts a_i to express the prior idea that countries with similar initial output and human capital will have similar levels of growth, but this would be redundant as we have already expressed this by letting these

variables enter the model linearly in equation (4.1).

The prior cross-country correlations $\rho_{i,j}$ are determined by the γ parameter through equation (4.9). This parameter is somewhat difficult to interpret directly, but it has a simple one-to-one relationship with the median of the correlations $\rho_{i,j}$:

$$\bar{\rho} = \text{median}(\rho_{i,j}) = \exp(-\gamma \bar{d}^2), \tag{4.12}$$

where $\bar{d}$ is the median of the distances $d_{i,j}$ in equation (4.9). If the median prior correlation $\bar{\rho}$ is set to one, then γ is equal to zero, all β_i parameters are assumed to be equal and our model reduces to the linear model specification of Fernandez et al. (2001b) and Sala-i-Martin et al. (2004). If on the other hand $\bar{\rho}$ approaches zero, γ grows infinite and the regression coefficients of the different countries approach complete independence. We would like to infer the right amount of dependence from the data and we therefore assume a uniform prior on $\bar{\rho}$, which implies an exponential prior on γ:

$$\bar{\rho} \sim \text{U}(0,1] \quad \Leftrightarrow \quad p(\gamma) = \bar{d}^2 \exp(-\bar{d}^2 \gamma), \text{for } \gamma \in [0, \infty). \tag{4.13}$$

Note that the support for $\bar{\rho}$ is $(0,1]$, since $\bar{\rho}$ is strictly positive (but arbitrarily small) for finite γ, but that it is exactly equal to one for $\gamma = 0$. Similarly, the support for γ is $[0, \infty)$, which is to be contrasted with some textbooks that exclude 0 from the support of the exponential distribution.

Finally, we finish the prior specification by adopting standard non-informative priors for the remaining parameters α and σ^2 as is often suggested in the literature on Bayesian model averaging (e.g. Fernandez et al., 2001a):

$$\begin{aligned} p(\sigma^2) &\propto 1/\sigma^2 \\ p(\alpha) &\propto 1. \end{aligned} \tag{4.14}$$

Conditional on s and γ, the posterior distributions and proportional marginal data likelihood can now be obtained analytically, details of which can be found in Appendix 4.A.1. The posterior distribution for σ^2 is inverted Gamma and that for β is multivariate Student's

t:

$$\begin{aligned} \sigma^2|y,s,\gamma &\sim \text{IG}(\nu/2,\tau/2), \quad \nu = n-1, \quad \tau = y'Dy \\ \beta|y,s,\gamma &\sim t_{np_s}(\nu,\mu,S), \quad \mu = CDy, \quad S = \frac{\tau}{\nu}(B_0 - CDC'), \end{aligned} \tag{4.15}$$

where n is the number of observations in y, and C is proportional to the prior covariance between β and y, characterized by $C_{i+pm,j} = \rho_{m+1,j}[(gX_s'X_s)^{-1}_{i,1}x^s_{j,1} + \cdots + (gX_s'X_s)^{-1}_{i,p}x^s_{j,p}]$ for integers $i \leq p, m < n, j \leq n$, with $x^s_{j,p}$ the element (j,p) of X_s. B_0 is the prior covariance of matrix β as defined in (4.8), and finally D is defined as

$$D = (K + \mathrm{I})^{-1} - \frac{(K+\mathrm{I})^{-1}\iota\iota'(K+\mathrm{I})^{-1}}{\iota'(K+\mathrm{I})^{-1}\iota}, \tag{4.16}$$

with ι an $n \times 1$ vector of ones, and K an $n \times n$ matrix defined in equation (4.19).

These posterior distributions are conditional on the variable selection s, and the coefficients of the variables that were not included have a posterior unit point mass at zero. The posterior distributions of σ^2 and β averaged over s and γ can be obtained by MCMC as described in Section 4.3.4.

The proportional marginal likelihood for given (s,γ) is given by

$$p(y|s,\gamma) \propto \frac{1}{\sqrt{\iota'(K+\mathrm{I})^{-1}\iota}} |K + \mathrm{I}\,|^{-1/2} (y'Dy)^{-(n-1)/2}. \tag{4.17}$$

This expression only gives us a proportionality as $p(y|s,\gamma)$ is not normalizable because of our improper priors on α and σ^2 in (4.14). Because these parameters are shared by all models indexed by s, expression (4.17) can nevertheless be used to derive weights for Bayesian model averaging and to construct Bayes factors comparing different values of γ.

4.3.3 Gaussian process priors

Thus far we have assumed a parametric model for economic growth and we have specified prior distributions for the parameters of that model. Another way of looking at this specification is by looking at the implied prior on the regression function itself. In equation

(4.1) we decomposed the growth rates into an explainable part and an error term:

$$y_i = f_i + \epsilon_i, \text{ with } f_i = \alpha + \beta_{i,1} x^s_{i,1} + \beta_{i,2} x^s_{i,2} + ... + \beta_{i,p_s} x^s_{i,p_s}, \tag{4.18}$$

where f_i can be interpreted as the latent regression function evaluated at x^s_i, the i-th row of the selected explanatory variables X_s. Stacking the f_i into an $n \times 1$ vector $f = (f_1, f_2, \ldots, f_n)'$ and integrating out the β parameters with respect to their prior, we find that our specification implies the following prior on the regression function:

$$\begin{aligned} f|s, \alpha, \gamma, \sigma^2 &\sim N(\alpha\iota, \sigma^2 K) \\ K_{i,j} &= K(x^s_i, x^s_j) = \rho_{i,j} x^s_i (g X'_s X_s)^{-1} x^{s\prime}{}_j, \end{aligned} \tag{4.19}$$

where ι is an $n \times 1$ vector of ones. Because we have specified a linear model combined with a Gaussian prior on its regression coefficients, the implied prior on the regression function is Gaussian as well. This prior is characterized by its mean α and its *covariance function* or *kernel* $K(x^s_i, x^s_j)$. Note that this covariance function can be evaluated at any two values of x, not just those that occurred in our finite sample. The covariance function thus encodes our prior distribution on the entire regression function, not just on its values for the countries in the sample. Such a distribution over functions is called a stochastic process, and a stochastic process of which any finite dimensional distribution is a (multivariate) Gaussian is called a Gaussian process. For this reason, priors of the type used here are often called *Gaussian process priors*.

The covariance function $K(x^s_i, x^s_j)$ captures our prior ideas about the smoothness properties of the regression function. By adjusting the covariance function we can allocate prior probability to many different kinds of nonlinearity, while inference remains analytically tractable due to the Gaussian nature of the prior. In fact, several well known methods of nonlinear regression such as smoothing splines and neural networks may be seen as special cases of Gaussian process priors (see e.g. Rasmussen and Williams, 2006; MacKay, 1998).

The most well known special case of a Gaussian process prior is of course the standard linear regression model with a normal prior on the regression coefficients. If the

prior covariance matrix of the regression coefficients in such a model is given by C, the covariance kernel on the regression function is the linear covariance kernel:

$$K_{\text{lin}}(x_i^s, x_j^s) = x_i^s C x^{s\prime}{}_j. \tag{4.20}$$

The linear or 'dot product' kernel has the defining property that it only allocates prior probability to linear regression functions, which means that the posterior of the regression function will always be linear. The most popular kernel for performing nonlinear regression using Gaussian process priors is the squared exponential covariance kernel

$$K_{\text{sq.exp.}}(x_i^s, x_j^s) = \exp(-\lambda ||x_i^s - x_j^s||^2), \tag{4.21}$$

where λ is a parameter that controls the smoothness of the regression function.

Many more useful types of covariance functions have been proposed in the literature. A review of the most popular ones can be found in Rasmussen and Williams (2006, Ch. 4). In practice, almost any function can be used as a covariance kernel, with the only restriction that the function has to be positive semi-definite, i.e. that it produces positive semi-definite covariance matrices. Interestingly, this means that the sum of two covariance kernels is always a valid covariance kernel, as is their product. In fact, the covariance function used in our analysis (4.19) is the product of a linear kernel $K_{\text{lin}}(x_i^s, x_j^s)$ with selected regressors x^s and a squared exponential kernel $K_{\text{sq.exp.}}(c_i, c_j)$ with initial conditions c. By adapting the covariance function (4.19), the approach presented here can easily be modified to allow for many different kinds of nonlinearity. Note that the marginal likelihood in equation (4.17) is already expressed in terms of the kernel matrix K, so it can easily be used to do Bayesian model averaging with general Gaussian process priors. A starting point in specifying alternative covariance functions may be the method of Koop and Poirier (2004), who explicitly discuss the specification of variants of commonly used classical semi-parametric methods in this form.

4.3.4 Posterior inference

Conditional on γ and the selection of variables s to include in the regression, the marginal data likelihood (4.17) and the parameter posterior distributions (4.15) can be calculated analytically. However, considering all possible subsets of explanatory variables, the full model space now contains $2^{67} \approx 1.5{*}10^{20}$ different models for the SDM data with $k = 67$, which makes it impossible to explicitly average over all candidate models. Fortunately, the work by Fernandez et al. (2001b) and Sala-i-Martin et al. (2004) suggests that the posterior probability mass is typically concentrated in a relatively small fraction of these models, making it feasible to simulate from the posterior distribution over models. To accomplish this, we use the MC^3 methodology of Madigan and York (1995), which is a Metropolis-Hastings algorithm on the model space. The MC^3 algorithm uses a uniform proposal distribution on the model space containing the current model and all models obtained by adding or removing a regressor. By using the stochastic Metropolis-Hastings acceptance criterion, the algorithm is ensured to have the posterior distribution over models as its stationary distribution. After each MC^3 step, we draw a new proposal value for γ from its prior in an independence chain Metropolis-Hastings step.

4.3.5 Data

We estimate the new nonlinear models on the data sets of Fernandez et al. (2001b) (FLS) and Sala-i-Martin et al. (2004) (SDM). Both data sets contain growth data on a cross section of countries, along with a number of explanatory variables. The FLS data set contains 72 countries and 41 explanatory variables, while the SDM data consist of 88 countries and 67 explanatory variables. All explanatory variables in these data sets were measured at the beginning of the sample period (1960) in order to avoid endogeneity problems, with the exception of the variables related to war, inflation and the openness of the economy. The economic growth rates for the countries in these data sets were computed over the period 1960-1992 (FLS) and 1960-1996 (SDM) respectively. The data sets contain countries in different stages of development and with a wide geographic dispersion. The explanatory variables cover a wide range of different factors, including data on economic development, social issues, health, geography, politics, education and

more. The FLS and SDM data sets contain a number of the same countries and do not represent independent data. However, by examining both data sets we can assess the sensitivity of our findings in Section 4.4 to the size of the data set and to data revisions (see Ciccone and Jarocinski, 2010). For further discussion of the data as well as a list of sources we refer the reader to Fernandez et al. (2001b) and Sala-i-Martin et al. (2004).

4.4 Model selection and parameter heterogeneity

In Section 4.3 we specified a nonlinear model capable of capturing the parameter heterogeneity caused by differences in the initial conditions for the countries in our sample. Of these initial conditions, output and human capital are believed to be the most important (e.g. Durlauf and Johnson, 1995; Liu and Stengos, 1999; Minier, 2007). In the SDM and FLS data sets initial output is captured by the log of the GDP per capita for each country. Both data sets contain 3 different variables that represent the human capital of a country: the primary schooling enrollment rate, the higher education enrollment rate and the public education expenditure as a fraction of GDP. We estimate several different specifications of the model from Section 4.3 that include different subsets of these variables in the vector of initial conditions c. Before inclusion, each of these variables is standardized to have unit variance to remove any effects of scaling, as is common practice in the application of Gaussian process priors (Rasmussen and Williams, 2006).

For each specification of the initial conditions, eleven million draws were generated from the posterior distribution over models and over γ, of which the first million were discarded as burn-in. The remaining draws contained around half a million unique models, depending on the specification of the initial conditions. This number was roughly the same for the SDM and FLS data: although the SDM data contain more explanatory variables they also contain more observations which leads to a higher concentration of posterior model probability. By numerically integrating out γ from the likelihood for these specifications we can compare the exact proportional posterior probabilities of these models to the number of times they were sampled (see Ley and Steel, 2009). The correlation between these measures is extremely close to one, indicating convergence of the sampling procedure. In addition, the trace plots for γ, $\bar{\rho}$, the model size, and the different

variable inclusion indicators support this conclusion.

For each specification of the initial conditions c, we perform a formal test of the nonlinear model specification (that is $\gamma > 0$) against the linear model ($\gamma = 0$) by constructing a Bayes Factor:

$$BF = \frac{p(y|\text{nonlinear model})}{p(y|\text{linear model})} = \frac{p(y)}{p(y|\gamma = 0)} = \frac{p(\gamma = 0)}{p(\gamma = 0|y)}, \tag{4.22}$$

where the last ratio is known as the Savage-Dickey density ratio (Dickey, 1971; Verdinelli and Wasserman, 1995). The numerator in this ratio is known a priori from (4.13) and is equal to $\bar{d}^2$. The posterior density in the denominator $p(\gamma = 0|y)$ has to be estimated by Monte Carlo methods. Since there is a one-to-one relationship between γ and $\bar{\rho}$, the Savage-Dickey density ratio can equivalently be stated as $p(\bar{\rho} = 1)/p(\bar{\rho} = 1|y) = 1/p(\bar{\rho} = 1|y)$, the inverse of the posterior density at $\bar{\rho} = 1$.

Calculation of the Bayes Factor above thus comes down to evaluating the posterior density $p(\gamma = 0|y)$, or $p(\bar{\rho} = 1|y)$. These densities can be calculated accurately using the method of Chib and Jeliazkov (2001), which uses an additional MCMC run with γ fixed to 0. By using this method, we can avoid the difficulties that occur when using kernel density estimation to determine a probability density on the boundary of its support. Details of our use of this method can be found in Appendix 4.A.2. We use this method to compute Bayes factors for each nonlinear specification against the linear model, for both the SDM and FLS data sets. The results are presented in Table 4.1. Note that all Bayes factors are computed against the same linear model, which means that the results can also be used to compare the different nonlinear specifications against each other: if model 1 has a Bayes factor of 2.4 against the linear model and model 2 has a Bayes factor of 1.2, the Bayes factor of model 1 against model 2 is equal to $2.4/1.2 = 2$. The Bayes factors in Table 4.1 can thus also be interpreted as the relative evidence in favor of each different specification of the initial conditions.

Initial conditions	Relative evidence on SDM data	Relative evidence on FLS data
GDP	43	4.5
primary schooling	24	52
GDP & primary schooling	44	34
GDP & higher education	42	1.8
GDP & education expenditure	1.1	0.70

Table 4.1: Bayes Factors nonlinear models vs linear model

Although there are some differences in the results for the two data sets, we can conclude that the model with the most support from the data is the nonlinear model with GDP and primary schooling as initial conditions. The nonlinear models with only GDP, only primary schooling, and GDP & higher education as initial conditions also do better than the linear model. The nonlinear model with GDP & education expenditure receives about the same support from the data as the linear model specification. A reason for this result may be that education expenditure is not a good predictor of education outcomes.

Overall, the Bayes factors in Table 4.1 provide evidence in support of parameter heterogeneity. This result is consistent with the conclusions of Durlauf and Johnson (1995) and Liu and Stengos (1999) who have documented this type of nonlinearity before.

One might suspect that the nonlinear models have an unfair advantage in this comparison since they always include an (indirect) influence of the variables selected for the initial conditions, while the linear models do not. However, restricting the linear models to always include these variables does not substantially increase their posterior probability. In addition, we find that the posterior distribution for γ is quite insensitive to changes in the prior distribution over models. The evidence in favor of nonlinearity presented in Table 4.1 is robust to many different settings of g and of the hyperprior on θ. We do find that the magnitude of the Bayes factors diminishes somewhat for very small values of g. For example, choosing $g = 1/k^2$ (as per Fernandez et al., 2001b) with GDP and primary schooling as initial conditions gives us a Bayes factor of 3 on the SDM data, rather than the 44 reported in Table 4.1. However, by constructing Bayes factors comparing different values of g, we find that such small values of g receive very little support from the data.

4.5 Posterior results

Since the nonlinear model with GDP and primary education as initial conditions is most supported by the data, we now characterize the posterior distribution under this specification. To avoid redundancy we only consider the SDM data. The corresponding posterior distributions for the median cross-country correlation $\bar{\rho}$ and for the model size are displayed in Figures 4.1 and 4.2. A summary of the posterior of the regression coefficients is given in Table 4.2. The posterior distribution of some regression coefficients is discussed more deeply in Section 4.5.2.

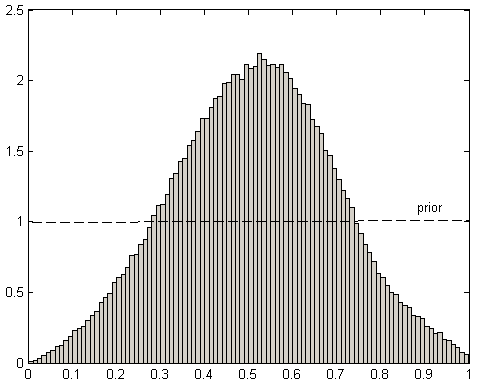


Figure 4.1: Posterior distribution $\bar{\rho}$

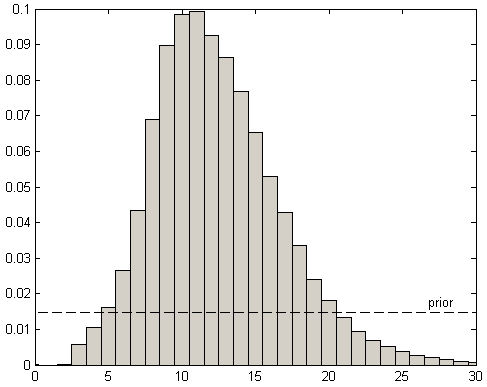


Figure 4.2: Posterior distribution model size

The posterior distribution of the median correlation $\bar{\rho}$, displayed in Figure 4.1, shows that the data support a median cross-country correlation of around 0.5 for the SDM data. The posterior density is low for both very small values of $\bar{\rho}$ as well as very large values, indicating that the regression coefficients are most likely not independent, but also not equal across countries.

The posterior correlation between $\bar{\rho}$ and the model size is -0.12, which is due to the fact that those draws with $\bar{\rho}$ very close to one have a somewhat smaller model size than the other draws. Using the methodology of George and McCulloch (1997) we find that 10 million draws is enough to cover about 60% of the posterior model probability. More importantly, the posterior statistics remain stable as we increase the number of draws, which indicates that the Monte Carlo sample is large enough.

4.5.1 Posterior summary of regression coefficients

The posterior distribution of the regression coefficients β is a mixture of multivariate Student's t densities, depending on the sampled γ and variable selections, and can be obtained analytically from equation (4.15). The characteristics of this posterior distribution are summarized in Table 4.2. The first column of the table lists the posterior inclusion probability of each variable, i.e. the sum of the posterior probabilities of all models including that variable. Conditional on the inclusion of each regressor, the table lists the posterior expectation and posterior standard deviation of the average parameter for that regressor, computed by averaging over all countries. This gives us a sense of the average directional effect of each explanatory variable and facilitates comparison with the earlier studies. Using the terminology of the preceding literature, a regressor can be considered robust if the bulk of the posterior mass of its average parameter lies either above or below zero. In order to determine this, the posterior confidence for the sign of the average parameter of each regressor is reported, which Sala-i-Martin et al. (2004) called the 'sign certainty probability'. Also reported is the expected fraction of countries with a parameter of this sign, which can be computed by summing the probabilities of each country having a parameter of the reported sign, and dividing by the number of countries. The individual probabilities in this sum can be obtained from the marginals of the posterior distribution given in (4.15). Finally, the table contains the number of countries that have at least a 90% posterior probability of having a parameter of this sign, which is obtained similarly.

The reported regressors are ordered according to posterior inclusion probability. The variables that were found to be 'significant' by Sala-i-Martin et al. (2004) are printed in bold face to facilitate comparison. They consider a variable significant if its posterior inclusion probability is above $7/k$, the prior inclusion probability in their specification. In their results, most of these variables also have a sign certainty probability above 0.975 and are thus considered 'robust'. Our prior inclusion probability is random, but a similar distinction can be made between the variables with above average posterior inclusion probability and those with below average posterior inclusion probability. In Table 4.2 these two groups of variables are separated by a horizontal line. All variable names used match those in Sala-i-Martin et al. (2004). Further description of these variables as well

as their source can be found in Table 1 of their paper.

Table 4.2: Parameter estimates regressors

Variable	Posterior inclusion probability	Posterior mean avg. parameter conditional on inclusion	Posterior stnd. dev. average parameter	Sign certainty probability	$\mathbb{E}$ fraction of countries with parameter of this sign	Number of countries with $>90\%$ prob of having a param. of this sign
Investment Price	0.911	-9.11e-5	4.90e-5	0.975	0.774	34
East Asian Dummy	0.814	0.0120	0.0078	0.943	0.757	16
Fertility in 1960s	0.749	-0.0060	0.0146	0.635	0.565	10
Life Expectancy in 1960	0.649	0.0007	0.0004	0.962	0.755	30
Political Rights	0.644	-0.0009	0.0014	0.739	0.564	11
Fraction of Tropical Area	0.570	-0.0093	0.0059	0.942	0.714	14
GDP in 1960 (log)	0.545	-0.0076	0.0052	0.930	0.701	26
Malaria Prevalence in 1960s	0.449	-0.0046	0.0079	0.712	0.609	8
Fraction Population Less than 15	0.412	0.1154	0.0651	0.964	0.708	20
Higher Education 1960	0.333	-0.1333	0.0606	0.991	0.768	33
Openness measure 1965-74	0.333	-0.0003	0.0055	0.522	0.493	5
Public Investment Share	0.332	-0.0003	0.0477	0.510	0.514	17
Primary Schooling in 1960	0.257	0.0341	0.0126	0.998	0.774	41
Fraction GDP in Mining	0.237	0.0322	0.0291	0.865	0.692	6
Terms of Trade Growth in 1960s	0.221	0.0330	0.0501	0.749	0.597	4
Spanish Colony	0.220	-0.0060	0.0068	0.815	0.662	0
Ethnolinguistic Fractionalization	0.212	-0.0067	0.0064	0.857	0.648	3
Fraction Confucian	0.208	-0.0034	0.0348	0.533	0.500	0
Socialist Dummy	0.197	-10.35e-5	0.0076	0.496	0.464	0
Air Distance to Big Cities	0.197	7.28e-8	9.11e-7	0.538	0.474	0
Latin American Dummy	0.194	-0.0051	0.0087	0.751	0.635	0
Fraction Spent in War 1960-90	0.177	-0.0056	0.0098	0.720	0.538	12
Fraction Population Over 65	0.161	-0.1522	0.1543	0.847	0.567	20
Years Open 1950-94	0.154	0.0028	0.0078	0.634	0.554	0
British Colony Dummy	0.147	0.0024	0.0038	0.731	0.605	4
Population Density Coastal in 1960s	0.142	2.19e-7	4.63e-6	0.521	0.533	7
African Dummy	0.141	-0.0088	0.0083	0.854	0.685	13
Fraction Muslim	0.140	0.0057	0.0071	0.794	0.628	7
Population Density 1960	0.131	8.54e-6	1.18e-5	0.768	0.637	5

Table 4.2: Parameter estimates continued

Variable	Posterior inclusion probability	Posterior mean avg. parameter conditional on inclusion	Posterior stnd. dev. average parameter	Sign certainty probability	$\mathbb{E}$ fraction of countries with parameter of this sign	Number of countries with $>$ 90% prob of having a param. of this sign
Real Exchange Rate Distortions	0.128	-6.96e-5	5.18e-5	0.839	0.659	4
Fraction Buddhist	0.125	0.0069	0.0132	0.710	0.593	18
Fraction Speaking Foreign Language	0.125	0.0058	0.0051	0.879	0.690	6
Fraction Protestants	0.116	-0.0159	0.0120	0.926	0.725	23
European Dummy	0.110	-0.0169	0.0152	0.883	0.659	8
Public Education Spending Share in GDP in 1960s	0.107	0.2405	0.2102	0.872	0.599	22
Civil Liberties	0.104	-0.0108	0.0082	0.912	0.689	14
Fraction Catholic	0.102	-0.0082	0.0090	0.816	0.647	9
Land Area	0.099	-7.96e-10	1.18e-9	0.691	0.549	0
Gov. Consumption Share 1960s	0.089	-0.0035	0.0472	0.554	0.521	0
Fraction Population In Tropics	0.085	-0.0045	0.0081	0.706	0.589	0
Nominal Government GDP Share 1960s	0.084	-0.0453	0.0327	0.920	0.714	9
Fraction of Land Area Near Navigable Water	0.075	-0.0049	0.0067	0.773	0.612	0
Tropical Climate Zone	0.073	-0.0038	0.0080	0.684	0.582	0
Government Share of GDP in 1960s	0.073	-0.0012	0.0494	0.511	0.508	0
Absolute Latitude	0.073	-4.68e-5	0.0003	0.534	0.502	0
Interior Density	0.068	-3.82e-5	2.52e-5	0.807	0.638	21
Capitalism	0.062	-3.56e-5	0.0014	0.519	0.514	0
Size of Economy	0.062	7.67e-5	0.0016	0.521	0.513	0
Oil Producing Country Dummy	0.060	-0.0012	0.0093	0.558	0.545	0
Population Growth Rate 1960-90	0.059	0.1366	0.3676	0.651	0.553	0
War Particpation 1960-90	0.055	0.0010	0.0029	0.635	0.554	0
Terms of Trade Ranking	0.055	-0.0117	0.0153	0.764	0.625	0
Timing of Independence	0.052	-0.0003	0.0020	0.553	0.510	0
Population in 1960	0.051	-3.8e-9	4.47e-8	0.522	0.496	0
Primary Exports 1970	0.049	-0.0037	0.0086	0.669	0.596	0
Fraction Othodox	0.049	-0.0046	0.0200	0.600	0.545	0
Average Inflation 1960-90	0.045	6.73e-5	0.0001	0.690	0.584	0
Fraction Hindus	0.041	-0.0046	0.0184	0.592	0.525	0

Table 4.2: Parameter estimates continued

Variable	Posterior inclusion probability	Posterior mean avg. parameter conditional on inclusion	Posterior stnd. dev. average parameter	Sign certainty probability	$\mathbb{E}$ fraction of countries with parameter of this sign	Number of countries with $>$ 90% prob of having a param. of this sign
Square of Inflation 1960-90	0.041	2.17e-7	2.10e-6	0.597	0.534	0
English Speaking Population	0.039	-0.0075	0.0126	0.729	0.601	0
Hydrocarbon Deposits in 1993	0.039	0.0003	0.0004	0.784	0.622	9
Religion Measure	0.039	0.0007	0.0079	0.543	0.523	0
Defense Spending Share	0.037	0.0022	0.0927	0.490	0.488	0
Revolutions and Coups	0.035	-0.0054	0.0071	0.781	0.632	0
Colony Dummy	0.032	0.0011	0.0065	0.558	0.531	0
Outward Orientation	0.032	9.34e-5	0.0028	0.515	0.504	0
Landlocked Country Dummy	0.022	-0.0071	0.0050	0.931	0.685	19

The results of our model averaging confirm the general conclusion of Fernandez et al. (2001b) and Sala-i-Martin et al. (2004), indicating the importance of a number of the same variables. Most of these variables also have average parameter estimates of the same sign as those in the earlier studies. However, we also find some important differences. The most striking difference is that in the earlier studies there was a very strong positive correlation between inclusion probability and sign certainty. For example, in the analysis of Sala-i-Martin et al. (2004) the twenty variables with the highest inclusion probability are also the twenty variables with the highest sign certainty. In our analysis this is very different: a number of variables have a high inclusion probability despite having low sign certainty and several variables with high sign certainty have a low posterior inclusion probability. On average, the sign certainty in our analysis is decreased compared to the results of SDM and FLS. This difference is a direct result of allowing the coefficients of the variables to differ over countries. In the linear model specification only those variables that have a similar effect across countries are likely to be included, while in our specification also variables with heterogeneous effects can have predictive power.

In addition to a reduced sign certainty for the average parameters, the two last columns of the table show that the parameter estimates for each country individually are much less

certain than under the linear model specification and that the expected signs of the parameters may differ strongly over countries. This means that many of the variables found to be robust in earlier studies do not have robust marginal effects across countries and variable selections under our model specification. Although this is to be expected when allowing the variables to be country-dependent, it contradicts the findings of Minier (2007) who finds that allowing for nonlinearity makes the parameter estimates more robust. However, as already discussed in Section 4.2, those results are based on only the linear components of the model and not on the full marginal relationships which makes comparison difficult.

The variable with the highest posterior inclusion probability in our analysis is the *investment price*. The importance of this variable was also apparent in the earlier studies of Fernandez et al. (2001b) and Sala-i-Martin et al. (2004). This variable is also one of the few with a high sign certainty: a high price for investment goods depresses economic growth. The dummy for *East Asian countries* also has a high posterior inclusion probability and sign certainty. The influence of this dummy variable reflects the high rate of economic growth in this region during the sample period. This dummy variable has a high correlation with the *fraction of the population that is Confucian*, which also belongs to the significant variables of Fernandez et al. (2001b) and Sala-i-Martin et al. (2004). In their linear specifications this variable has a positive parameter with high sign certainty, while the posterior mean of its average parameter is negative in our analysis (with low sign certainty). This can be explained by the extremely high posterior inclusion probability of the East Asian dummy in our analysis: the majority of models with the Confucian variable also include the East Asian variable. In the analysis of Sala-i-Martin et al. (2004) the inclusion probability for the latter is a little lower (0.82) and it is not part of the explanatory variables considered by Fernandez et al. (2001b). When we look at the subset of our models that exclude the East Asian dummy, we recover the positive posterior mean found in these earlier studies. When both variables are included, the Confucian variable seems to enter with a negative sign. These two variables are good examples of the importance of geographic location to economic growth, as are the variables for the *fraction of tropical area*, the *Spanish Colony dummy*, the *air distance to big cities* and the *Latin American dummy*. Also consistent with the earlier studies are the findings that education (*primary schooling*, *higher education*), investment (*investment price*, *public investment share*) and

health (*life expectancy*, *Malaria prevalence*) are important to economic growth.

4.5.2 Parameter posterior means

Since the parameters in our specification are allowed to vary with GDP and primary schooling, the posterior means of the parameters are functions of these initial conditions. These posterior mean functions are summarized in Table 4.2, but this summary does not fully describe the heterogeneity present in the posterior mean functions of the parameters. To offer a more detailed look at this heterogeneity, this section shows and discusses the full posterior mean functions of selected regression coefficients, conditional on inclusion of the regressor. The variance around these posterior mean functions is reasonably high in most cases, so the comments below are mostly illustrative.

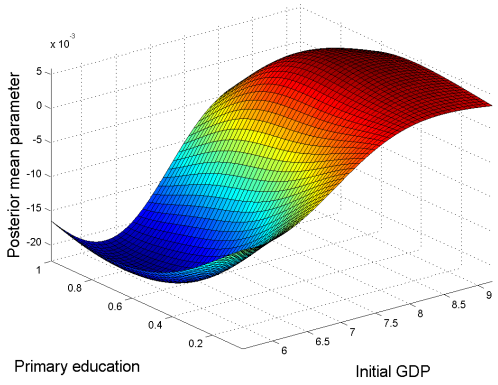


Figure 4.3: GDP in 1960

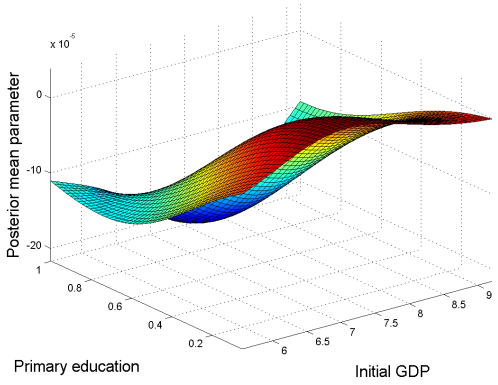


Figure 4.4: Investment Price

The first posterior mean function shown is that for the *initial log GDP per capita*. Figure 4.3 shows that the marginal effect of initial GDP on growth is largely negative, but that as initial GDP rises the marginal effect rises with it. This shows that the nonlinearity in this variable, found in several earlier studies, does not come only from its interaction with the other variables but also through its direct relationship with economic growth. This is consistent with the results of Liu and Stengos (1999) who find that initial GDP has an additive nonlinear effect on growth. The negative sign of the posterior mean of this coefficient provides evidence for the conditional convergence hypothesis. All else being equal, poor countries grow more rapidly than rich countries, and the marginal effect of initial GDP grows larger as initial income moves down. This suggests that very

poor countries catch up rapidly, ceteris paribus, and that this effect slows down as they get richer. However, as Quah (1993) points out, this does not necessarily imply that the income distribution over countries will eventually collapse into a single point. The negative effect of initial GDP on growth may also be explained by a model where the cross sectional income distribution is fixed, but the individual countries experience reversion to the mean around their long term growth rates.

The *price of investment goods* is the variable with the highest posterior inclusion probability in our analysis, with a negative posterior mean coefficient for all countries and a high sign certainty for the average effect over countries. Despite the uniformity in the sign of the posterior mean, Figure 4.4 still suggests strong parameter heterogeneity for this variable. In particular, the marginal benefit of a decrease in investment price seems to be much greater if a country has good primary schooling. This finding is consistent with several earlier studies on the effect of foreign direct investment on growth (Borensztein et al., 1998; Bengoa and Sanchez-Robles, 2003) that find that foreign direct investment only boosts a country's long term growth if a sufficient level of human capital is present. Apparently, this human capital is needed to fully take advantage of an increase in physical capital and technology.

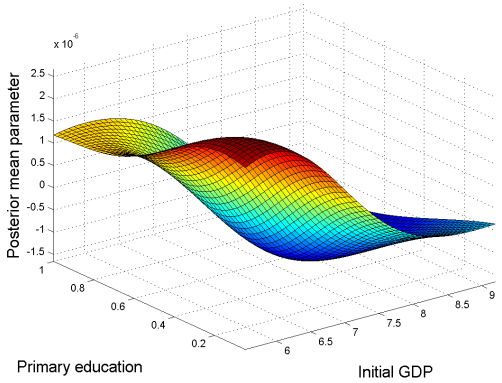


Figure 4.5: Air distance to big cities

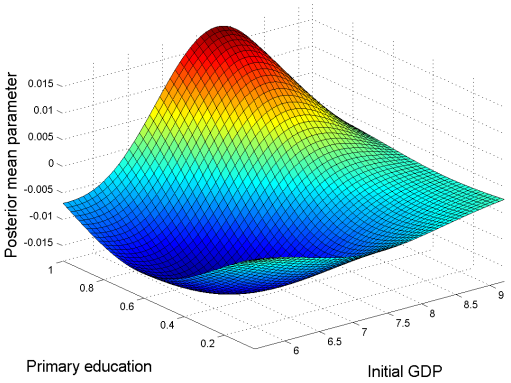


Figure 4.6: Openness measure 1965-74

The *air distance to big cities* variable measures the minimal log distance of a country to either New York, Rotterdam or Tokyo. The idea behind the construction of this variable is that a smaller distance to these cities is a proxy for better access to the American, European and Japanese markets. Earlier studies (e.g. Moreno and Trehan, 1997) find that

being close to a big market promotes economic growth through trade and technological spillovers. Figure 4.5 suggests that this effect is only beneficial for the richer half of the countries in our sample.

The same pattern can be seen for the *Openness measure 1965-74*, which represents the ratio of exports plus imports to GDP, averaged over 1965 to 1974. The posterior mean coefficient of this variable is increasing in both initial GDP and primary schooling, as can be seen in Figure 4.6. These findings are consistent with those of Yanikkaya (2003) who finds that having a closed economy with trade barriers might well be better for developing countries. They are also consistent with theoretical work that indicates that economic integration may be detrimental for certain individual countries, even if it is beneficial on average (see Grossman and Helpman, 1991; Rivera-Batiz and Xie, 1993). However, we cannot discount the possibility that this variable is also correlated with some unobserved aspect of a country's institutions. For this reason we should be careful in interpreting the estimated relationship as causal.

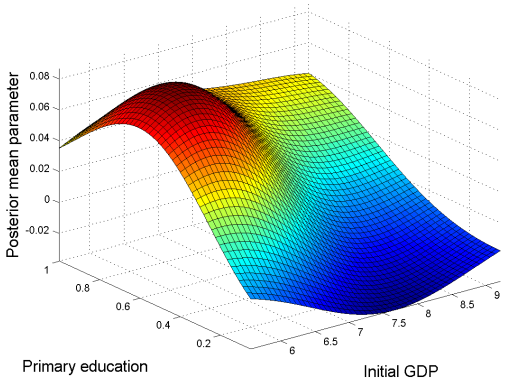


Figure 4.7: Fraction GDP in mining

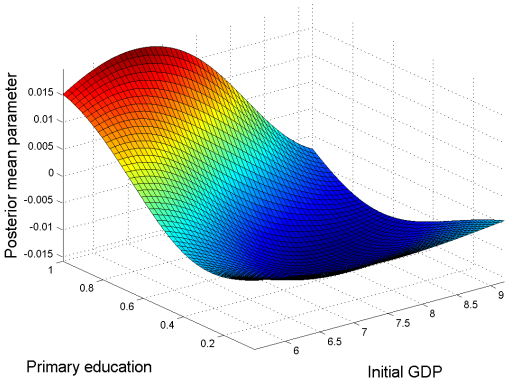


Figure 4.8: Oil producing country dummy

The *fraction of GDP in mining* is an important variable with high posterior inclusion probability in our analysis, as well as in the analysis of Sala-i-Martin et al. (2004). The SDM data set contains a few countries with high mining activity and high economic growth. The most notable example is Botswana, a country in Southern Africa that derives over half of its GDP from the mining of diamonds. Another example is Chile, one of the richest countries in Latin America, that also derives a large part of its income from mining. However, many other countries with high mining activity have experienced very

low growth. This phenomenon is known as the 'natural resources curse' (see Sachs and Warner, 2001) and can be explained by the rent-seeking behavior of countries with large endowments of natural resources. Figure 4.7 suggests that this contradiction may be explained by parameter heterogeneity. While the posterior mean of this parameter is positive for some countries, it is negative for others. Countries with high primary schooling enrollment rates seem to be less affected by the natural resources curse than countries with low enrollment rates. This result seems plausible as poor schooling is often cited as one of the mediating factors of the natural resources curse (e.g. Papyrakis and Gerlagh, 2004; Wood and Berge, 1997). The same pattern of parameter heterogeneity can be seen in Figure 4.8 for the *oil producing country dummy*, where the countries with high primary schooling enrollment rates also grew faster than those with low enrollment rates. It is important to note that this pattern is not apparent if we use higher education in the initial conditions instead of primary schooling, however this may be because there are simply no countries in the sample with high mining activity or oil production and high enrollment rates for higher education. The heterogeneity in the effect of natural resources is a striking example of nonlinearity in economic growth and deserves further research.

4.6 Conclusion

A rigorous regression analysis of cross-country economic growth data should jointly take into account the model uncertainty present in the variable selection problem as well in the functional form specification. Although both sources of model uncertainty have separately received much attention in the literature, joint treatments are unfortunately very rare. This chapter presents such an integrated analysis.

We address the model uncertainty relating to the functional form of the regression function by introducing a new flexible growth regression model based on a Gaussian process prior. The new model explicitly allows for parameter heterogeneity as suggested by new growth theory, while nesting the linear model specification as a special case. We solve the variable selection problem by performing Bayesian model averaging with the new model using different sets of explanatory variables. As argued earlier by Fernandez et al. (2001b) and Sala-i-Martin et al. (2004), this approach provides a theoretically sound

and practical way of considering a large class of different variable selections.

A formal model comparison provides evidence supporting the existence of parameter heterogeneity, consistent with the conclusions of Durlauf and Johnson (1995) and Liu and Stengos (1999) who have documented this type of nonlinearity before. The results do not support the conclusions of Minier (2007) who finds that allowing for nonlinearity makes the parameter estimates more robust to the variable selection. In addition, our results show that many of the explanatory variables do not have robust partial correlations to growth across countries.

The proposed method enables us to perform Bayesian model averaging over general Gaussian process priors, and can easily be adapted to allow for many different kinds of nonlinearity in the functional form of the regression function. By changing the covariance kernel used in the analysis, we can allocate prior probability to different forms of nonlinearity. Rasmussen and Williams (2006) provide an overview of the most commonly used covariance functions in the Gaussian process literature. Another starting point is the work of Koop and Poirier (2004), who explain how to specify variants of commonly used classical semi-parametric methods in a form compatible with the framework presented here. Exploring these different specifications is a promising direction for future research. In particular, it would be worthwhile to compare different types of nonlinear covariance functions to see which particular new growth theories are supported by the data and which are not.

Another possible extension of this work would be to allow for heteroskedasticity in the regression models. The assumption of homoskedasticity was made here to allow comparison with the earlier studies (Fernandez et al., 2001b; Sala-i-Martin et al., 2004; Minier, 2007) and because there is no strong evidence that the considered data are heteroskedastic. However, relaxing this assumption would be a logical next step. A possible starting point for an extension in this direction is the work of Giordani et al. (2009) who present a regression model allowing for both nonlinearity and heteroskedasticity.

4.A Appendix

4.A.1 Marginal likelihood and posterior distributions

The specification presented in section 4.3 is similar to the standard conjugate Bayesian specification for linear regression models, but with some subtle differences. For completeness we therefore present the derivations of the marginal likelihood and posterior distributions for this specification here. We derive all results in terms of the kernel matrix K defined in section 4.3.3 so that these results can easily be used to do Bayesian model averaging with general Gaussian process priors, simply by changing the covariance function of the prior.

Our starting point is the (improper) full joint distribution of the data and all model parameters, conditional on the distance parameter γ and the variable selection s:

$$p(y, \alpha, \beta, \sigma^2|s, \gamma) = p(y|\alpha, \beta, \sigma^2, s)p(\alpha)p(\beta|\sigma^2, s, \gamma)p(\sigma^2). \tag{4.23}$$

Using the decomposition on the right-hand side, we can easily integrate out β from this expression in the same way we integrated out these parameters in section 4.3.3. Doing so yields

$$\begin{aligned} p(y, \alpha, \sigma^2|s, \gamma) &= p(y|\alpha, \sigma^2, s, \gamma)p(\alpha)p(\sigma^2) \\ &\propto |\Sigma|^{-1/2} \exp[-0.5(y - \alpha\iota)'\Sigma^{-1}(y - \alpha\iota)] \times 1 \times \frac{1}{\sigma^2}, \end{aligned} \tag{4.24}$$

where ι is an $n \times 1$ vector of ones and $\Sigma = \sigma^2(K + \mathrm{I})$ is the covariance matrix of y, with the kernel matrix K as defined in equation (4.19).

The next step is to decompose the above expression into a univariate normal density function in α, multiplied by a function that does not depend on α:

$$\begin{aligned} p(y, \alpha, \sigma^2|s, \gamma) &\propto \sqrt{\iota'\Sigma^{-1}\iota} \exp\left(-\frac{1}{2}\iota'\Sigma^{-1}\iota\left[\alpha - \frac{\iota'\Sigma^{-1}y}{\iota'\Sigma^{-1}\iota}\right]^2\right) \\ &\quad \times \frac{1}{\sqrt{(\iota'\Sigma^{-1}\iota)|\Sigma|}} \exp\left(-\frac{1}{2}y'\left(\Sigma^{-1} - \frac{\Sigma^{-1}\iota\iota'\Sigma^{-1}}{\iota'\Sigma^{-1}\iota}\right)y\right) \times 1 \times \frac{1}{\sigma^2} \\ &\propto p(\alpha|y, \sigma^2, s, \gamma)p(y, \sigma^2|s, \gamma). \end{aligned} \tag{4.25}$$

The first term in this decomposition gives us the posterior distribution of α, conditional on σ^2:

$$p(\alpha|y,\sigma^2,s,\gamma) = N\left(\frac{\iota'\Sigma^{-1}y}{\iota'\Sigma^{-1}\iota}, \frac{1}{\iota'\Sigma^{-1}\iota}\right) = N\left(\frac{\iota'(K+\mathrm{I})^{-1}y}{\iota'(K+\mathrm{I})^{-1}\iota}, \frac{\sigma^2}{\iota'(K+\mathrm{I})^{-1}\iota}\right). \quad (4.26)$$

Since the first term in (4.25) is a proper density function, α can now easily be integrated out by simply dropping this term from the expression. Doing this, and now explicitly writing $\Sigma = \sigma^2(K+\mathrm{I})$, we are left with

$$\begin{aligned} p(y,\sigma^2|s,\gamma) \quad &\propto \quad \left(\frac{1}{\sigma^2}\right)^{1+(n-1)/2} \frac{1}{\sqrt{(\iota'(K+\mathrm{I})^{-1}\iota)|K+\mathrm{I}|}} \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}y'\left[(K+\mathrm{I})^{-1} - \frac{(K+\mathrm{I})^{-1}\iota\iota'(K+\mathrm{I})^{-1}}{\iota'(K+\mathrm{I})^{-1}\iota}\right]y\right) \end{aligned} \quad (4.27)$$

This expression can be recognized as the kernel of an inverse Gamma probability density function in σ^2, giving it the following posterior distribution:

$$\sigma^2|y,s,\gamma \sim \mathrm{IG}(\nu/2,\tau/2), \quad \nu = n-1, \quad \tau = y'(K+\mathrm{I})^{-1}y - \frac{(y'(K+\mathrm{I})^{-1}\iota)^2}{\iota'(K+\mathrm{I})^{-1}\iota}. \quad (4.28)$$

The normalizing constant of (4.27) gives us the proportional marginal likelihood:

$$p(y|s,\gamma) \propto \frac{\tau^{-\nu}}{\sqrt{(\iota'(K+\mathrm{I})^{-1}\iota)|K+\mathrm{I}|}}. \quad (4.29)$$

Note that these expressions only depend on the kernel matrix K. The analysis presented here can thus very easily be used to do Bayesian model averaging with general Gaussian process priors, simply by changing the covariance function.

For the purpose of deriving the posterior distribution for $\beta = (\beta_1', \ldots, \beta_n')'$, first define

$$Z_s = \begin{bmatrix} x_1^s & & 0 \\ & x_2^s & \\ & & \ddots \\ 0 & & x_n^s \end{bmatrix}, \quad (4.30)$$

with x_i^s the i-th row of X_s, and recall the normal prior distribution $p(\beta|\sigma^2,s,\gamma) = N(0,\sigma^2 B_0)$

specified in section 4.3. The full conditional distribution of β then follows from standard results (see Greenberg, 2008, pp. 45):

$$\begin{aligned} p(\beta|y,\alpha,\sigma^2,s,\gamma) &= N(\bar{\beta}_1,\sigma^2 B_1) \\ B_1 &= (Z_s'Z_s + B_0)^{-1} \\ \bar{\beta}_1 &= B_1 Z_s'(y-\alpha\iota) \end{aligned} \tag{4.31}$$

The conditional posterior mean of β depends linearly on α, which means that we can easily integrate α out with respect to its posterior distribution given in equation (4.26):

$$\begin{aligned} p(\beta|y,\sigma^2,s,\gamma) &= \int p(\beta|y,\alpha,\sigma^2,s,\gamma)p(\alpha|y,\sigma^2,s,\gamma)d\alpha \\ &= N(\bar{\beta}_2,\sigma^2 B_2) \\ \bar{\beta}_2 &= B_1 Z_s'\left(y - \frac{\iota'(K+\mathrm{I})^{-1}y}{\iota'(K+\mathrm{I})^{-1}\iota}\iota\right) \\ B_2 &= B_1 + \frac{B_1 Z_s'\iota\iota' Z_s B_1}{\iota'(K+\mathrm{I})^{-1}\iota} \end{aligned} \tag{4.32}$$

The expression above is hard to work with directly as it is defined in terms of B_1 which is the inverse of a matrix of dimension $np_s \times np_s$. However, we can simplify things by making use of the Woodbury matrix identity (Woodbury, 1950) to perform this inversion, and then realizing that $Z_s B_0 Z_s' = K$. This gives us

$$\bar{\beta}_2 = CDy, \quad B_2 = B_0 - CDC', \tag{4.33}$$

with

$$D = (K+\mathrm{I})^{-1} - \frac{(K+\mathrm{I})^{-1}\iota\iota'(K+\mathrm{I})^{-1}}{\iota'(K+\mathrm{I})^{-1}\iota}, \tag{4.34}$$

and where $\mathcal{C}$ is proportional to the prior covariance between β and y, characterized by $C_{i+pm,j} = \rho_{m+1,j}[(gX_s'X_s)^{-1}_{i,1}x^s_{j,1} + \cdots + (gX_s'X_s)^{-1}_{i,p}x^s_{j,p}]$ for integers $i \leq p, m < n, j \leq n$.

Finally, we can integrate out the variance parameter σ^2 from $p(\beta|y,\sigma^2,s,\gamma)$ with respect to its posterior distribution (4.28) to give a multivariate Student's t posterior distri-

bution for β:

$$\begin{aligned} p(\beta|y,s,\gamma) &= \int p(\beta|y,\sigma^2,s,\gamma)p(\sigma^2|y,s,\gamma)d\sigma^2 \\ &= t_{np_s}(\nu,\bar{\beta}_2,\frac{\tau}{\nu}B_2) \end{aligned} \tag{4.35}$$

4.A.2 Bayes factor calculation

In Section 4.4 we formally compare the proposed model with the linear model by means of a Bayes factor. Since the new model is equivalent to the linear model when we set its γ parameter equal to zero, this Bayes factor is given as follows:

$$BF = \frac{p(y)}{p(y|\gamma=0)} = \frac{p(\gamma=0)}{p(\gamma=0|y)}. \tag{4.36}$$

This equality is known as the Savage-Dickey density ratio (Dickey, 1971; Verdinelli and Wasserman, 1995). Here, the numerator is equal to $\bar{d}^2$, as given by equation (4.13). Calculation of the Bayes factor thus comes down to determining the posterior density $p(\gamma=0|y)$. Since 0 is on the boundary of the support of γ, and since the posterior density at this point is very low, this density cannot be estimated accurately using standard kernel density estimation. In order to do determine this density accurately, we use the methodology of Chib and Jeliazkov (2001). The central equality used in this method is

$$p(\gamma^*|y) = \frac{E_1\ \alpha(\gamma,\gamma^*|y,s)q(\gamma,\gamma^*)}{E_2\ \alpha(\gamma^*,\gamma|y,s)}, \tag{4.37}$$

where y denotes the data, s the selection of explanatory variables, $q(\gamma,\gamma^*)$ is the probability density of proposing a move from γ to γ^* in the Metropolis-Hastings algorithm, and $\alpha(\gamma,\gamma^*|y,s)$ is the conditional probability of accepting such a proposal. The first expectation E_1 is taken with respect to the posterior distribution $p(\gamma,s|y)$ and can be evaluated using the samples from the main MCMC run. The second expectation E_2 is taken with respect to the distribution $p(s|y,\gamma^*)q(\gamma^*,\gamma)$, which is evaluated by running an additional MCMC chain, keeping γ fixed to γ^*. In our case, this second chain is run conditional on $\gamma=0$, which eliminates any problems caused by having a low posterior density at this point.

An alternative and conceptually simpler method of obtaining the Bayes factor is to perform inference in a mixture containing the linear model and the nonlinear model, and to allow the Markov chain to mix between the two. (see for example Carlin and Chib, 1995; Green, 1995) Since the linear model is obtained for $\gamma = 0$, this is equivalent to using the following prior on γ:

$$p(\gamma) = w\delta(0) + (1-w)\bar{d}^2 \exp(-\bar{d}^2\gamma), \quad \text{for } \gamma \in [0, \infty), \tag{4.38}$$

where $\delta(0)$ is a unit point mass at $\gamma = 0$ and w and $1 - w$ are the mixture weights. The Bayes factor can then be determined by running MCMC as described in Section 4.3 and counting the number of times that the linear model ($\gamma = 0$) is selected (and correcting for the mixture weights). This method is less efficient than the method based on Chib and Jeliazkov (2001) and requires tuning: the w parameter in the mixture weights has to be set to a suitably high value to make sure that the linear models are visited often enough. In turn, the proposal density for γ then has to be tuned since sampling from the prior (as in Section 4.3) would lead to many rejected proposals for $\gamma = 0$. After the necessary tuning we were able to use this method to confirm that the results presented in Section 4.4 are correct.

Finally, we could compare the linear and nonlinear models by constructing partial Bayes factors (O'Hagan, 1995). Partial Bayes factors work by splitting the data in an estimation set and a test set, and using the posterior obtained from the estimation set as a prior to form a Bayes factor on the test set. This approach was used in an earlier version of this chapter and the results are in line with those for the full Bayes factor. Details can be obtained from the author upon request.

Chapter 5

Fixed-form variational posterior approximation through stochastic linear regression

Joined work with David Knowles

5.1 Introduction

In Bayesian analysis the posterior distribution is often of non-standard form. To obtain quantities of interest under such a distribution, such as moments or marginal distributions, we typically need to use Monte Carlo methods or approximate the posterior with a more convenient distribution. A popular method of obtaining such an approximation is *structured* or *fixed-form Variational Bayes*, which works by numerically minimizing the Kullback-Leibler divergence of an approximating distribution in the exponential family to the intractable target distribution (Attias, 2000; Beal and Ghahramani, 2006; Wainwright and Jordan, 2003). For certain problems, algorithms exist that can solve this optimization problem in much less time than it would take to approximate the posterior using Monte Carlo methods (see e.g. Honkela et al., 2010). However, these methods usually rely on analytic solutions to certain integrals and need conditional conjugacy in the model specification, i.e. the full conditionals of the posterior distribution must be standard exponential family distributions for these methods to be applicable. This makes this class of methods

limited in the types of approximations and posteriors they can handle.

We show that solving the optimization problem of fixed-form Variational Bayes is equivalent to performing a linear regression with the sufficient statistics of the approximation as explanatory variables and the (unnormalized) log posterior density as the dependent variable. Inspired by this result, we present an efficient stochastic approximation algorithm for solving this optimization problem. In contrast to earlier work, our approach does not require any analytic calculation of integrals, which allows us to extend the fixed-form Variational Bayes approach to problems where it was previously not applicable. Our method can be used to approximate any posterior distribution, provided that it is given in closed form up to the proportionality constant. The type of approximating distribution can be any distribution in the exponential family or any mixture of such distributions, which means that our approximations can in principle be made arbitrarily precise. While our method somewhat resembles performing stochastic gradient descent on the variational objective function in parameter space (Paisley et al., 2012; Nott et al., 2012), the linear regression view gives insights which allow a more computationally efficient approach.

Section 5.2 introduces fixed-form variational posterior approximation, the optimization problem to be solved, and the notation used in the remainder of the chapter. In Section 5.3 we provide a new way of looking at variational posterior approximation by re-interpreting its optimization problem as a linear regression problem. We propose a stochastic approximation algorithm to perform the optimization in Section 5.4. In Section 5.5 we discuss how to assess the quality of our posterior approximations and how to use the proposed methods to approximate the marginal likelihood of a model. These sections represent the core of the ideas behind this chapter.

To make our approach more general and computationally efficient we provide a number of extensions in two separate sections. Section 5.6 discusses modifications of our stochastic approximation algorithm to improve efficiency. Up to this point, all sections assume that our posterior approximation is in the exponential family. This is generalized to mixtures of exponential family distributions in Section 5.7. Section 5.8 gives some examples of using our method in practice. Here we show that despite its generality, the efficiency of our algorithm is highly competitive with more specialized approaches. Finally, Section 5.9 concludes.

5.2 Fixed-form Variational Bayes

Let x be a vector of unknown parameters and/or latent random effects for which we have specified a prior distribution $p(x)$, and let $p(y|x)$ be the likelihood of observing a given set of data y. Upon observing y, we can use Bayes' rule to obtain our updated state of belief, the posterior distribution:

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}. \tag{5.1}$$

An equivalent definition of the posterior distribution is

$$p(x|y) = \arg\min_{q(x)} \mathbb{E}_{q(x)} \left[\log \frac{q(x)}{p(x,y)} \right] = \arg\min_{q(x)} D[q(x)|p(x|y)], \tag{5.2}$$

where the optimization is over all proper probability distributions $q(x)$, and where $D[q(x)|p(x|y)]$ denotes the Kullback-Leibler divergence between $q(x)$ and $p(x|y)$. The KL-divergence is always non-negative and has a unique minimizing solution $q(x) = p(x|y)$ almost everywhere, at which point the divergence is zero. Note that the solution of (5.2) does not depend on the normalizing constant $p(y)$ of the posterior distribution, but that we do obtain it as a by-product of solving $D[q(x)|p(x|y)] = 0$.

The posterior distribution given in (5.1) is the exact solution of the *variational* optimization problem in (5.2), but except for certain special cases it is not very useful by itself because it is of non-standard form. This means that we do not have analytical expressions for the posterior moments of x, or for the marginals $p(x_i|y)$ of the multivariate posterior distribution, nor can we determine the normalizing constant $p(y)$. One method of solving this problem is to approximate these quantities using Monte Carlo simulation. A different approach, which we will pursue here, is to restrict the optimization problem in (5.2) to a reduced set of more convenient distributions Q. If $p(x,y)$ is of conjugate exponential form, choosing Q to be the set of factorized distributions $q(x) = q(x_1)q(x_2)\dots q(x_k)$ often leads to a tractable optimization problem that can be solved efficiently using an algorithm called Variational Bayes Expectation Maximization (VBEM, Beal and Ghahramani, 2002). Such a factorized solution is attractive because it makes the variational optimization problem easy to solve, but it is also very restrictive: it requires a conjugate

exponential model and prior specification and it assumes posterior independence between the different blocks of parameters x_i. This means that this factorized approach can be used with few models, and that the solution $q(x)$ may be a poor approximation to the exact posterior (see e.g. Turner et al., 2008).

A different approach to simplifying the variational optimization problem is to restrict the solution set Q to only include distributions of a certain parametric form $q_\eta(x)$, where η denotes the vector of parameters governing the shape of the posterior approximation. This approach is known as *structured* or *fixed-form* Variational Bayes (Honkela et al., 2010; Storkey, 2000; Saul and Jordan, 1996). Usually, the posterior approximation is chosen to be a specific member of the exponential family of distributions:

$$q_\eta(x) = \exp[T(x)\eta - U(\eta)]\nu(x), \tag{5.3}$$

where $T(x)$ is a $1 \times k$ vector of sufficient statistics, $U(\eta)$ takes care of normalization, and $\nu(x)$ is a base measure. The $k \times 1$ vector η is often called the set of *natural parameters* of the exponential family distribution $q_\eta(x)$. Using this approach, the variational optimization problem in (5.2) reduces to a parametric optimization problem in η:

$$\hat{\eta} = \arg\min_\eta \mathbb{E}_{q_\eta(x)}[\log q_\eta(x) - \log p(x, y)]. \tag{5.4}$$

If our posterior approximation is of a standard form, the $\mathbb{E}_{q(x)}[\log q(x)]$ term in (5.4) can often be evaluated analytically. If we can then also determine $\mathbb{E}_{q(x)}[\log p(x, y)]$ and its derivatives with respect to η, the optimization problem can be solved using gradient-based optimization or fixed-point algorithms. Posterior approximations of this type are often much more accurate than a factorized approximation, but the requirement that $q_\eta(x)$ is of standard form is still restrictive, as is the requirement of being able to evaluate $\mathbb{E}_{q(x)}[\log p(x, y)]$. In addition, existing optimization algorithms for fitting this type of approximation can be much slower than the EM type algorithms used for factorized approximation, reducing somewhat their advantage with respect to Monte Carlo methods. In the next section, we develop an algorithm that can efficiently solve the variational optimization problem for almost any type of approximating distribution $q_\eta(x)$ and exact posterior $p(x|y)$. The only requirements we impose on $\log p(x, y)$ is that it is given in

closed form. The main requirement on $q_\eta(x)$ is that we can sample from it. For simplicity, Sections 5.3 and 5.6 will also assume that $q_\eta(x)$ is in the exponential family. Section 5.7 will then show how we can extend this to include mixtures of exponential family distributions. By using these mixtures and choosing $q_\eta(x)$ to be of a rich enough type, we can in principle make our approximation arbitrarily precise.

5.3 Variational Bayes as linear regression

For notational convenience we will write our posterior approximation in the following adjusted form:

$$\tilde{q}_{\tilde{\eta}}(x) = \exp[\tilde{T}(x)\tilde{\eta}]\nu(x), \tag{5.5}$$

where we have removed the normalizer $U(\eta)$, and we have replaced it by adding a constant to the vector of sufficient statistics, i.e. $\tilde{T}(x) = (1, T(x))$ and $\tilde{\eta} = (\eta_0, \eta')'$. If η_0 is equal to $-U(\eta)$, equation (5.5) describes the same (normalized) distribution function as does equation (5.3). If η_0 is different from $U(\eta)$ it describes a rescaled (unnormalized) version of this distribution function.

To work with $\tilde{q}_{\tilde{\eta}}(x)$, we use the unnormalized version of the KL-divergence, which is given by

$$\begin{aligned} D[\tilde{q}_{\tilde{\eta}}(x)|p(x,y)] &= \int \tilde{q}_{\tilde{\eta}}(x) \log \frac{\tilde{q}_{\tilde{\eta}}(x)}{p(x,y)} d\nu(x) - \int \tilde{q}_{\tilde{\eta}}(x) d\nu(x) \\ &= \int \exp[\tilde{T}(x)\tilde{\eta}][\tilde{T}(x)\tilde{\eta} - \log p(x,y)] d\nu(x) - \int \exp[\tilde{T}(x)\tilde{\eta}] d\nu(x) \end{aligned} \tag{5.6}$$

At the minimum this gives $\eta_0 = \mathbb{E}_q[\log p(x,y) - \log q(x)]$ as shown in Appendix 5.A.1, which is the usual bound on the log evidence. The other parameters η have the same minimum as in the normalized case.

Taking the gradient of (5.6) with respect to the natural parameters $\tilde{\eta}$ we have

$$\nabla_{\tilde{\eta}} D[\tilde{q}_{\tilde{\eta}}(x)|p(x,y)] = \int \tilde{q}_{\tilde{\eta}}(x)[\tilde{T}(x)'\tilde{T}(x)\tilde{\eta} - \tilde{T}(x)' \log p(x,y)] d\nu(x). \tag{5.7}$$

Setting this expression to zero in order to find the minimum gives

$$\tilde{\eta} = \left[\int \tilde{q}_{\tilde{\eta}}(x)\tilde{T}(x)'\tilde{T}(x)d\nu(x)\right]^{-1}\left[\int \tilde{q}_{\tilde{\eta}}(x)\tilde{T}(x)'\log p(x,y)d\nu(x)\right], \tag{5.8}$$

or in its normalized form

$$\tilde{\eta} = \mathbb{E}_q[\tilde{T}(x)'\tilde{T}(x)]^{-1}\mathbb{E}_q[\tilde{T}(x)'\log p(x,y)]. \tag{5.9}$$

Note that we have implicitly assumed that the Fisher information matrix, $\mathbb{E}_q[\tilde{T}(x)'\tilde{T}(x)]$ is non-singular, which will be the case for any identifiable approximating exponential family distribution q. Our key insight is to notice the similarity between (5.9) with the maximum likelihood estimator for linear regression. Recall that in classical linear regression we have that the dependent variable $\{y_n \in \mathbb{R} : n = 1, .., N\}$ is distributed as $N(Y|X\beta, \sigma^2 I)$ where X is the $N \times D$ design matrix, β is the $D \times 1$ vector of regression coefficients and σ^2 is the noise variance. The maximum likelihood estimator for β is then

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{5.10}$$

To see the relation between (5.9) and (5.10), associate the design matrix X with the sufficient statistics $\tilde{T}$, the dependent variable Y with the unnormalized log posterior $\log p(x, y)$, and the regression coefficients β with the vector of natural parameters $\tilde{\eta}$. If we then consider Monte Carlo estimates of the expectations in (5.9) the analogy is very fitting indeed. A similar analogy is used by Richard and Zhang (2007) in the context of importance sampling. Appendix 5.A.3 discusses the connection between their work and ours.

Note that in equation (5.9), unlike equation (5.10), the right-hand side depends on the unknown parameters. This means that equation (5.9) in itself does not constitute a solution to our variational optimization problem. The next section introduces a stochastic approximation algorithm to perform this optimization.

5.4 A stochastic approximation algorithm

The link between variational Bayes and linear regression in itself is interesting, but it does not yet provide us with a solution to the variational optimization problem of equation (5.4). We propose solving this optimization problem using the stochastic approximation algorithm presented below. The basic idea is to draw a single sample from our posterior approximation $q(x)$ at a a time, and then to update this approximation using equation (5.9), while taking small enough steps to ensure convergence of the algorithm.

Algorithm 2 Stochastic Optimization for Fixed-Form Variational Bayes

Require: An unnormalized posterior distribution $p(x, y)$
Require: A type of approximating posterior $q_\eta(x)$
Require: The total number of iterations N
  Initialize η to a first guess, for example by matching the prior $p(x)$
  Initialize $C = \mathbb{E}_{q_\eta}[\tilde{T}(x)'\tilde{T}(x)]$, or a diagonal approximation of this matrix
  Initialize $g = C\eta$
  Initialize $\bar{C} = \mathbf{0}$
  Initialize $\bar{g} = \mathbf{0}$
  Step-size $w = 1/\sqrt{N}$
  for $t = 1 : N$ **do**
    Set $\eta = C^{-1}g$
    Simulate a draw x^* from the current approximation $q_\eta(x)$
    Set $\hat{g}_t = \tilde{T}(x^*)' \log p(x^*, y)$, or another unbiased estimate of $\mathbb{E}_{q_\eta}[\tilde{T}(x)' \log p(x, y)]$
    Set $\hat{C}_t = \tilde{T}(x^*)'\tilde{T}(x^*)$, or another unbiased estimate of $\mathbb{E}_{q_\eta}[\tilde{T}(x)'\tilde{T}(x)]$
    Set $g = (1 - w)g + w\hat{g}_t$
    Set $C = (1 - w)C + w\hat{C}_t$
    if $t > N/2$ **then**
      Set $\bar{g} = \bar{g} + \hat{g}_t$
      Set $\bar{C} = \bar{C} + \hat{C}_t$
    end if
  end for
  return $\hat{\eta} = \bar{C}^{-1}\bar{g}$

Algorithm 2 is inspired by a long line of research on stochastic approximation, starting with the seminal work of Robbins and Monro (1951). In fact, up to first order it can be considered a relatively standard stochastic gradient descent algorithm. At each iteration, we have $\eta_t = C_t^{-1}g_t$, where we use the subscript t to indicate the values of η, C and g during iteration t of Algorithm 2. We then update η_t to

$$\eta_{t+1} = [(1 - w)C_t + w\hat{C}_t]^{-1}[(1 - w)g_t + w\hat{g}_t] = [C_t + \lambda\hat{C}_t]^{-1}[g_t + \lambda\hat{g}_t],$$

where $\hat{g}_t$ and $\hat{C}_t$ are the stochastic estimates generated during iteration t, w is the step-size in our algorithm, and $\lambda = w/(1-w)$ is the effective step-size as it usually defined in the stochastic approximation literature. To characterize this update for small values of λ we perform a first order Taylor expansion of η_{t+1} around $\lambda = 0$, which gives

$$\eta_{t+1} = \eta_t - \lambda C_t^{-1}(\hat{C}_t \eta_t - \hat{g}_t) + \mathcal{O}(\lambda^2). \tag{5.11}$$

Comparison with equation (5.7) shows that the stochastic term in this expression ($\hat{C}_t \eta_t - \hat{g}_t$) is an unbiased estimate of the gradient of the KL-divergence $D[q_{\eta_t}(x)|p(x,y)]$. Up to first order, the update equation in (5.11) thus represents a stochastic gradient descent step, pre-conditioned with the C_t^{-1} matrix. Since this pre-conditioner is independent of the stochastic gradient approximation at iteration t, this gives a valid adaptive stochastic gradient descent algorithm, to which all the usual convergence results apply (see e.g. Amari, 1997).

If we take small steps, the pre-conditioner C_t^{-1} of equation (5.11) will be close to the Riemannian metric $\mathbb{E}_{q_t}\hat{C}_t = \mathbb{E}_{q_t}[T(x)'T(x)]$ used in natural gradient descent algorithms like that of Honkela et al. (2010). For certain exponential family distributions this metric can be calculated analytically, which would suggest performing stochastic natural gradient descent optimization with updates of the form

$$\eta_{t+1} = \eta_t - \lambda \left(\eta_t - \mathbb{E}_{q_t}[T(x)'T(x)]^{-1}[T(x^*)' \log p(x^*, y)]\right),$$

where the $\mathbb{E}_{q_t}[T(x)' \log p(x,y)]$ term is approximated using Monte Carlo, but $\mathbb{E}_{q_t}[T(x)'T(x)]$ is calculated analytically. At first glance, our approach of approximating $\mathbb{E}_{q_t}[T(x)'T(x)]$ using Monte Carlo only seems to add to the randomness of the gradient estimate, and using the same random numbers to approximate both $\mathbb{E}_{q_t}[T(x)' \log p(x,y)]$ and $\mathbb{E}_{q_t}[T(x)'T(x)]$ leads to biased pre-conditioned gradient approximations at that (although that bias disappears as $\lambda \to 0$). However, it turns out that approximating both terms using the same random draws in fact increases the efficiency of our algorithm dramatically. This reason for this is similar to the reason that the optimal estimator in linear regression is given by $(X'X)^{-1}X'y$ and not $\mathbb{E}[X'X]^{-1}X'y$: by using the same randomness for both the $X'X$ and $X'y$ terms, a large part of the noise in their product cancels out.

A particularly interesting example of this is when the true posterior distribution is of the same functional form as its approximation, say $p(x, y) = \exp[\tilde{T}(x)\xi]$, in which case Algorithm 2 will recover the true posterior exactly in $2(k+1)$ iterations, with k the number of sufficient statistics in q and p. Assuming the last $k+1$ samples $x_i, i = 1, ..., k+1$ generated by our algorithm are unique (which holds almost surely for continuous distributions q), we have

$$\begin{aligned} \hat{\eta} &= \left(\sum_{i=1}^{k+1} \tilde{T}(x_i)'\tilde{T}(x_i)\right)^{-1} \sum_{i=1}^{k+1} \tilde{T}(x_i)' \log[p(x_i, y)] \\ &= \left(\sum_{i=1}^{k+1} \tilde{T}(x_i)'\tilde{T}(x_i)\right)^{-1} \sum_{i=1}^{k+1} \tilde{T}(x_i)'\tilde{T}(x_i)\xi = \xi. \end{aligned} \tag{5.12}$$

If the algorithm is run for additional iterations after the true posterior is recovered, the approximation will not change. This is to be contrasted with other stochastic gradient descent algorithms which have non-vanishing variance for a finite number of samples, and is due to the fact that our regression in itself is *noise free*: only its support points are stochastic. This exact convergence will not hold for cases of actual interest, where p and q will not be of the exact same functional form, but we generally still observe much improvement when using Algorithm 2 instead of more conventional stochastic gradient descent algorithms. A deeper analysis of the variance of our stochastic approximation is given in Appendix 5.A.4.

Contrary to most applications in the literature, Algorithm 2 uses a fixed step size $w = 1/\sqrt{N}$ rather than a declining one in updating our statistics. The analyses of Robbins and Monro (1951) and Amari (1997) show that a sequence of learning rates $w_t = ct^{-1}$ is asymptotically efficient in stochastic gradient descent as the number of iterations N goes to infinity, but this conclusion rests on very strong assumptions on the functional form of the objective function (e.g. strong convexity) that are not satisfied for the problems we are interested in. Moreover, with a finite number of iterations N, the effectiveness of a sequence of learning rates that decays this fast is highly dependent on the proportionality constant c. If we choose c either too low or too high, it may take an extremely long time to reach the efficient asymptotic regime of this learning rate sequence.

Nemirovski et al. (2009) show that a more robust approach is to use a constant learning

rate $w = 1/\sqrt{N}$ and that this is optimal for finite N without putting stringent requirements on the objective function. In order to reduce the variance of the last iterate with this non-vanishing learning rate, they propose to use an average of the last L iterates as the final output of the optimization algorithm. The value of L should grow with the total number of iterations, and is usually chosen to be equal to $N/2$. Remarkably, they show that such an averaging procedure can match the asymptotic efficiency of the optimal learning sequence $w_t = ct^{-1}$.

For our particular optimization problem we have observed excellent results using constant learning rate $w = 1/\sqrt{N}$, and averaging starting half-way into the optimization. Note that we perform this averaging on the statistics g and C, rather than on the parameters $\eta = C^{-1}g$, which is statistically more efficient for our application. Using this set-up, g and C are actually weighted MC estimates where the weight of the j-th MC sample during the t-th iteration ($j \leq t$) is given by $w(1-w)^{t-j}$. Since $w \in (0, 1)$, this means that the weight of earlier MC samples declines as the algorithm advances, which is desirable since we expect q to be closer to optimal later in the algorithm's progression.

If the initial guess for η is very far from the optimal value, or if the number of steps N is very small, it can sometimes occur that the algorithm proposes a new value for η that does not define a proper distribution, for example because the η values correspond to a negative variance. This is a sign that the number of iterations should be increased: since our algorithm becomes a pre-conditioned gradient descent algorithm as the number of steps goes to infinity, the algorithm is guaranteed to converge if the step size is small enough. In addition, note that the exact convergence result presented in equation (5.12) suggests that divergence is very unlikely if $q_\eta(x)$ and $p(x, y)$ are close in functional form: choosing a good type of approximation will thus also help to ensure fast convergence. Picking a good first guess for η also helps the algorithm to converge more quickly. For very difficult cases it might therefore be worthwhile to base this guess on a first rough approximation of the posterior, for example by choosing η to match the curvature of $\log p(x, y)$ at its mode. For all our applications we found that a simple first guess for η and a large enough number of iterations was sufficient to guarantee a stable algorithm.

5.5 Marginal likelihood and approximation quality

The stochastic approximation algorithm presented in the last section serves to minimize the Kullback-Leibler divergence between $q_\eta(x)$ and $p(x|y)$, given by

$$D(q_\eta|p) = \mathbb{E}_{q_\eta}\left[\log \frac{q_\eta(x)}{p(x|y)}\right] = \mathbb{E}_{q_\eta}\left[\log \frac{q_\eta(x)}{p(x,y)}\right] + \log p(y),$$

which shows that we need to know the marginal likelihood $p(y)$ (the normalizing constant of the posterior) in order to evaluate this Kullback-Leibler divergence. As discussed before, we do not need to know this constant in order to *minimize* $D(q_\eta|p)$ as $p(y)$ does not depend on η, but we do need to know it if we want to determine the quality of the approximation, as measured by the final KL-divergence. In addition, the constant $p(y)$ is also essential for performing Bayesian model comparison or model averaging.

When our algorithm has converged, we have the following identity

$$\log p(x,y) = \hat{\eta}_0 + \log q_\eta(x) + r(x),$$

where $r(x)$ is the 'residual' or 'error term' in the linear regression of $\log p(x,y)$ on the sufficient statistics of $q_\eta(x)$. The intercept of the regression is

$$\hat{\eta}_0 = \mathbb{E}_{q_\eta}\left[\log p(x,y) - \log q_\eta(x)\right],$$

the usual VB lower bound on the marginal likelihood. Exponentiating this term yields

$$p(x,y) = \exp(\hat{\eta}_0) q_\eta(x) \exp(r(x)),$$

which we need to integrate with respect to x in order to find the marginal likelihood $p(y)$. Doing so gives

$$p(y) = \exp(\hat{\eta}_0)\mathbb{E}_{q_\eta}[\exp(r(x))]. \tag{5.13}$$

At convergence we have that $\mathbb{E}_{q_\eta}[r(x)] = 0$. Jensen's inequality then tells us that

$$\mathbb{E}_{q_\eta}[\exp(r(x))] \geq 1,$$

which shows that $\hat{\eta}_0$ is indeed a lower bound on the log marginal likelihood as we claimed earlier. If our approximation is perfect, the KL-divergence is zero and $r(x)$ is zero almost everywhere. In that case the residual term vanishes and the lower bound will be tight, otherwise it will underestimate the true marginal likelihood. The lower bound $\hat{\eta}_0$ is often used in model comparison, which works well if the KL-divergence between the approximate and true posterior distribution is of approximately the same size for all models that are being compared. However, if we compare two very different models this will often not be the case, and the model comparison will be biased as a result. In addition, as opposed to the exact marginal likelihood, the lower bound gives us no information on the quality of our posterior approximation. It would therefore be useful to obtain a better estimate of the marginal likelihood.

One approach to doing this would be to evaluate the expectation in (5.13) using Monte Carlo sampling. Some analysis shows that this corresponds to approximating $p(y)$ using importance sampling, with $q_\eta(x)$ as the candidate distribution. It is well known that this estimator of the marginal likelihood may have infinite variance, unless $r(x)$ is bounded from above. In general, we cannot guarantee the boundedness of $r(x)$ for our approach, so we will instead approximate the expectation in (5.13) using something that is easier to calculate.

At convergence, we know that the mean of $r(x)$ is zero when sampling from $q_\eta(x)$. The variance of $r(x)$ can easily be estimated using the mean squared error of the regressions we perform during the optimization, with relatively low variance. We denote our estimate of this variance by s^2. The assumption we will then make in order to approximate $\log p(y)$ is that $r(x)$ is approximately distributed as a normal random variable with these two moments. This leads to the following simple estimate of the log marginal likelihood

$$\log p(y) \approx \hat{\eta}_0 + \frac{1}{2}s^2.$$

That is, our estimate of the marginal likelihood is equal to its lower bound plus a correction term that captures the error in our posterior approximation $q_\eta(x)$. Similarly, we can

approximate the KL-divergence of our posterior approximation as

$$D(q_\eta|p) \approx \frac{1}{2}s^2.$$

The KL-divergence is approximately equal to half the mean squared error in the regression of $\log p(x, y)$ on the sufficient statistics of the approximation. This relationship should not come as a surprise: this mean squared error is exactly what we minimize when we do a linear regression. Our experiments indicate that this approximation of the KL-divergence can be quite accurate (see Section 5.8.3), especially when the approximation $q_\eta(x)$ is reasonably good.

Note that the scale of the KL-divergence is highly dependent on the amount of curvature in $\log p(x|y)$ and is therefore not easily comparable across different problems. If we scale the approximate KL-divergence to account for this curvature, this naturally leads to the R-squared measure of fit for regression modeling:

$$R^2 = 1 - \frac{s^2}{\mathrm{Var}_q[\log p(x, y)]}$$

The R-squared measure corrects for the amount of curvature in the posterior distribution and is therefore comparable across different models and data sets. In addition it is a well-known measure and easily interpretable. We therefore propose to use the R-squared as the measure of approximation quality for our variational posterior approximations. Although we find the R-squared to be a useful measure for the majority of applications, it is important to realize that it mostly contains information about the mass of the posterior distribution and its approximation, and not directly about their moments. It is therefore possible to construct pathological examples in which the R-squared is relatively high, yet the (higher) moments of the posterior and its approximation are quite different. This may for example occur if the posterior distribution has very fat tails.

The discussion up to this point represents the core of the ideas behind this chapter. To make our approach more general and computationally efficient we now provide a number of extensions in two separate sections. Section 5.6 discusses modifications of our stochastic approximation algorithm to improve efficiency, and Section 5.7 generalizes the exponential family approximations $q(x)$ used so far to include mixtures of exponential

family distributions. Some examples of using our method in practice are given in Section 5.8. Finally, Section 5.9 concludes.

5.6 Extensions I: Improving algorithmic efficiency

Algorithm 2 approximates the regression statistics $\mathbb{E}_{q_\eta}[\tilde{T}(x)' \log p(x,y)]$ and $\mathbb{E}_{q_\eta}[\tilde{T}(x)'\tilde{T}(x)]$ by simply drawing a sample x^* from $q_\eta(x)$ and using this sample to calculate

$$\hat{g}_t = \tilde{T}(x^*)' \log p(x^*, y) \tag{5.14}$$

$$\hat{C}_t = \tilde{T}(x^*)'\tilde{T}(x^*) \tag{5.15}$$

This works remarkably well because, as Section 5.4 explains, using the same random draw x^* to form both estimates, part of the random variation in $\eta = C^{-1}g$ cancels out. However, it is certainly not the only method of obtaining unbiased approximations of the required expectations, and in this section we present alternatives that often work even better. In addition, we also present alternative methods of parameterizing our problem, and we discuss ways of speeding up the regression step of our algorithm.

5.6.1 Making use of conditional independencies

For most statistical problems, the log posterior can be decomposed into a number of additive factors, i.e. $\log p(x,y) = \sum_{j=1}^{N} \log \phi_j(x,y)$. The optimality condition in equation (5.9) can then also be written as a sum:

$$\tilde{\eta} = \sum_{j=1}^{N} \mathbb{E}_q[\tilde{T}(x)'\tilde{T}(x)]^{-1}\mathbb{E}_q[\tilde{T}(x)' \log \phi_j(x,y)]$$

This means that rather than performing one single linear regression we can equivalently perform N separate regressions.

$$\hat{\eta} = \sum_{j=1}^{N} \hat{\eta}^j \tag{5.16}$$

$$\hat{\eta}^j = \mathbb{E}_q[\tilde{T}(x)'\tilde{T}(x)]^{-1}\mathbb{E}_q[\tilde{T}(x)' \log \phi_j(x,y)] \tag{5.17}$$

The practical benefit of this is that these separate regressions are often of much lower dimension: We know that element i of $\hat{\eta}_j$ will only be non-zero if the i-th sufficient statistic $\tilde{T}_i(x)$ has non-zero partial correlation to $\log \phi_j(x, y)$. Since the separate factors $\log \phi_j(x, y)$ often involve only a subset of the variables in x, this means that we can omit many of the sufficient statistics in performing each regression. That is, we have

$$\hat{\eta}_R^j = \mathbb{E}_q[\tilde{T}_R(x)'\tilde{T}_R(x)]^{-1}\mathbb{E}_q[\tilde{T}_R(x)' \log \phi_j(x, y)]$$

with $\tilde{T}_R(x)$ the relevant subset of $\tilde{T}(x)$, and $\hat{\eta}_R^j$ the corresponding subset in $\hat{\eta}^j$. The remaining elements in $\hat{\eta}^j$ will be zero. By performing these lower dimensional regressions we can reduce the variance of the stochastic approximation algorithm, as well as reduce the overhead needed to store and invert $C = \mathbb{E}_q[\tilde{T}(x)'\tilde{T}(x)]$.

In those rare cases where there are no conditional independencies in the posterior and we have to use the full C matrix, computing C^{-1} explicitly (which is $\mathcal{O}(k^3)$, with k the number of sufficient statistics) is not recommended, but if desired then C^{-1} should be updated each iteration using rank-one updates (i.e. using the matrix inversion lemma) which cost $\mathcal{O}(k^2)$. Similar low cost updates could be used to maintain Cholesky decompositions since C is symmetric, which is a numerically stable and efficient option. In very high dimensions one could also use conjugate gradients to solve $C^{-1}g$ approximately, using the current variational parameters η for a warm start.

5.6.2 Using the gradient of the log posterior

Using the Frisch-Waugh-Lovell theorem (Lovell, 2008), we can remove the constant from the sufficient statistics $\tilde{T}(x)$ and rewrite the optimality condition (5.9) in its normalized form (this is shown for our particular application in Appendix 5.A.1):

$$\hat{\eta} = \mathrm{Cov}_q[T(x), T(x)]^{-1} \mathrm{Cov}_q[T(x), \log p(x, y)]. \tag{5.18}$$

Furthermore, using the properties of the exponential family of distributions, we know that

$$\mathrm{Cov}_q[T(x), T(x)] = \nabla_\eta \mathbb{E}_{q_\eta}[T(x)] \tag{5.19}$$

and

$$\mathrm{Cov}_q[T(x), \log p(x,y)] = \nabla_\eta \mathbb{E}_{q_\eta}[\log p(x,y)] \tag{5.20}$$

Both $\mathbb{E}_{q_\eta}[T(x)]$ and $\mathbb{E}_{q_\eta}[\log p(x,y)]$ can be approximated without bias using Monte Carlo. By differentiating these Monte Carlo approximations we can then obtain unbiased estimates of their derivatives. This is easy to do as long as the pseudo-random draw x^* from $q_\eta(x)$ is a differentiable function of the parameters η, given our random number seed z^*.

$$x^* = f(\eta, z^*), \text{ with } z^* \text{ such that } x^* \sim q_\eta(x) \tag{5.21}$$

$$\hat{g} = \nabla_\eta \log p(f(\eta, z^*), y) = \nabla_\eta f(\eta, z^*) \nabla_x \log p(x^*, y) \tag{5.22}$$

$$\hat{C} = \nabla_\eta T(f(\eta, z^*)) = \nabla_\eta f(\eta, z^*) \nabla_x T(x^*) \tag{5.23}$$

By using the same random number seed z^* in both Monte Carlo approximations we once again get the beneficial variance reduction effect described in Section 5.4. Empirically, we find that using gradients often leads to a more efficient stochastic optimization algorithm. For some applications the posterior distribution will not be differentiable in some of the elements of x, for example when x is discrete. In that case the stochastic approximations presented here may be combined with the basic approximation of Section 5.4.

Note that for many samplers $\nabla_\eta f(\eta, z^*)$ is not defined, e.g. rejection samplers. However, for the gradient approximations it does not matter what type of sampler is actually used to draw x^*, only that it is from the correct distribution. A correct strategy is therefore to draw x^* using any desired sampling algorithm, and then proceeding as if we had used a different sampling algorithm for which $\nabla_\eta f(\eta, z^*)$ is defined. For example, we may calculate expression (5.21) as if we had used an inverse-transform sampler to sample x^*, for which we have

$$\nabla_\eta f(\eta, z^*) = -\frac{\nabla_\eta \Phi_\eta(x^*)}{\phi_\eta(x^*)}$$

with $\Phi_\eta(x^*)$ the CDF and $\phi_\eta(x^*)$ the pdf of the sampler.

5.6.3 Using the Hessian of the log posterior

When we have both first and second order gradient information for $\log p(x,y)$ and if we choose our approximation to be multivariate Gaussian, i.e. $q_\eta(x) = N(m(\eta), V(\eta))$, we

have a third option for approximating the statistics used in the regression. For Gaussian $q(x)$ and twice differentiable $\log p(x, y)$, Minka (2001b) and Opper and Archambeau (2009) show that

$$\nabla_m \mathbb{E}_q[\log p(x, y)] = \mathbb{E}_q[\nabla_x \log p(x, y)] \tag{5.24}$$

and

$$\nabla_V \mathbb{E}_q[\log p(x, y)] = \frac{1}{2}\mathbb{E}_q[\nabla_x \nabla_x \log p(x, y)] \tag{5.25}$$

where $\nabla_x \nabla_x \log p(x, y)$ denotes the Hessian matrix of $\log p(x, y)$ in x.

For the multivariate Gaussian distribution we know that the natural parameters are given as $\eta_1 = V^{-1}m$ and $\eta_2 = V^{-1}$. Using this relationship, we can derive Monte Carlo estimators $\hat{g}$ and $\hat{C}$ using the identities (5.19, 5.20). We find that these stochastic approximations are often even more efficient than the ones in Section 5.6.2, provided that the Hessian matrix of $\log p(x, y)$ can be calculated cheaply.

5.6.4 Using analytic expectations where possible

In many cases it is possible to calculate the contributions of some of the factors $\log \phi_i(x)$ to the stochastic approximations C and g analytically, while for others it is not. For example, this occurs when part of $\log p(x, y)$ (most often the prior) is conjugate to the posterior approximation $q_\eta(x)$. Even for some non-conjugate factors it might be possible to calculate certain expectations analytically. Using these exact expectations rather than their stochastic estimates can help reduce the variance of the approximations as well as reduce the time required to compute them, both of which increase the efficiency of the optimization procedure.

5.6.5 Subsampling the data: double stochastic approximation

The stochastic approximations derived above are all linear functions of $\log p(x, y)$ and its first and second derivatives. This means that these estimates are still unbiased even if we take $\log p(x, y)$ to be a noisy unbiased estimate of the true log posterior, rather than the exact log posterior. For most statistical applications $\log p(x, y)$ itself is a separable additive function of a number of independent factors, i.e. $\log p(x, y) = \sum_{i=1}^{N} \log \phi_i(x)$. These

$\log \phi_i(x)$ terms can be the likelihood contributions of individual observed data points, but they can also arise through conditional independencies between the x variables in the posterior. Using this fact we can construct an unbiased stochastic approximation of $\log p(x, y)$ as

$$\log \tilde{p}(x, y) = \frac{N}{K} \sum_{j=1}^{K} \log \phi_j(x) \tag{5.26}$$

where the K factors $\log \phi_j(x)$ are randomly selected from the total N factors. This approach was previously proposed for online learning of topic models by Hoffman et al. (2010). Since $\log \tilde{p}(x, y)$ has $\log p(x, y)$ as its expectation, performing stochastic approximation based on $\tilde{p}(x, y)$ converges to the same solution as when using $p(x, y)$, provided we resample the factors in $\log \tilde{p}(x, y)$ at every iteration. By subsampling the $K \ll N$ factors in the model the individual steps of the optimization procedure become more noisy, but since we can calculate $\tilde{p}(x, y)$ faster than we can $p(x, y)$, we can perform a larger number of steps in the same amount of time. If the number of factors in the posterior is especially large, this tradeoff often favors using subsampling. This principle has been used in many successful applications of stochastic gradient descent, see e.g. Bottou (2010).

5.6.6 Linear transformations of the regression problem

It is well known that classical linear least squares regression is invariant to invertible linear transformations of the explanatory variables. We can use the same principle in our stochastic approximation algorithm to allow us to work with alternative parameterizations of the approximate posterior $q(x)$. These alternative forms can be easier to implement or lead to more efficient algorithms, as we show in this section.

In classical linear least squares regression, we have an $N \times D$ matrix of explanatory variables X, and an $N \times 1$ vector of dependent variables Y. Instead of doing a linear regression with these variables directly, we may equivalently perform the linear regression using a transformed set of explanatory variables $\tilde{X} = XK'$, with K any invertible matrix of size $D \times D$. The least squares estimator $\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$ of the transformed problem can then be used to give the least squares estimator of the original problem as $\hat{\beta} = K'\tilde{\beta}$:

$$\hat{\beta} = K'(KX'XK')^{-1}KX'Y = (KX'X)^{-1}KX'Y = (X'X)^{-1}X'Y.$$

Using the same principle, we can rewrite the optimality condition of equation (5.9) as

$$\tilde{\eta} = \mathbb{E}_{q_\eta}[K(\eta)\tilde{T}(x)'\tilde{T}(x)]^{-1}\mathbb{E}_{q_\eta}[K(\eta)\tilde{T}(x)' \log p(x,y)], \tag{5.27}$$

for any invertible matrix K, which may depend on the variational parameters η. Instead of solving our original least squares regression problem, we may thus equivalently solve this transformed version. When we perform the linear regression of equation (5.27) for a fixed set of parameters η, the result will obviously be identical to that of the original regression with $K(\eta) = \mathrm{I}$, as long as we use the same random numbers for both regressions. However, when the Monte Carlo samples ('data points' in our regression) are generated using different values of η, as is the case with the proposed stochastic approximation algorithm, the two regressions will not necessarily give the same solution. If the true posterior $p(x|y)$ is of the same functional form as the approximation q_η, the exact convergence result of Section 5.4 holds for any invertible $K(\eta)$, so it is not immediately obvious which $K(\eta)$ is best for general applications.

We hypothesize that certain choices of $K(\eta)$ may lead to statistically more efficient stochastic approximation algorithms for certain specific problems, but we will not pursue this idea here. What we will discuss is the observation that the stochastic approximation algorithm may be easier to implement for some choices of $K(\eta)$ than for others, and that the computational costs are not identical for all $K(\eta)$. In particular, it is worth noting that the transformation $K(\eta)$ allows us to use different parameterizations of the variational approximation. Let q_ϕ be such a reparameterization of the approximation, let the new parameter vector $\phi(\eta)$ be an invertible and differentiable transformation of the original parameters η, and set $K(\eta)$ equal to the inverse Jacobian of this transformation, i.e. $K(\eta) = [\nabla_\eta \phi(\eta)]^{-1}$. Using the properties of the exponential family of distributions, we can then show that

$$K(\eta)\,\mathrm{Cov}_{q_\phi}[T(x), h(x)] = \nabla_\phi \mathbb{E}_{q_\phi}[h(x)], \tag{5.28}$$

for any differentiable function $h(x)$. Using this result, the stochastic approximations of

Section 5.6.2 for the transformed regression problem are found to be

$$x^* = f(\phi, z^*), \text{ with } z^* \text{ such that } x^* \sim q_\phi(x) \tag{5.29}$$

$$\hat{g} = \nabla_\phi \log p(f(\phi, z^*), y) \tag{5.30}$$

$$\hat{C} = \nabla_\phi T(f(\phi, z^*)). \tag{5.31}$$

These new expressions for $\hat{g}$ and $\hat{C}$ may be easier to calculate than the original ones (5.21), and the resulting $\hat{C}$ may have a structure making it easier to invert in some cases. A particularly striking example of this occurs when we use a Gaussian approximation in combination with the stochastic approximations of Section 5.6.3, using the gradient and Hessian of $\log p(x, y)$. In this case we may work in the usual natural parameterization, but doing so gives a dense matrix $\hat{C}$ with dimensions proportional to p^2, where p is the dimension of x. For large p, such a stochastic approximation is expensive to store and invert. However, using the stochastic approximations above, we may also parameterize our approximation in terms of the mean m and variance V, and work with these parameters directly. Doing so leads to the following sparse regression algorithm, as derived in Appendix 5.A.2.

Algorithm 3 Stochastic Approximation for Gaussian Variational Approximation

Require: An unnormalized, twice differentiable posterior distribution $p(x, y)$
Require: The total number of iterations N
Initialize the mean and variance of the approximation (m, V) to a first guess, for example by matching the prior $p(x)$
Initialize $z = m$, $P = V^{-1}$ and $a = 0$
Initialize $\bar{z} = 0$, $\bar{P} = \mathbf{0}$ and $\bar{a} = \mathbf{0}$
Step-size $w = 1/\sqrt{N}$
for $t = 1 : N$ **do**
  Set $V = P^{-1}$ and $m = Va + z$
  Generate a draw x^* from $N(m, V)$
  Calculate the gradient g_t and Hessian H_t of $\log p(x, y)$ at x^*
  Set $a = (1 - w)a + wg_t$
  Set $P = (1 - w)P - wH_t$
  Set $z = (1 - w)z - wx^*$
  if $t > N/2$ **then**
    Set $\bar{a} = \bar{a} + g_t$
    Set $\bar{P} = \bar{P} - H_t$
    Set $\bar{z} = \bar{z} + x^*$
  end if
end for
Set $V = \bar{P}^{-1}$ and $m = V\bar{a} + \bar{z}$
return m, V

Instead of storing and inverting the full C matrix, this algorithm uses the sparsity induced by the transformation $K(\eta)$ to work with the precision matrix P instead. The dimensions of this matrix are equal to p, rather than its square, providing great savings. Moreover, while the C matrix in the original parameterization is always dense, P will have the same sparsity pattern as the Hessian of $\log p(x, y)$, which may reduce the costs of storing and inverting it even further for many applications. An example using this algorithm can be found in Section 5.8.1.

5.7 Extensions II: Using mixtures of exponential family distributions

So far, we have assumed that the approximating distribution $q_\eta(x)$ is a member of the exponential family. Here we will relax that assumption. If we choose a non-standard approximation, this most likely means that certain moments or marginals of $q_\eta(x)$ are

no longer available analytically, which should be taken into account when choosing the type of approximation. However, if we can at least sample directly from $q_\eta(x)$, it is often still much cheaper to approximate these moments using Monte Carlo than it would be to approximate the corresponding moments of $p(x|y)$ using MCMC or other indirect sampling methods. We have identified two general strategies for constructing useful non-standard posterior approximations which are discussed in the following two sections.

5.7.1 Hierarchical approximations

If we split our vector of unknown parameters x into p non-overlapping blocks, our approximating posterior may be decomposed as

$$q(x) = q(x_1)q(x_2|x_1)q(x_3|x_2, x_1)\dots q(x_p|x_{p-1}, \dots, x_1).$$

If we then choose every conditional posterior $q(x_i|x_{i-1}, x_{i-2}, \dots, x_1)$ to be of a standard form, we can easily sample from the joint $q(x)$, while still having much more freedom in capturing the dependence between the different blocks of x. In practice, such a conditionally standard approximation can be achieved by specifying the sufficient statistics of each standard block $q(x_i|x_{i-1}, x_{i-2}, \dots, x_1)$ to be a function of the preceding elements $x_{i-1}, x_{i-2}, \dots, x_1$. This leads to a natural type of approximation for hierarchical Bayesian models, where the hierarchical structure of the prior often suggests a good hierarchical structure for the posterior approximation.

If every conditional $q(x_i|x_{i-1}, x_{i-2}, \dots, x_1)$ is in the exponential family, the joint may not be if the normalizing constant of $q(x_i|x_{i-1}, x_{i-2}, \dots, x_1)$ is a non-separable function of $x_{i-1}, x_{i-2}, \dots, x_1$ and the variational parameters. However, because the conditionals are still in the exponential family, our optimality condition still holds separately for the variational parameters of each conditional with only slight modification. In that case we therefore propose applying the optimization procedures separately to each of these conditionals. Without loss of generalization, consider the case where our posterior approximation consists of two factors: $q(x) = q_{\eta_1}(x_1)q_{\eta_2}(x_2|x_1)$. In its normalized form

(see Section 5.6.2), the optimality condition for the first factor is then given as

$$\eta_1 = \mathrm{Var}_q[T(x_1)]^{-1} \mathrm{Cov}_q[T(x_1), \log p(x, y) - \log q_{\eta_2}(x_2|x_1)],$$

where $T(x_1)$ denotes the sufficient statistics of $q_{\eta_1}(x_1)$. The optimality condition for the second block is

$$\eta_2 = \mathbb{E}_{q(x_1)}[\mathrm{Var}_{q(x_2|x_1)}(T(x_2))]^{-1}\mathbb{E}_{q(x_1)}[\mathrm{Cov}_{q(x_2|x_1)}(T(x_2), \log p(x, y) - \log q_{\eta_1}(x_1))],$$

where $T(x_2)$ denotes the sufficient statistics of $q_{\eta_2}(x_2|x_1)$. By making use of the conditional independencies discussed in Section 5.6.1 we can often simplify these expressions further for given problems.

Using this type of approximation, the marginals $q(x_i)$ will generally be mixtures of exponential family distributions, which is where the added flexibility of this method comes from. By allowing the marginals $q(x_i)$ to be mixtures with dependency on the preceding elements of x, we can achieve much better approximation quality than by forcing them to be of a standard form. A practical example of this in a hierarchical Bayesian model is given in Section 5.8.2.

5.7.2 Using auxiliary variables

Another method of constructing flexible posterior approximations is by using the conditionally standard approximation of Section 5.7.1, but by letting the first block of variables be a vector of *auxiliary variables* z, that are not part of the unknowns x. Doing this, the posterior approximation has the form

$$q(x, z) = q(z)q(x|z).$$

The factors $q(z)$ and $q(x|z)$ should both be of standard form, which allows the marginal approximation $q(x)$ to be a general mixture of exponential family distributions, like a mixture of normals for example. If we use enough mixture components, the approximation $q(x)$ could then in principle be made arbitrarily close to $p(y|x)$. This mixture approxima-

tion can then be fitted by performing the standard KL-divergence minimization:

$$\hat{\eta} = \arg\min_{\eta} \mathbb{E}_{q_\eta}[\log q_\eta(x) - \log p(x, y)] \tag{5.32}$$

From (5.32) it becomes clear that an additional requirement of this type of approximation is that we can integrate out the auxiliary variables z from the joint $q(x, z)$ in order to evaluate the marginal density $q(x)$ at a given point x. Fortunately this is easy to do for many interesting approximations, such as discrete mixtures of normals or continuous mixtures like Student's t distributions. Also apparent from equation (5.32) is that we cannot use this approximation directly with the stochastic approximation algorithms proposed in the last sections since $q(x)$ is itself not part of the exponential family of distributions. However, we can rewrite (5.32) as

$$\hat{\eta} = \arg\min_{\eta} \mathbb{E}_{q_\eta}[\log q_\eta(x, z) - \log \tilde{p}(x, y, z)], \tag{5.33}$$

with $\tilde{p}(x, y, z) = p(x, y)q_\eta(z|x)$, and

$$q_\eta(z|x) = \frac{q_\eta(x|z)q_\eta(z)}{\int q_\eta(x|z)q_\eta(z)dz}.$$

Equation (5.33) now once again has the usual form of a KL-divergence minimization with an approximation ($q_\eta(x, z)$) in the exponential family. By including the auxiliary variables z in the 'true' posterior density, we can thus once again make use of our efficient stochastic optimization algorithms. Note that including z in the posterior did not change the marginal posterior $p(x|y)$ which is what we are interested in. A practical example of this approach, using an approximation consisting of a mixture of normals, can be found in Section 5.8.3.

5.8 Examples

We demonstrate our proposed methodology on three problems from the literature.

5.8.1 Binary probit regression

Binary probit (and logistic) regression is a classic model in statistics, also referred to as binary classification in the machine learning literature. Here we take a Bayesian approach to probit regression to demonstrate the performance of our methodology relative to existing variational approaches. We have N observed data pairs $(y_i \in \{0, 1\}, \mathbf{x}_i \in \mathbb{R}^P)$, and we model $y|\mathbf{x}$ as $P(y = 1|\mathbf{x}, \mathbf{w}) = \phi(\mathbf{w}'\mathbf{x})$ where $\phi(.)$ is the standard Gaussian cdf and $\mathbf{w} \in \mathbb{R}^P$ is a vector of regression coefficients, for which we assume an elementwise Gaussian prior $N(0, 1)$. This is in fact a model for which existing approaches are straightforward so it is interesting to compare the performance. Of course the major benefit of our approach is that it can be applied in a much wider class of models.

We use data simulated from the model, with $N = 100$ and $P = 5$, to be able to show the performance averaged over many datasets (1000 in fact). We compare Algorithm 3 to the VBEM algorithm of Ormerod and Wand (2010) which makes use of the fact that the expectations required for this model can in fact be calculated analytically. We choose not to do this for our method to investigate how effective our MC estimation strategy can be. For completeness we also compare to variational message passing (VMP, Winn and Bishop, 2006), a message passing implementation of VBEM, and expectation propagation (EP, Minka, 2001b), which is known to have excellent performance on binary classification problems (Nickisch and Rasmussen, 2008). These last two alternatives are both implemented in Infer.NET (Minka et al., 2010) a library for probabilistic inference in graphical models, where we implement the first two methods ourselves in MATLAB.

Since this experiment is on synthetic data we are able to assess performance in terms of the method's ability to recover the known regression coefficients $\mathbf{w}$, which we quantify as the root mean squared error (RMSE) between the variational mean and the true regression weights, and the "log score": the log density of the true weights under the approximate variational posterior. The log score is useful because it rewards a method for finding good estimates of the posterior variance as well as the mean, which should of course be central to any approximate Bayesian method.

The results, shown in Figure 5.1 and 5.2, demonstrate that our method is able to outperform the standard analytic VBEM algorithm in terms of speed accuracy tradeoff. The

improvement in the RMSE is noticeable, but the difference in log score is dramatic, showing that Algorithm 3 gives significantly better estimates of the variance that VBEM. In fact our results are very similar to those of EP, which obtained an RMSE of 0.261 and log score of 0.079, but took an average of 18.2 milliseconds per run (note the system set ups are not completely comparable: EP was run on a laptop rather than a desktop, and Infer.NET is implemented in C# rather than Matlab). As expected VMP gave consistent results with VBEM: a RMSE of 0.268 and a log score of -4.85. The average R-squared obtained by our variational approximation was 0.97, indicating a close fit to the exact posterior distribution.

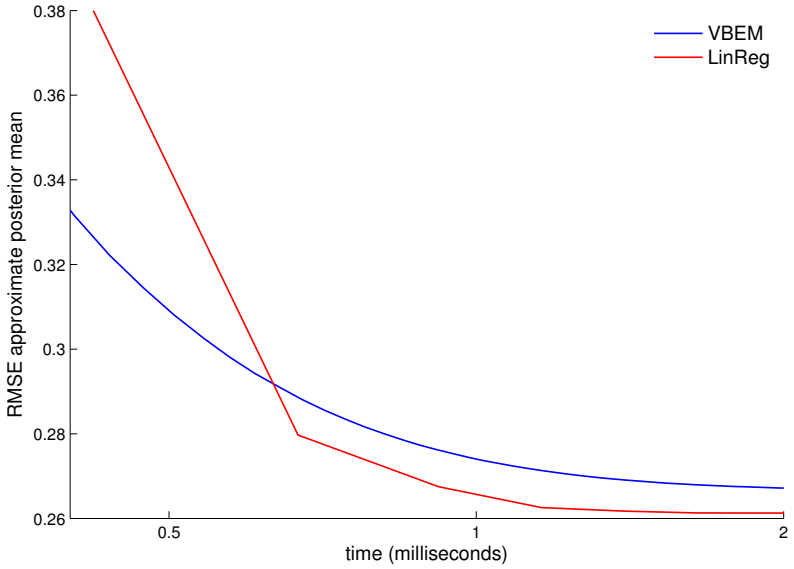


Figure 5.1: RMSE approximate posterior mean - Stochastic Linear Regressions v.s. VBEM

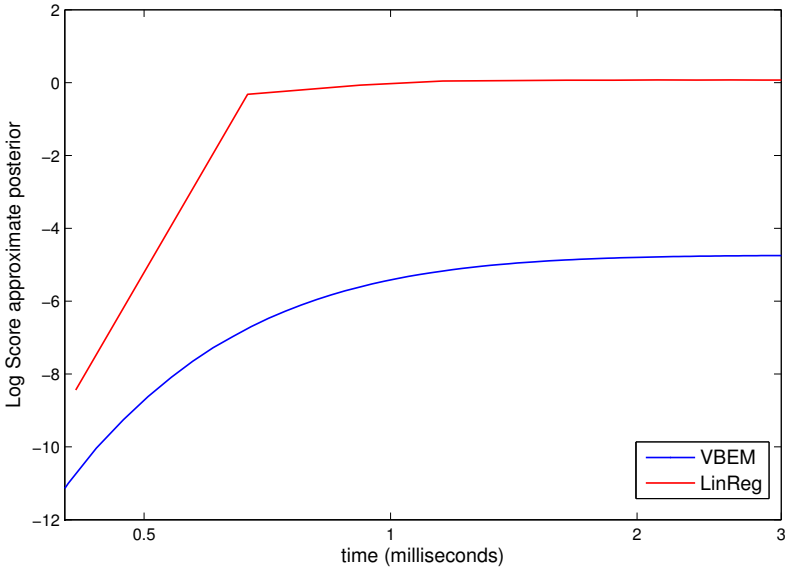


Figure 5.2: Log-Score of approximate posterior - Stochastic Linear Regressions v.s. VBEM

5.8.2 A stochastic volatility model

Stochastic volatility models for signals with time varying variances are considered extremely important in finance. Here we apply our methodology to the model and prior specified in Girolami and Calderhead (2011). The data we will use, from Kim et al. (1998), is the percentage change y_t in GB Pound v.s. US Dollar exchange rate, modeled as:

$$y_t = \epsilon_t \beta \exp(v_t/2).$$

The relative volatilities, v_t are governed by the autoregressive AR(1) process

$$v_{t+1} = \phi v_t + \xi_{t+1}, \text{ with } v_1 \sim N[0, \sigma^2/(1-\phi^2)].$$

The distributions of the error terms are given by $\epsilon_t \sim N(0,1)$ and $\xi_t \sim N(0,\sigma^2)$. The prior specification is as in Girolami and Calderhead (2011):

$$p(\beta) \propto \beta^{-1}, \qquad (\phi+1)/2 \sim \text{Beta}(20, 1.5), \qquad \sigma^2 \sim \text{Inv-Gamma}(5, 0.25)$$

Following Section 5.7.1 we use the hierarchical structure of the prior to suggest a hierarchical structure for the approximate posterior:

$$q_\eta(\phi, \sigma^2, \beta, v) = q_\eta(\phi) q_\eta(\sigma^2|\phi) q_\eta(\beta, v|\phi, \sigma^2).$$

The prior of ϕ is in the exponential family, so we choose the posterior approximation $q_\eta(\phi)$ to be of the same form:

$$q_\eta[(\phi + 1)/2] = \text{Beta}(\eta_1, \eta_2).$$

The prior for σ^2 is inverse-Gamma, which is also in the exponential family. We again choose the same functional form for the posterior approximation, but with a slight modification in order to capture the posterior dependency between ϕ and σ^2:

$$q_\eta(\sigma^2|\phi) \sim \text{Inv-Gamma}(\eta_3, \eta_4 + \eta_5\phi^2),$$

where the extra term $\eta_5\phi^2$ was chosen by examining the functional form of the exact full conditional $p(\sigma^2|\phi, v)$.

The conditional prior $p[\log(\beta), v|\phi, \sigma^2]$ can be seen as the diffuse limit of a multivariate normal distribution. We therefore also use a multivariate normal conditional approximate posterior:

$$q_\eta[(\log(\beta), v)|\phi, \sigma^2] = N(m, V),$$

with

$$V^{-1} = P(\phi, \sigma^2) + \eta_6 \text{ and } m = V^{-1}\eta_7$$

where $P(\phi, \sigma^2)$ is the precision (inverse covariance) matrix of $p[(\log(\beta), v)|\phi, \sigma^2]$, η_6 is a $T \times T$ matrix, and η_7 is a $T \times 1$ vector. Furthermore, an analysis following Opper and Archambeau (2009) shows that only a relatively small number of the elements of η_6 will be non-zero: all elements on the diagonal of η_6 and all elements in the column and row belonging to $\log(\beta)$.

Using the GB Pound vs US Dollar exchange rate data, the approximation above has almost 2000 free variational parameters to be optimized. This seems like a problemat-

ically large number, but is easily feasible by using algorithm 3 to fit $q_\eta[\log(\beta), v|\phi, \sigma^2]$ and the algorithm using only gradients (Section 5.6.2) to fit $q_\eta(\phi)$ and $q_\eta(\sigma^2|\phi)$. Expectations and normalizing constants for $q_\eta[\log(\beta), v|\phi, \sigma^2]$ can be calculated efficiently using the Kalman filter and smoother (see e.g. Durbin and Koopman, 2001). For the current application we therefore only need to sample ϕ and σ^2 each iteration, and not β and v.

We compare the results against the "true" posterior, provided by a very long run of the MCMC algorithm of Girolami and Calderhead (2011).

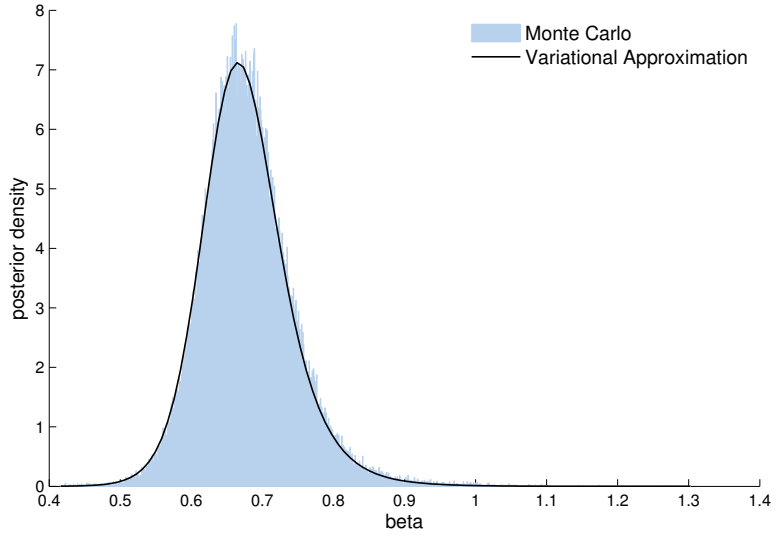


Figure 5.3: Exact and approximate posterior for the stochastic volatility model - β parameter

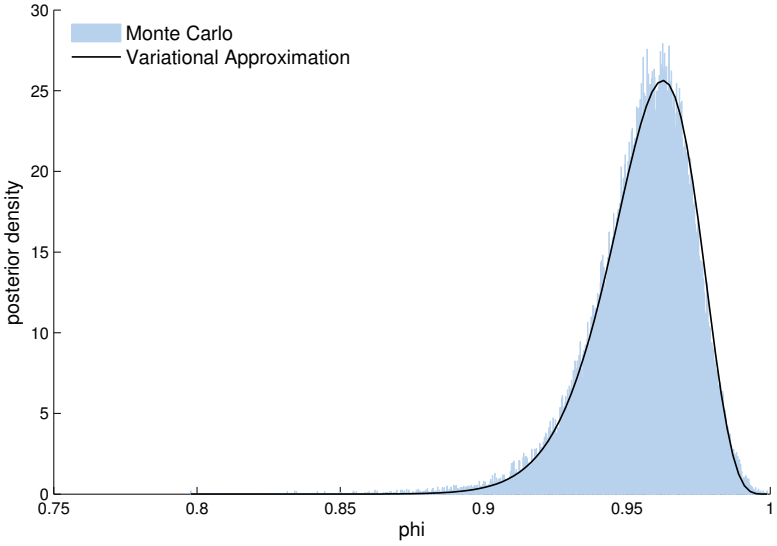


Figure 5.4: Exact and approximate posterior for the stochastic volatility model - ϕ parameter

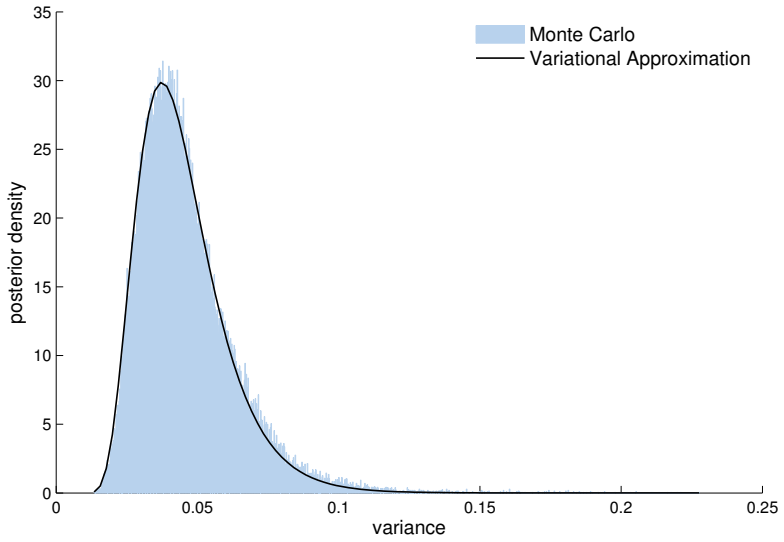


Figure 5.5: Exact and approximate posterior for the stochastic volatility model - σ^2 parameter

As can be seen from Figures 5.3, 5.4 and 5.5, the posterior approximations are nearly exact. The posterior approximation for β seems especially good (Figure 5.3), which is due to β being in the last block of the hierarchical posterior approximation. Similarly, the posterior approximations for the latent volatilities v (not shown) are also indistinguishable from the exact posterior.

Initializing $q_\eta(\phi)$ and $q_\eta(\sigma^2|\phi)$ to the prior, the above results can be obtained using 500 iterations of our algorithm, with a single (ϕ, σ^2) sample per iteration. Using these settings, the single-threaded Matlab implementation of our stochastic optimization algorithm requires just under a second to complete on a 3Ghz processor. This is more than two orders of magnitude faster than the running time required by advanced MCMC algorithms for this problem.

Our approach to doing inference in the stochastic volatility model shares some similarities with the approach of Liesenfeld and Richard (2008). They fit a Gaussian approximation to the posterior of the volatilities for given ϕ, σ^2, β parameters, using the importance sampling algorithm of Richard and Zhang (2007), which is based on auxiliary regressions somewhat similar to those in Algorithm 2. They then infer the model parameters using MCMC methods. The advantage of our method is that we are able to leverage the information in the gradient and Hessian of the posterior, and that our stochastic approximation algorithm allows us to fit the posterior approximation very quickly for all volatilities si-

multaneously, while their approach requires optimizing the approximation one volatility at a time. Unique to our approach is also the ability to concurrently fit a posterior approximation for the model parameters ϕ, σ^2, β and have the approximate posterior of the volatilities depend on these parameters, while Liesenfeld and Richard (2008) need to reconstruct their approximation every time a new set of model parameters is considered. As a result, our approach is significantly faster for this problem.

5.8.3 A beta-binomial model for overdispersion

Albert (2009, Section 5.4) considers the problem of estimating the rates of death from stomach cancer for the largest cities in Missouri. This cancer mortality data is available from the R package LearnBayes, and consists of 20 pairs (n_j, y_j) where n_j contains the number of individuals that were at risk in city j, and y_j is the number of cancer deaths that occurred in that city. The counts y_j are overdispersed compared to what one could expect under a binomial model with constant probability, so Albert (2009) assumes the following beta-binomial model with mean m and precision K

$$P(y_j|m, K) = \binom{n_j}{y_j} \frac{B(Km + y_j, K(1-m) + n_j - y_j)}{B(Km, K(1-m))},$$

where $B(\cdot, \cdot)$ denotes the Beta-function. The parameters m and K are given the following improper prior

$$p(m, K) \propto \frac{1}{m(1-m)} \frac{1}{(1+K)^2}.$$

The resulting posterior distribution is non-standard and extremely skewed. In order to ameliorate this, Albert (2009) proposes to use the reparameterization

$$\theta_1 = \text{logit}(m), \text{ and } \theta_2 = \log(K).$$

The form of the posterior distribution $p(\theta|n, y)$ still does not resemble any standard distribution, so we will approximate it using a finite mixture of L bivariate Gaussians. In order to do this, we first introduce an auxiliary variable z, to which we assign a categorical

approximate posterior distribution with L possible outcomes.

$$\log q_\eta(z) = \delta(z = 1)\eta_1 + \delta(z = 2)\eta_2 + \cdots + \delta(z = L)\eta_L,$$

where $\delta(.)$ is the indicator function.

Conditional on z, we assign θ a Gaussian approximate posterior

$$q_\eta(\theta|z = i) = N(\mu_i, \Sigma_i)$$

By adapting the true posterior as described in Section 5.7.2, we can fit this approximate posterior to $p(\theta|n, y)$. We do this by using the basic algorithm of Section 5.4. The regression statistics C and g used in the resulting algorithm depend linearly on the indicator vector $\delta(z^* = i)$, which denotes whether or not component i was used to sample θ^* in each iteration. Rather than using this indicator function directly, we use its Rao-Blackwellized version $\mathbb{E}[\delta(z^* = i)|\theta^*]$, where θ^* are the sampled parameters. The resulting stochastic estimates will have the same expectation as when using $\delta(z^* = i)$ itself, but with lower variance at no additional computational cost.

We fit these approximations using a varying number of mixture components L and examine the resulting KL-divergence from the true posterior density. Since this is a low dimensional problem, we can obtain this divergence very precisely using quadrature methods. This also allows us to assess the accuracy of the KL-divergence approximation derived in Section 5.5. Finally, we present a contour plot that visually shows that a good approximation can indeed be obtained using a large enough number of mixture components.

Figures 5.6 and 5.7 show that we can indeed approximate this skewed and fat-tailed density very well using a large enough number of Gaussians. The R-squared of the mixture approximation with 8 components is 0.997. Also apparent is the inadequacy of an approximation consisting of a single Gaussian for this problem, with an R-squared of only 0.82. This clearly illustrates the advantages of our approach which allows us to use much richer types of approximations than was previously possible. Furthermore, Figure 5.6 shows that the KL-divergence of the approximation to the true posterior can be approximated quite accurately using the measure developed in Section 5.5, especially if

the posterior approximation is reasonably good.

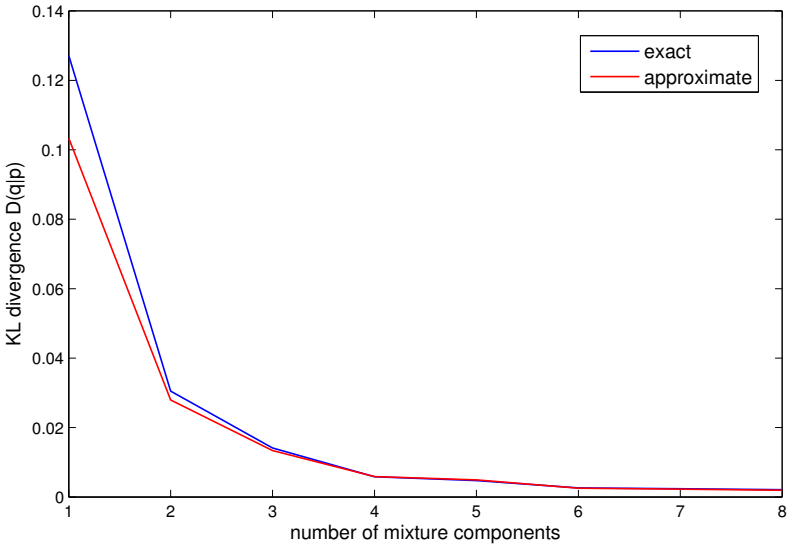


Figure 5.6: KL-divergence between the variational approximation and the exact posterior density for an increasing number of mixture components.

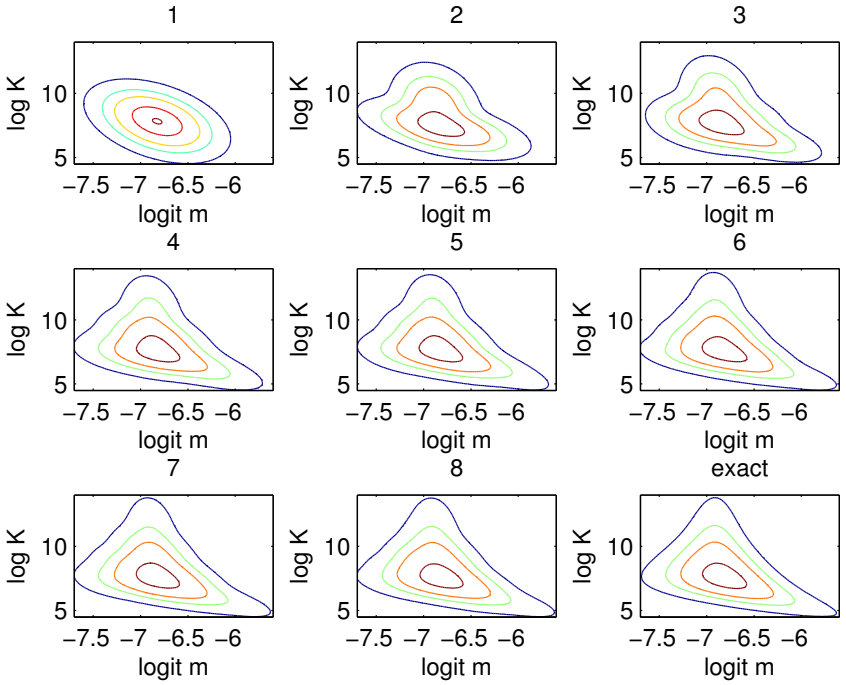


Figure 5.7: Contour plots of posterior approximations using 1-8 mixture components, with the exact posterior at the bottom-right.

5.9 Conclusion and future work

We have introduced a stochastic optimization scheme for variational inference inspired by a novel interpretation of linear regression of the target log density against the sufficient statistics of the approximating family. Our scheme allows very generic implementation for a wide class of models since in its most basic form only the unnormalized density of the target distribution is required, although we have shown how gradient or even Hessian information can be used if available. The generic nature of our methodology would lend itself naturally to a software package for Bayesian inference along the lines of Infer.NET (Minka et al., 2010) or WinBUGS (Gilks et al., 1994), and would allow inference in a considerably wider range of models. Incorporating automatic differentiation in such a package could clearly be beneficial. Automatic selection of the approximating family would be very appealing from a user perspective, but could be challenging in general.

Variational inference usually requires that we use conditionally conjugate models: since our method removes this restriction several possible avenues of research are opened. For example, for MCMC methods collapsed versions of models (i.e. with certain parameters or latent variables integrated out) sometimes permit much more efficient inference (Porteous et al., 2008) but adapting variational methods to work with collapsed models is complex and requires custom per model methodology (Teh et al., 2007). However, our methods is indifferent to whether the model is collapsed or not, so it would be straightforward to experiment with different representations of the same model.

We have shown it is straightforward to extend our methodology to use hierarchical structured approximations and more flexible approximating families such as mixtures. This closes the gap considerably relative to MCMC methods. Perhaps the biggest selling point of MCMC methods is their asymptotic limits: in practice this means simply running the MCMC chain for longer can give greater accuracy, an option not available to a researcher using variational methods. However, if we use a mixture approximating family then we can tune the computation time vs. accuracy trade off simply by varying the number of mixture components used. Another interesting direction of research along this line would be to use low rank approximating families such as factor analysis models.

It should be noted that it is possible to mix our method with VBEM, for example

using our method for any non-conjugate parts of the model and VBEM for variables that happen to be conjugate. This is closely related to the non-conjugate variational message passing (NCVMP) algorithm of Knowles and Minka (2011) implemented in Infer.NET, which aims to fit conjugate models while maintaining the convenient message passing formalism. Note that NCVMP only specifies how to perform the variational optimization, not how to approximate required integrals: in Infer.NET where analytic expectations are not available quadrature or secondary variational bounds are used, unlike the Monte Carlo approach proposed here.

5.A Appendix

5.A.1 Unnormalized to normalized optimality condition

The unnormalized optimality condition in (5.8) is

$$\tilde{\eta} = \left[\int \tilde{q}_{\tilde{\eta}}(x)\tilde{T}(x)'\tilde{T}(x)d\nu(x)\right]^{-1}\left[\int \tilde{q}_{\tilde{\eta}}(x)\tilde{T}(x)'\log p(x,y)d\nu(x)\right]. \tag{5.34}$$

Clearly we can replace $\tilde{q}$ by its normalized version $q(x) = \tilde{q}/Z(\eta)$ since the normalizing terms will cancel. Recalling $\tilde{T}(x) = (1, T(x))$ and $\tilde{\eta} = (\eta_0, \eta')'$ we then have

$$\begin{bmatrix} 1 & \mathbb{E}[T] \\ \mathbb{E}[T'] & \mathbb{E}[T'T] \end{bmatrix}^{-1} \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[TY] \end{pmatrix} = \begin{pmatrix} \eta_0 \\ \eta \end{pmatrix} \tag{5.35}$$

where $Y := \log p(x, y)$. Rearranging gives

$$\begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[TY] \end{pmatrix} = \begin{bmatrix} 1 & \mathbb{E}[T] \\ \mathbb{E}[T'] & \mathbb{E}[T'T] \end{bmatrix} \begin{pmatrix} \eta_0 \\ \eta \end{pmatrix} \tag{5.36}$$

Solving for η_0 easily gives

$$\eta_0 = \mathbb{E}[Y] - \mathbb{E}[T]\eta \tag{5.37}$$

$$\eta = \left(\mathbb{E}[T'T] - \mathbb{E}[T']\mathbb{E}[T]\right)^{-1}\left(\mathbb{E}[TY] - \mathbb{E}[T]\mathbb{E}[Y]\right) \tag{5.38}$$

$$= \mathrm{Cov}(T)^{-1}\,\mathrm{Cov}(T, Y) \tag{5.39}$$

5.A.2 Derivation of Gaussian variational approximation

For notational simplicity we will derive our stochastic approximation algorithm for Gaussian variational approximation (Algorithm 3) under the assumption that x is univariate. The extension to multivariate x is conceptually straightforward but much more tedious in terms of notation.

Let $p(x, y)$ be the unnormalized posterior distribution of a univariate random variable x, and let $q(x) = N(m, V)$ be its Gaussian approximation with sufficient statistics, $T(x) = (x, -0.5x^2)'$. In order to find the mean m and variance V that minimize the KL-divergence between $q(x)$ and $p(x|y)$ we solve the transformed regression problem defined in Equation (5.27), i.e.

$$\begin{aligned}\tilde{\eta} &= \mathbb{E}_{q_\eta}[K(\eta)\tilde{T}(x)'\tilde{T}(x)]^{-1}\mathbb{E}_{q_\eta}[K(\eta)\tilde{T}(x)' \log p(x, y)\\ &= C^{-1}g\end{aligned}$$

where

$$K(\eta) = [\nabla_\eta \phi(\eta)]^{-1},$$

with $\phi = (\phi_1, \phi_2)' = (m, V)'$ the usual mean-variance parameterization and where the natural parameters are given by $\eta = (V^{-1}m, V^{-1})'$. Recall identity (5.24) which states that

$$\nabla_{\phi_1}\mathbb{E}_{q_\phi}[h(x)] = \mathbb{E}_{q_\phi}[\nabla_x h(x)],$$

with $\phi_1 = m$ the first element of the parameter vector ϕ, and $g(x)$ any differentiable function. Similarly, identity (5.25) reads

$$\nabla_{\phi_2}\mathbb{E}_{q_\phi}[h(x)] = -\frac{1}{2}\mathbb{E}_{q_\phi}[\nabla_x \nabla_x h(x)],$$

with $\phi_2 = V$ the second element of the parameter vector. Using these identities we find

that the regression statistics for this optimization problem are given by

$$\begin{aligned} C &:= K(\eta)\,\mathrm{Cov}_{q_\phi}[T(x), T(x)] = \nabla_\phi \mathbb{E}_{q_\phi}[T(x)] \\ &= \mathbb{E}_{q_\phi}[\nabla_x T(x)] = \mathbb{E}_{q_\phi}\begin{bmatrix} 1 & -x \\ 0 & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & -\mathbb{E}_{q_\phi}[x] \\ 0 & -\frac{1}{2} \end{bmatrix}, \end{aligned}$$

and

$$\begin{aligned} g &:= K(\eta)\,\mathrm{Cov}_{q_\phi}[T(x), \log p(x,y)] \\ &= \nabla_\phi \mathbb{E}_{q_\phi}[\log p(x,y)] \\ \Rightarrow \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} &= \begin{bmatrix} \mathbb{E}_q[\nabla_x \log p(x,y)] \\ -\frac{1}{2}\mathbb{E}_q[\nabla_x \nabla_x \log p(x,y)] \end{bmatrix} \end{aligned}$$

Now since $\eta = C^{-1} g$ we have

$$\begin{aligned} \begin{bmatrix} Pm \\ P \end{bmatrix} := \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} &= \begin{bmatrix} 1 & -\mathbb{E}_{q_\phi}[x] \\ 0 & -\frac{1}{2} \end{bmatrix}^{-1} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \\ \Rightarrow \eta_2 = P &= -2g_2 = \mathbb{E}_q[\nabla_x \nabla_x \log p(x,y)] \\ \eta_1 = Pm &= g_1 + P^{-1}\mathbb{E}_q[x] = \mathbb{E}_q[\nabla_x \log p(x,y)] + P^{-1}\mathbb{E}_q[x] \end{aligned}$$

where Pm and P are the natural parameters (mean times precision and precision) of the approximation. Thus the quantities we need to stochastically approximate are

$$\begin{aligned} a &:= \mathbb{E}_q[\nabla_x \log p(x,y)] \\ H &:= \mathbb{E}_q[\nabla_x \nabla_x \log p(x,y)] \\ z &:= \mathbb{E}_q[x] \end{aligned}$$

so we have $P = H$ and $m = P^{-1}a + z$.

5.A.3 Connection to Efficient Importance Sampling

It is worth pointing out the connection between fixed-form variational Bayes and Richard and Zhang's (2007) *Efficient Importance Sampling* (EIS) algorithm. Although these authors take a different perspective (that of importance sampling) their goal of approximating the intractable posterior distribution with a more convenient distribution is identical to the goal of variational Bayes. Specifically, Richard and Zhang (2007) choose their posterior approximation to minimize the variance of the log-weights of the resulting importance sampler. This leads to an optimization problem obeying a similar fixed-point condition as in (5.9), but with the expectation taken over $p(x|y)$ instead of $q(x)$. Since sampling from $p(x|y)$ directly is not possible, they evaluate this expectation by sampling from $q(x)$ and weighting the samples using importance sampling. In practice however, these 'weights' are often kept fixed to one during the optimization process in order to improve the stability of the algorithm. When all weights are fixed to one, Richard and Zhang's (2007) fixed-point condition becomes identical to that of (5.9) and the algorithm is in fact fitting a variational posterior approximation.

The connection between EIS and variational Bayes seems to have gone unnoticed until now, but it has some important consequences. It is for example well known (e.g. Minka, 2005; Nickisch and Rasmussen, 2008; Turner et al., 2008) that the tails of variational posterior approximations tend to be thinner than those of the actual posterior unless the approximation is extremely close, which means that using EIS with the importance-weights fixed to one is not to be recommended for general applications: In case the posterior approximation is nearly exact, one might as well use it directly instead of using it to form another approximation using importance sampling. In cases where the approximation is not very close, the resulting importance sampling algorithm is likely to suffer from infinite variance problems. The literature on variational Bayes offers some help with these problems. Specifically, de Freitas et al. (2001) propose a number of ways in which variational approximations can be combined with Monte Carlo methods, while guarding for the aforementioned problems.

Much of the recent literature (e.g. Teh et al., 2007; Honkela et al., 2010) has focused on the computational and algorithmic aspects of fitting variational posterior approximations,

and this work might also be useful in the context of importance sampling. Algorithmically, the 'sequential EIS' approach of Richard and Zhang (2007) is most similar to the non-conjugate VMP algorithm of Knowles and Minka (2011). As these authors discuss, such an algorithm is not guaranteed to converge, and they present some tricks that might be used to improve convergence in some difficult cases.

The algorithm presented in this chapter for fitting variational approximations is provably convergent, as discussed in Section 5.4. Furthermore, Sections 5.5 and 5.6 present multiple new strategies for variance reduction and computational speed-up that might also be useful for importance sampling. In this chapter we will not pursue the application of importance sampling any further, but exploring these connections more fully is a promising direction for future work.

5.A.4 Choosing an estimator

As discussed in Section 5.4, the particular estimator used in our stochastic approximation is not the most obvious choice, but it seems to provide a lower variance approximation than other choices. In this section we consider three different MC estimators for approximating (5.9) to see why this might be the case.

The first separately approximates the two integrals and then calculates the ratio:

$$\hat{\eta}_1 = \left(\frac{1}{S} \sum_r \tilde{T}(x_r)' \tilde{T}(x_r) \right)^{-1} \frac{1}{S} \sum_s \tilde{T}(x_s)' \log p(x_s, y), \quad x_r, x_s \sim_{iid} q(x), \tag{5.40}$$

with S the number of Monte Carlo samples. The second approximates both integrals using the same samples from q

$$\hat{\eta}_2 = \left(\frac{1}{S} \sum_s \tilde{T}(x_s)' \tilde{T}(x_s) \right)^{-1} \frac{1}{S} \sum_s \tilde{T}(x_s)' \log p(x_s, y), \quad x_s \sim_{iid} q(x). \tag{5.41}$$

Note that only this estimator is directly analogous to the linear regression estimator. The third estimator is available only when the first expectation is available analytically:

$$\hat{\eta}_a = \mathbb{E}_q \left[\tilde{T}(x)' \tilde{T}(x) \right]^{-1} \frac{1}{S} \sum_s \tilde{T}(x_s)' \log p(x_s, y), \quad x_s \sim_{iid} q(x). \tag{5.42}$$

We wish to understand the bias/variance tradeoff inherent in each of these estimators. To keep notation manageable consider the case with only $k = 1$ sufficient statistic[1] and let

$$a(x) = \tilde{T}(x)'\tilde{T}(x) = \tilde{T}(x)^2 \tag{5.43}$$

$$b(x) = \tilde{T}(x) \log p(x, y) \tag{5.44}$$

We can now write the three estimators of η more concisely as

$$\hat{\eta}_1 = \frac{\frac{1}{S}\sum_r b(x_r)}{\frac{1}{S}\sum_s a(x_s)}, \qquad x_r, x_s \sim_{iid} q(x) \tag{5.45}$$

$$\hat{\eta}_2 = \frac{\frac{1}{S}\sum_s b(x_s)}{\frac{1}{S}\sum_s a(x_s)}, \qquad x_s \sim_{iid} q(x) \tag{5.46}$$

$$\hat{\eta}_a = \frac{\frac{1}{S}\sum_s b(x_s)}{\mathbb{E}[a]}, \qquad x_s \sim_{iid} q(x) \tag{5.47}$$

Using a simple Taylor series argument it is straightforward to approximate the bias and variance of these estimators. We first consider the bias. Consider the multivariate Taylor expansion of $f : \mathbb{R}^K \to \mathbb{R}$ around the point $\bar{y} \in \mathbb{R}^K$:

$$f(y) \approx f(\bar{y}) + (y - \bar{y})' f'(\bar{y}) + \frac{1}{2}\text{tr}((y - \bar{y})(y - \bar{y})'\nabla^2 f(\bar{y})) \tag{5.48}$$

From this we can derive expressions for the expectation of $f(y)$:

$$\mathbb{E}[f] \approx f(\bar{y}) + \frac{1}{2}\text{tr}(\text{Cov}(y) f''(\bar{y})) \tag{5.49}$$

where we have chosen to perform the Taylor expansion around the mean $\bar{y} = \mathbb{E}[y]$. For the first estimator let $y = \frac{1}{S}\sum_s a(x_s)$ and $f(y) = 1/y$, then we find

$$\mathbb{E}[\hat{\eta}_1] = \mathbb{E}\left[\left(\frac{1}{S}\sum_s a(x_s)\right)^{-1}\right]\mathbb{E}[b] \tag{5.50}$$

$$\approx \left(\frac{1}{\mathbb{E}[a]} + \frac{\text{Var}(a)}{S\mathbb{E}[a]^3}\right)\mathbb{E}[b] \tag{5.51}$$

$$= \mathbb{E}[\eta] + \frac{\text{Var}(a)\mathbb{E}[b]}{S\mathbb{E}[a]^3} \tag{5.52}$$

[1] These results extend in a straightforward manner to the case where $k > 1$

since $\mathrm{Var}(y) = \mathrm{Var}(a)/S$. We see that the bias term depends on the ratio $\mathrm{Var}(a)/\mathbb{E}[a]^2$, i.e. the spread of the distribution of a relative to its magnitude.

Now for the second estimator let

$$y = \begin{bmatrix} \frac{1}{S}\sum_s a(x_s) \\ \frac{1}{S}\sum_s b(x_s) \end{bmatrix} \tag{5.53}$$

so that $\eta_2 = f(y) = \frac{y_2}{y_1}$. Note that $\mathrm{Cov}(y) = \frac{1}{S}\mathrm{Cov}([a, b]')$ and

$$\nabla^2 f(y) = \begin{bmatrix} \frac{2y_2}{y_1^3} & -\frac{1}{y_1^2} \\ -\frac{1}{y_1^2} & 0 \end{bmatrix} \tag{5.54}$$

Putting everything together we have

$$\mathbb{E}[\hat{\eta}_2] \approx \eta + \frac{\mathrm{Var}(a)\mathbb{E}b}{S\mathbb{E}[a]^3} - \frac{\mathrm{Cov}(a, b)}{S\mathbb{E}[a]^2} \tag{5.55}$$

Note that we recover the expression for $\mathbb{E}\hat{\eta}_1$ if $\mathrm{Cov}(a, b) = 0$, which makes sense because if we use different randomness for calculating $\mathbb{E}[a]$ and $\mathbb{E}[b]$ then a, b have 0 covariance in our MC estimate. Finally the analytic estimator is unbiased:

$$\mathbb{E}\hat{\eta}_a = \eta \tag{5.56}$$

We now turn to the variances. The analytic estimator is a standard MC estimator with variance

$$\mathrm{Var}(\hat{\eta}_a) = \frac{\mathrm{Var}(b)}{S\mathbb{E}[a]^2} \tag{5.57}$$

Consider only the linear terms of the Taylor expansion:

$$f(y) \approx f(\bar{y}) + (y - \bar{y})' f'(\bar{y}) \tag{5.58}$$

Substituting this into the formula for variance gives

$$\begin{aligned}\operatorname{Var}[f(y)] &= \mathbb{E}[(f(y) - \mathbb{E}[f(y)])(f(y) - \mathbb{E}[f(y)])'] && (5.59)\\ &\approx \mathbb{E}[f'(\bar{y})'(y - \bar{y})(y - \bar{y})' f'(\bar{y})] && (5.60)\\ &= f'(\bar{y})' \operatorname{Var}(y) f'(\bar{y}) && (5.61)\end{aligned}$$

We will calculate the variance of the second estimator and derive the variance of the first estimator from this. Again let y be as in (5.53). Note that $\operatorname{Var}(y) = \operatorname{Cov}(a, b)/S$. We find

$$\operatorname{Var} \hat{\eta}_2 \approx \frac{1}{S} \left(\frac{\mathbb{E}[b]^2 \operatorname{Var} a}{\mathbb{E}[a]^4} - 2 \frac{\mathbb{E}[b] \operatorname{Cov}(a, b)}{\mathbb{E}[a]^3} + \frac{\operatorname{Var} b}{\mathbb{E}[a]^2} \right) \tag{5.62}$$

The final term is equal to that for the analytic estimator. The second term is not present in the variance of the first estimator, since then a and b have no covariance under the sampling distribution, i.e.

$$\operatorname{Var} \hat{\eta}_1 \approx \frac{1}{S} \left(\frac{\mathbb{E}[b]^2 \operatorname{Var} a}{\mathbb{E}[a]^4} + \frac{\operatorname{Var} b}{\mathbb{E}[a]^2} \right) \tag{5.63}$$

The first term is always positive, suggesting that $\hat{\eta}_1$ is dominated by the analytic estimator.

Summarizing these derivations, we have

$$\begin{aligned}\operatorname{bias}(\hat{\eta}_1) &\approx \frac{\operatorname{Var}(a)\mathbb{E}[b]}{S\mathbb{E}[a]^3} \\ \operatorname{bias}(\hat{\eta}_2) &\approx \frac{\operatorname{Var}(a)\mathbb{E}[b]}{S\mathbb{E}[a]^3} - \frac{\operatorname{Cov}(a, b)}{S\mathbb{E}[a]^2}. \end{aligned} \tag{5.64}$$

Note that the first term is shared, but the first estimator does not have the covariance term as a result of the independent sampling in approximating the numerator and denominator. In contrast $\hat{\eta}_a$ is unbiased. Now consider the variances

$$\operatorname{Var}(\hat{\eta}_1) \approx \frac{1}{S} \left(\frac{\mathbb{E}[b]^2 \operatorname{Var}(a)}{\mathbb{E}[a]^4} + \frac{\operatorname{Var}(b)}{\mathbb{E}[a]^2} \right) \tag{5.65}$$

$$\operatorname{Var}(\hat{\eta}_2) \approx \frac{1}{S} \left(\frac{\mathbb{E}[b]^2 \operatorname{Var}(a)}{\mathbb{E}[a]^4} - 2 \frac{\mathbb{E}[b] \operatorname{Cov}(a, b)}{\mathbb{E}[a]^3} + \frac{\operatorname{Var}(b)}{\mathbb{E}[a]^2} \right) \tag{5.66}$$

$$\operatorname{Var}(\hat{\eta}_a) = \frac{\operatorname{Var}(b)}{S\mathbb{E}[a]^2} \tag{5.67}$$

All three estimators have the same final term (the variance of the "analytic" estimator). Again the second estimator has an additional term resulting from the covariance between a and b which we find is typically beneficial in that it results in the variance of $\hat{\eta}$ being significantly smaller. It is worth recalling that the mean squared error (MSE) of an estimator is given by

$$\mathbb{E}[(\eta - \hat{\eta})^2] = \mathrm{Var}(\hat{\eta}) + \mathrm{bias}(\hat{\eta})^2. \tag{5.68}$$

Since both the variance and bias are $O(1/S)$, the variance contribution to the MSE is $O(1/S)$ whereas the bias contribution is $O(1/S^2)$, so the variance is actually a greater problem than the bias. From these expressions it is still not immediately obvious which estimator we should use. However, consider the case when the target distribution p is in the same exponential family as q, i.e. when $\log p(x, y) = \tilde{T}(x)\lambda$. It is then straightforward to show that

$$\mathrm{bias}(\hat{\eta}_1) \approx \frac{\lambda \,\mathrm{Var}(\tilde{T}^2)}{S\mathbb{E}[\tilde{T}^2]^2}, \qquad \mathrm{Var}(\hat{\eta}_1) \approx 2\frac{\lambda^2 \,\mathrm{Var}(\tilde{T}^2)}{S\mathbb{E}[\tilde{T}^2]^2} \tag{5.69}$$

$$\mathrm{bias}(\hat{\eta}_2) \approx 0, \qquad \mathrm{Var}(\hat{\eta}_2) \approx 0 \tag{5.70}$$

$$\mathrm{bias}(\hat{\eta}_a) = 0, \qquad \mathrm{Var}(\hat{\eta}_a) = \frac{\lambda^2 \,\mathrm{Var}(\tilde{T}^2)}{S\mathbb{E}[\tilde{T}^2]^2} \tag{5.71}$$

We see that in this case for $\hat{\eta}_2$ the positive and negative contributions to both the bias and variance cancel. While this result will not hold exactly for cases of interest, it suggests that for exponential families which are capable of approximating p reasonably well, $\hat{\eta}_2$ should perform significantly better than $\hat{\eta}_1$ or even $\hat{\eta}_a$. If q and p are of the same exponential family, it is actually possible to see that $\hat{\eta}_2$ will in fact give the exact solution in $k + 1$ samples (with k the number of sufficient statistics), while the other estimators have non-vanishing variance for a finite number of samples. This means that the approximate equality in (5.70) can be replaced by exact equality. Using $k+1$ samples $x_i, i = 1, ..., k+1$, assumed to be unique (which holds almost surely for continuous distributions q), we have

$$\hat{\eta}_2 = \left(\sum_{i=1}^{k+1} \tilde{T}(x_i)'\tilde{T}(x_i)\right)^{-1} \sum_{i=1}^{k+1} \tilde{T}(x_i)'\tilde{T}(x_i)\lambda = \lambda \tag{5.72}$$

That is, the algorithm has recovered $p(x, y)$ exactly with probability one. If we assume we know how to normalize q, this means we also have $p(x|y)$ exactly in this case. Note that we recover the exact answer here because the $p(x, y)$ function evaluations are in themselves *noise free*, so the regression analogy really corresponds to a noise free regression.

It is instructive to consider a toy example: fitting an exponential distribution $p(x) = \lambda e^{-\lambda x}$, about the simplest possible demonstration of the exact fitting phenomenon shown in (5.72). We assume that we are unaware that p happens to be normalized. Our variational approximation has $\tilde{T} = [1, x]'$ and rate η, i.e. $q(x) = \eta e^{-\eta x}$. Note that this is an example where it is straightforward to calculate

$$\mathbb{E}_q[\tilde{T}(x)'\tilde{T}(x)] = \begin{bmatrix} 1 & -\eta^{-1} \\ -\eta^{-1} & \eta^{-2} \end{bmatrix}$$

We test the three estimators in (5.40), (5.41) and (5.42) when the true exponential rate is $\lambda = 1.5$, and sampling from the optimal q distribution with $\eta = 1.5$. The results confirm that $\hat{\eta}_2$ finds the exact rate using just $S = 2$ MC samples, as predicted by (5.72). We would expect $\hat{\eta}_a$ to be unbiased, and this is borne out by the results shown in Figure 5.8. The estimator $\hat{\eta}_1$ has both poor bias and such large variance that it often gives an invalid negative rate if fewer than 10 MC samples are used. While this is clearly a very simple example it hopefully emphasizes the potential benefit to be gained from using estimators related to $\hat{\eta}_2$.

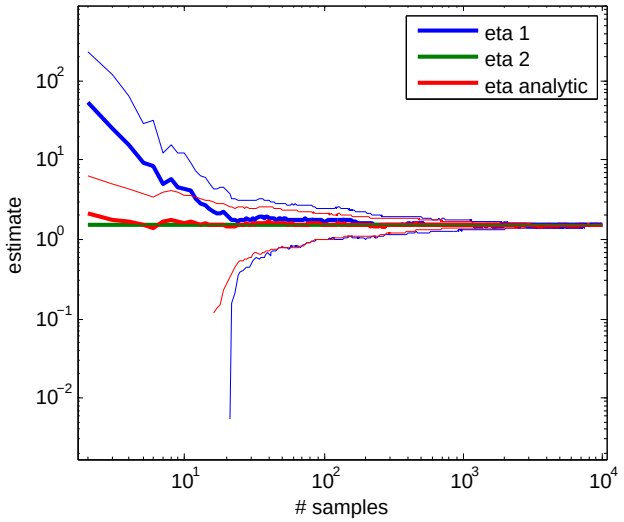


Figure 5.8: Comparison of three estimators for fitting a variational posterior q to a simple exponential distribution p. 50 repeats were used to estimate the mean and variance of the estimator: the thick line shows the mean and the thin lines show $\pm$ one standard deviation. The x-axis indicator the number of MC samples, S, used. As expected in this case $\hat{\eta}_2$ gives the correct solution of 1.5 using $S \geq 2$ samples.

Chapter 6

A preference ranking model for making product recommendations

Joined work with Ulrich Paquet and Thore Graepel. Most of the work towards this chapter was done during an internship at Microsoft Research, Cambridge, UK.

6.1 Introduction

With the advent of the internet, online product recommendation systems have become big business. Companies like Amazon, Netflix, and Microsoft all depend on online product recommendations to drive a significant part of their sales: We all recognize Amazon's "customers who bought x also bought y" advertisements that are shown after a customer makes a purchase. Other well-known but less obvious examples of recommendation systems include the "you may know these people: " messages on Facebook and LinkedIn, the personalized news feed of Google news, or personalized online radio channels such as Last.FM. These recommender systems are very interesting from a research perspective: A user's purchasing history, demographics, and interactions with a website provide a small but diverse data set from which we can learn about that person's tastes and preferences. Building a statistical model that efficiently takes this information into account and that can use it to make useful recommendations is a great challenge. We can recommend Jannach et al. (2010) for a more elaborate introduction on the topic.

Research on recommender systems received a major boost when the Netflix prize was

held between 2006 and 2009: a million dollar contest for developing a model to accurately predict the ratings Netflix' customers give to the movies they watch. Besides providing a monetary incentive to work on this problem, the contest also provided a large data set and benchmark problem for the research community. A comprehensive summary of the contest and a discussion of its impact on research can be found in Paterek (2012).

During the Netflix prize period recommender systems research was mostly focused on explicit feedback in the form of ratings. Later research also explored using implicit binary feedback such as observed purchases or clicks. Often, however, real world data sources lie in between these extremes. Explicit ratings of items are rare and hard to obtain, but often our information is richer than a simple binary signal such as click/non-click. For example, users may express relative value judgments in comparing two different products, or they may provide a partial preference ranking over available items. Such rankings can be explicit such as lists of favorite songs, or inferred from implicit information such as the play counts for these songs. To make efficient use of such information, we propose in this chapter a new bilinear factor model that maps latent user preferences to observed pairwise comparisons or rankings over items. Since feedback is relative to other items, this modeling approach is more robust than models of user preferences on an absolute scale. Yet it makes more efficient use of available data compared to methods that only allow for binary feedback. Research also shows that people find it easier to formulate their preferences in such a relative way (Jaeger et al., 2008, among others). An additional advantage is that modeling preference rankings directly leads to a ranking of items to be recommended to users, which is the end goal of most recommendation systems.

We present the new user preference model in Section 6.2. The basis of our model is formed by a bilinear factor model, similar to the type of models that were successful in the Netflix prize. Rather than relying upon absolute feedback, however, we couple this model with a likelihood function that takes into account the relative feedback provided by a user. Performing inference with this model is discussed in Section 6.3. Here we develop both an efficient Gibbs sampling algorithm that can be parallelized to allow for very large data sets, and a deterministic approximation method that is a hybrid of two different types of approximation algorithms: Expectation Propagation and Variational Bayes Expectation Maximization. In Section 6.4 we present two applications of our model. The first

application is recommendation of games on Microsoft's Xbox Live Marketplace. The 35 million users of this service do not explicitly state their preferences with respect to the games in the Marketplace catalog, but we do observe their interaction with these games. We interpret their playing time for each game as revealing a preference ranking over the games the user owns, and we use this to allow our model to recommend games that the user does not yet own. In our second example we apply our model to the ranking of sushi items. We use the publicly available data set of Kamishima and Akaho (2006) in order to be able to compare our approach to competing methods. This data set describes the stated preferences over different types of sushi for 5,000 Japanese survey correspondents. The new model outperforms the previous state of the art on this problem. In Section 6.5 we discuss the potential of the model to guide a more active learning strategy, where we actively and selectively ask the user for relative feedback on different items. Finally, Section 6.6 concludes.

6.2 The Model

The classic problem adressed by recommender systems is to predict the level of satisfaction or *utility* a given *user* i will receive from purchasing a given *product* j. One of the most popular approaches taken to model this problem is to assign a vector of latent variables (also called features or factors) to both the users and the items, and to take the utility for each user/item combination to be the inner product of these user and item vectors. Inferring the latent user and item factors then comes down to constructing a low rank approximation to the user-item-utility matrix. For this reason this class of models is known as *matrix factorization* in the recommender systems literature. Factorizing a matrix is closely related to calculating its singular value decomposition, so this class of methods is also sometimes referred to as SVD (singular value decomposition). Koren et al. (2009) provide an introduction to this class of methods. The model presented here also falls in the class of matrix factorization methods and is most closely related to the Bayesian approaches of Stern et al. (2009) and Paquet et al. (2011).

In our model, as in other matrix factorization models, each of the N users and M items are represented with low-rank factors: user i with a latent $K \times 1$ feature vector $\boldsymbol{u}_i$,

and item j with a latent $K \times 1$ feature vector $\boldsymbol{v}_j$. These feature vectors represent the latent characteristics of the users and items. As some items are predominantly more popular than others, each item is also given a univariate bias parameter b_j.

The user and item features are combined into a latent *score* or *utility* $s_{i,j}$,

$$s_{i,j} \sim N(\boldsymbol{u}_i' \boldsymbol{v}_j + b_j, 1) \,, \tag{6.1}$$

which represents how much user i likes item j.

The relative ordering of a set of scores determines a user's preference of one item over the next. If user i prefers items $j_1 \succ j_2$ ($\succ$ meaning "is preferred to"), we require that $s_{i,j_1} > s_{i,j_2}$. We interpret our relative feedback data as observations on a number of these pairwise comparisons between the different latent scores $s_{i,j}$. For each user, we denote these comparisons by $\boldsymbol{C}^i$, a sparsely filled matrix of dimension $M \times M$, with elements

$$\begin{aligned} c^i_{j,j'} &= 1 && \text{if } s_{i,j} > s_{i,j'} \\ c^i_{j,j'} &= -1 && \text{if } s_{i,j} < s_{i,j'} \\ c^i_{j,j'} &= \text{empty} && \text{if unknown} \,. \end{aligned} \tag{6.2}$$

These observed preferences can be explicitly provided by the user in the form of a ranking or a number of pairwise preference statements, or they can be inferred from the behavior of the user, for example by ordering the time spent interacting with different items. Section 6.4 gives example applications, where we use the model to learn stated preferences over sushi items and inferred preferences over Xbox games. Importantly, if the preferences are expressed as a transitive ranking of items, one might always find a set $\{s_{i,j}\}$ that is consistent over the ranked items j, and hence consistent with user i's observations $\boldsymbol{C}^i$. The data likelihood is then given by

$$p(\boldsymbol{C}|\boldsymbol{S}) = \prod_{i=1}^{N} \prod_{(j,j') \in c^i} \mathbb{I}\Big[c^i_{j,j'}(s_{i,j} - s_{i,j'}) > 0\Big] \,, \tag{6.3}$$

where $\mathbb{I}[\text{true}] = 1$ and $\mathbb{I}[\text{false}] = 0$.

We assign independent normal priors to the user and item vectors, as well as the bias

terms, with

$$\begin{aligned} \boldsymbol{u}_i &\sim N(\beta, \mathrm{B}) \,, && \text{for } i = 1, \ldots, N \\ \boldsymbol{v}_j &\sim N(\gamma, \Gamma) \,, && \text{for } j = 1, \ldots, M \\ b_j &\sim N(0, \psi) \,, && \text{for } j = 1, \ldots, M \,. \end{aligned} \tag{6.4}$$

Although very popular in the machine learning literature, this type of factor model has not been discussed much in the marketing or economics literature. The main difference between this type of model and the more commonly used discrete choice models in econometrics (e.g. the probit and logit variants) is that the model is symmetric: both the preferences of the users and the characteristics of the items are latent, and have to be inferred from the data, whereas the more common approach in economics is to infer only the user's preferences, while the item's characteristics are observed and thus fixed.

In order to relate the model presented here to the more well-studied choice models of econometrics, it is instructive to examine the distribution of the utilities $\boldsymbol{S}$, conditional on the item characteristics $\boldsymbol{V}, \boldsymbol{b}$. For convenience, we collect the relevant terms from (6.1) and (6.4) above:

$$s_{i,j} = b_j + \boldsymbol{v}_j' \boldsymbol{u}_i + \epsilon_{i,j}, \text{ where } \epsilon_{i,j} \sim N(0,1) \tag{6.5}$$

$$\boldsymbol{u}_i \sim N(\beta, \mathrm{B}). \tag{6.6}$$

If we condition on the $\boldsymbol{v}_j, b_j$ terms, the model above can be recognized as a relatively standard mixed probit model: (6.5) is the standard probit model, where $\boldsymbol{v}_j$ can be interpreted as a vector of explanatory variables, and where b_j is an item-dependent intercept. Equation 6.6 then represents the multivariate normal mixing distribution over the subject parameters $\boldsymbol{u}_i$. If we would replace the standard normal distribution on $\epsilon_{i,j}$ with a type-1 extreme value distribution, we would recover the popular mixed logit specification as discussed by Revelt and Train (1998) and many others in the marketing literature.

Alternatively, we can integrate out the user parameters $\boldsymbol{u}_i$ over their prior distribution to obtain the conditional distribution of the $M \times 1$ vector of utilities $\boldsymbol{S}_i$ for user i given

$\boldsymbol{V}, \boldsymbol{b}$:

$$\begin{aligned} \boldsymbol{S}_i &= \boldsymbol{b} + \boldsymbol{V}\beta + \boldsymbol{\eta}_i, \text{ where} \\ \boldsymbol{\eta}_i|\boldsymbol{V}, \boldsymbol{b}. &\sim N(0, \boldsymbol{V}\,\mathrm{B}\,\boldsymbol{V}' + \mathrm{I}_M), \end{aligned} \tag{6.7}$$

where the item features are stacked into an $M \times N$ matrix $\boldsymbol{V}$, and the biases into an $M \times 1$ vector $\boldsymbol{b}$. Conditional on $\boldsymbol{V}, \boldsymbol{b}$, equation 6.7 can then be recognized as a standard multinomial probit model, with $\boldsymbol{V}$ the matrix of explanatory variables and β their coefficients. The covariance matrix of the utility errors $\boldsymbol{\eta}_i$ in this model has a low-rank factor structure, as is common for this type of models (see e.g. Yai et al., 1997).

The comparison to mixed probit and multinomial probit models allows us to better understand the preference model presented here. The difference between our model and these standard models is that in our case the "explanatory variables" $\boldsymbol{v}_j$ are latent instead of observed, but our model still shares many of the properties of the more standard probit models. In particular, our model does not have the independence of irrelevant alternatives (IIA) property. For additional analysis of this type of discrete choice models see Train (2003). Furthermore, note that although our model is very flexible in modeling the characteristics of the items, the user's mean utilities are still linear functions of the continuous attributes of the items. This means that the model cannot represent arbitrary monotone preferences, as is possible with a nonparametric approach such as the one used by Geweke (2012).

For the applications in this chapter, we set the prior means of $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ to zero for all i, j, and their prior covariance matrices to $\pi\,\mathrm{I}_K$. This leaves us with two scalar prior parameters π and ψ. These parameters are set manually for our examples, but can also be inferred from the data, as explained in Paquet et al. (2011). For brevity we do not consider this here. We set the feature dimensionality K to a default value of 20. As with other applications of matrix factorization we find that the results are generally better for large values of K than for small values, but that the improvement levels off as K is increased. Since the computational work demanded continues to grow at least linearly in K, a value between 10 and 100 is recommended for most applications.

A graphical representation of the full Bayesian network of our model is given in Figure

6.1.

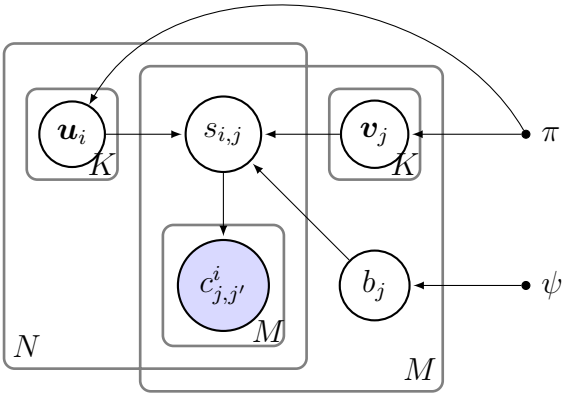


Figure 6.1: The proposed Bayesian factor model for learning preference rankings

The next section discusses how to obtain the posterior distribution of the parameters $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}$ conditional on the observed preferences.

6.3 Bayesian Inference

The model proposed in the last section does not admit a closed form posterior distribution for the parameters $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}$ that we need in order to make recommendations. We therefore propose two strategies for approximating this posterior distribution: a Gibbs sampling algorithm to generate samples from the posterior distribution, and a deterministic approximation algorithm that minimizes local divergence measures between the posterior distribution and a factored approximation. The performance and scaling of these two algorithms is evaluated on real world data in Section 6.4.

6.3.1 Gibbs Sampling

We can generate correlated samples from the posterior distribution using a Gibbs sampling algorithm that iteratively samples from the conditional distributions $p(\boldsymbol{u}_i|\boldsymbol{V}, \boldsymbol{b}, \boldsymbol{S})$, $p(\boldsymbol{v}_j, b_j|\boldsymbol{U}, \boldsymbol{S})$ and $p(s_{i,j}|\boldsymbol{s}_{i,/j}, \boldsymbol{u}_i, \boldsymbol{v}_j, b_j, \boldsymbol{C}^i)$, for all i, j, where $\boldsymbol{s}_{i,/j}$ denotes the vector of all scores for user i excluding the j-th. The conditional distributions $p(\boldsymbol{u}_i|\boldsymbol{V}, \boldsymbol{b}, \boldsymbol{S})$ and $p(\boldsymbol{v}_j, b_j|\boldsymbol{U}, \boldsymbol{S})$ are Gaussian and have been used by several authors before. See Paquet

et al. (2011) for their precise form. The full conditionals $p(s_{i,j}|\boldsymbol{s}_{i,/j}, \boldsymbol{u}_i, \boldsymbol{v}_j, b_j, \boldsymbol{C}^i)$ are univariate truncated normal, which follows from the Gaussian conditional prior (6.1) and the truncating likelihood (6.3).

The $s_{i,j}$ are most efficiently updated by first performing a forward pass over all scores for a given user i, sampling the scores $s_{i,j}$ in the order of the observed preference ranking, followed by a backward pass sampling in the reversed order. (Observe that we do not have to sample those $s_{i,j}$ for which we have no feedback.) We find that this updating schedule does a good job of sampling the relative differences between the scores, but that it is slow in changing the overall level of the scores. To further improve the mixing of the Gibbs sampling algorithm we therefore follow the forward and backward pass by an additional Monte Carlo step that simultaneously shifts all scores for a given user, while leaving the stationary distribution of the Markov chain invariant. The update equation for this step is given as

$$\begin{aligned} s_{i,j} &\leftarrow s_{i,j} + d_i, \text{ with } j = 1, \ldots, L_i \\ d_i &\sim N(\bar{f}_i - \bar{s}_i, L_i^{-1}), \end{aligned} \tag{6.8}$$

with L_i the number of items for which user i has provided feedback, $\bar{s}_i$ the mean sampled score for those items, and where $\bar{f}_i$ the mean of the predicted scores for those items:

$$\bar{f}_i = \frac{1}{L_i} \sum_j \boldsymbol{u}_i' \boldsymbol{v}_j + b_j.$$

Since sampling the scores using the steps outlined here is relatively quick compared to sampling $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{b}$, we find that the most efficient implementation of Gibbs sampling resamples $\boldsymbol{S}$ multiple times per iteration.

6.3.2 Hybrid VB/EP posterior approximation

The Gibbs sampling algorithm outlined in the last section is relatively fast and can be applied at quite a large scale, however for very large data sets a deterministic approximation of the posterior distribution may provide a better tradeoff between accuracy and computational cost. An additional advantage of such a deterministic approximation is that it con-

verges to a single mode of the posterior distributions and that it can be represented more compactly than the Gibbs sampling approximation, which reduces the computational cost of generating new recommendations for users given the posterior approximation. We develop a new algorithm to construct such a deterministic approximation, making use of Expectation Propagation (EP) (Minka, 2001a) for the ranking likelihood (6.3) and Variational Bayes for the latent factor model. EP provides an excellent approximation for the unimodal posterior resulting from the truncated Gaussian in (6.3), whereas Variational Bayes picks and locally approximates the posterior mode resulting from the product factor in (6.1) as described in Lim and Teh (2007) among others.

We approximate the posterior distribution $p(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}|\boldsymbol{C})$ with a fully factorized Gaussian

$$q(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}) = \prod_{i,k} q(u_{i,k}) \prod_{j,k} q(v_{j,k}) \prod_{j} q(b_j), \tag{6.9}$$

although our inference algorithm can also be used with a Gaussian approximation that preserves some of these dependencies, e.g. $q(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}) = \prod_j q(\boldsymbol{v}_j, b_j) \prod_i q(\boldsymbol{u}_i)$. In order to optimize this approximate posterior distribution we first approximate the likelihood term $p(\boldsymbol{C}|\boldsymbol{S})$ by a product of univariate Gaussian density functions in $s_{i,j}$, i.e.

$$q(\boldsymbol{C}|\boldsymbol{S}) = \prod_{i,j} \phi(s_{i,j}; \mu_{i,j}, \sigma^2_{i,j}), \tag{6.10}$$

with $\phi(\cdot)$ a Gaussian pdf, which we initialize to have infinite variance. The parameters of the likelihood approximation, $\mu_{i,j}$ and $\sigma^2_{i,j}$ are then set using EP. This EP step starts with the construction of a Gaussian 'pseudo prior' on the $s_{i,j}$:

$$\begin{aligned} q(s_{i,j}) &\propto \phi(s_{i,j}; \mu^*_{i,j}, \sigma^{2*}_{i,j}) / \phi(s_{i,j}; \mu_{i,j}, \sigma^2_{i,j}), \\ \mu^*_{i,j} &= \mathbb{E}_q s_{i,j} \text{ and } \sigma^{2*}_{i,j} = \mathrm{Var}_q\, s_{i,j}, \end{aligned} \tag{6.11}$$

where the expectations $\mathbb{E}_q$ are taken with respect to the current posterior approximation $q(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b})$. Under the pseudo prior $q(s_{i,j})$, the algorithm for determining the approximate likelihood terms $\phi(s_{i,j}; \mu_{i,j}, \sigma^2_{i,j})$ is a special case of algorithm 1 as presented in Chapter 3, where the pairwise restrictions of (6.3) form the non-Gaussian likelihood terms. We refer the reader to that chapter for additional information. This approximation step is also very

similar to the approach taken by Dangauthier et al. (2008), whose work can be consulted for additional details.

After the EP step, we optimize the posterior approximation using Variational Bayes, i.e. we choose our posterior approximation to solve

$$\max_{q(\boldsymbol{U},\boldsymbol{V},\boldsymbol{b})} \mathbb{E}_q \log q(\boldsymbol{C}|\boldsymbol{S}) p(\boldsymbol{U},\boldsymbol{V},\boldsymbol{b}) - \log q(\boldsymbol{U},\boldsymbol{V},\boldsymbol{b}) \,, \tag{6.12}$$

under the restriction that $q(\boldsymbol{U},\boldsymbol{V},\boldsymbol{b})$ is of the form specified in (6.9). This constrained optimization problem can be solved efficiently using the Variational Bayes Expectation Maximization (VBEM) algorithm. Since our pseudo likelihood terms $q(\boldsymbol{C}|\boldsymbol{S})$ are now i.i.d. Gaussian, the update equations for the optimization are identical to those used in earlier factor models with absolute feedback (see Lim and Teh (2007) for their exact form). This approach shares some characteristics with other works using variational approximations in which the $s_{i,j}$ are assumed unobserved (e.g. Paquet et al., 2011), however, note that with our approach the expectations with respect to $s_{i,j}$ in (6.12) follow from $q(\boldsymbol{U},\boldsymbol{V},\boldsymbol{b})$ and the model in (6.1) rather than from a separate posterior approximation on the $s_{i,j}$ as is more commonly used. By avoiding this explicit approximation of $p(\boldsymbol{S}|\boldsymbol{C})$, the posterior approximation $q(\boldsymbol{U},\boldsymbol{V},\boldsymbol{b})$ gains in accuracy without increasing computational cost. The VBEM and EP steps are repeated until convergence. For typical applications with a small to medium number of comparisons per user we find that around 50 iterations is usually sufficient.

6.3.3 Parallel Computation

For many real world applications of recommendation algorithms, both the number of users as well as the number of items is very large, necessitating the use of parallel computation to speed up inference. Both algorithms described above can be completely parallelized over users when updating $\boldsymbol{S}$ and $\boldsymbol{U}$, and over items when updating $\boldsymbol{V}$ and $\boldsymbol{b}$, which is an important advantage over the message passing algorithm used in Stern et al. (2009) for a similar problem. Since in our applications the number of items is relatively small compared to the number of users, we found it most efficient to distribute the users and their feedback over multiple threads. Within each thread, $\boldsymbol{S}$ and $\boldsymbol{U}$ can then be updated

without requiring any communication across threads. Every update of $\boldsymbol{V}$ and $\boldsymbol{b}$ then requires each thread to submit the sufficient statistics for the update of these variables and to receive the updated values. Since the number of items is relatively low this adds very little overhead and it allows us to speed up inference almost linearly with the number of available computate nodes. Using this strategy for spreading the computation across compute nodes we were able to perform inference very quickly on massive data sets.

6.4 Recommendation

We now present two real world applications of our model and compare its performance against alternative methods.

6.4.1 Xbox marketplace

The motivating application for the current work is the recommendation of games on Microsoft's Xbox Live Marketplace. The 35 million users of this service do not explicitly state their preferences with respect to the games in the Marketplace catalog, but we do observe their interaction with these games. Specifically we observe how much time each user spends playing each game. The old recommendation engine used for this service only used this data to infer which user owned which game and then interpreted ownership as positive feedback and non-ownership as negative feedback. This works quite well in predicting whether or not a user owns a game, however this is not the task that we are trying to solve: we want to generate good recommendations for games the user does not yet have. Specifically, a recommendation is good if a user is happy with the recommended game after buying it. The users do not provide this type of feedback explicitly, but if a user is happy with the recommended game we expect him or her to spend a lot of time playing it. Our statistical problem thus comes down to predicting the time a user spends playing a game, conditional on owning it. We could model this play time directly, but this is hard to do as different users can have very different behaviour: some users own many games and play multiple hours a day, while others own only a few and play only once a week. Furthermore, modeling the play time on an absolute scale is unnecessary since we only need a ranking of predicted play times to generate our recommendations.

A more robust approach is thus to model the time a user spends on a game relatively to how much the user spends on other games. We do this by constructing a ranking from the observed play time and we use this to infer user preferences using the model presented in section 6.2.

We evaluate our model for this task by collecting a data set containing all Xbox Live Marketplace users that own at least five games. Of the games each user owns we select two into a holdout set and we use the rest for model estimation. We then use the model to predict whether each user spends more time playing the first holdout game or the second holdout game. The model is able to pick the correct game in 72% of the cases, which is a small but meaningful improvement over a simple benchmark based on popularity.

In addition, we can check that our probabilistic model is appropriate by comparing the confidence of these predictions to the empirical fraction of correctly predicted games. Figure 6.2 shows that these two measures correspond quite closely, but that our more confident predictions are somewhat overconfident, which may be due to our approximation of the posterior distribution. Overall our model seems reasonably well calibrated and the probabilistic model works well in modeling the ranking of play time.

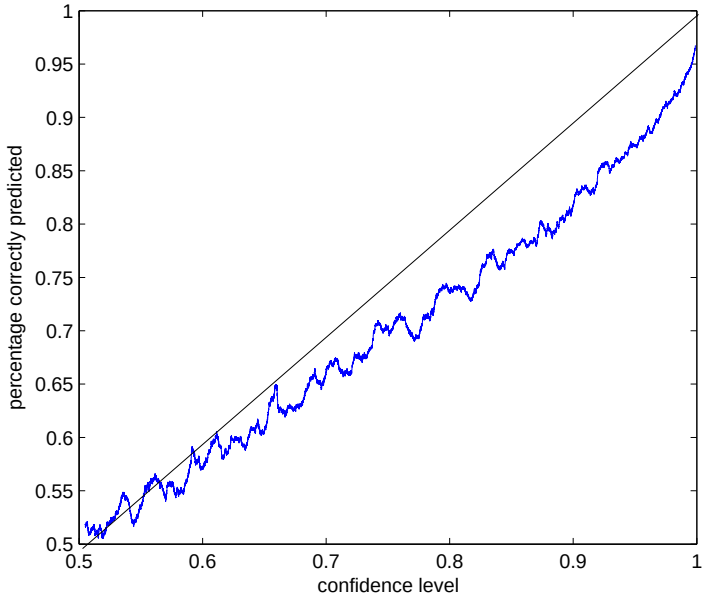


Figure 6.2: Xbox games: Posterior predicted probability versus correctly predicted fraction, obtained using the deterministic approximation

6.4.2 Learning Sushi preferences

In order to compare the new algorithms to existing methods we evaluate them on the publically available sushi preference data of Kamishima and Akaho (2006). This data set was generated by asking 5,000 Japanese survey correspondents to order a subset of 100 sushi types according to their preferences. Each correspondent provided two such ordered lists containing 10 different sushi types. Kamishima and Akaho (2006) evaluate their collaborative ranking approach by estimating their model on list 'B' and using the model to predict the order of list 'A'. They measure the performance of their method by the average Spearman correlation between the predicted and realized rankings. We use this measure to compare the performance of the new method to the 'Nantonac' algorithm of Kamishima and Akaho (2006), and also to compare our two inference algorithms against each other. Using this measure, we found that the maximum predictive accuracy was reached after about 1000 draws of the Gibbs sampler after a burn-in period of 100 draws, or after 50 iterations of the VB/EP algorithm. After that, additional draws or iterations no longer significantly changed the accuracy of the predictions. The corresponding results are shown in Table 6.1 below.

Table 6.1: Prediction accuracy of different methods on Sushi preference data

METHOD	SPEARMAN COR. TEST
NEW FACTOR MODEL, GIBBS	0.56
NEW FACTOR MODEL, VB/EP	0.54
NANTONAC (KAMISHIMA AND AKAHO, 2006)	0.49

The results in Table 6.1 show that the new method compares favorably to that of Kamishima and Akaho (2006): The Gibbs sampling version of the new algorithm improves the Spearman correlation of the predictions with the test set by 0.07 in comparison with the Nantonac method, while the deterministic posterior approximation gives an improvement of 0.05. The relatively small performance difference between the Gibbs sampling inference algorithm and the deterministic posterior approximation suggests that the latter is the more practical choice for real world applications, taking into account its benefits discussed in Section 6.3.2.

6.5 Active Learning

In order to improve our recommendations we may actively ask users to provide explicit feedback on certain items. This is most commonly done on an absolute rating scale, i.e. by asking the users to rate items. However, some studies indicate that people are better able to formulate their preferences in a relative way, by ranking multiple items, see e.g. Jaeger et al. (2008). Such relative preference statements can be used directly by the model presented in Section 6.2.

Asking the user for feedback is costly as it will take time for the user to think about his or her preferences. In addition, users may find it difficult to provide a full ranking of a very large list of items, so the number of items we can enquire about is limited. When selecting this limited number of items we should take into account that not every item will be equally informative.

The question of how to optimally select the items to present to the user for feedback has been well studied in the field of Marketing, in the context of optimal design of conjoint analyses. Specifically, Vermeulen et al. (2007) consider optimal design of rank-order conjoint experiments, and Sandor and Wedel (2005) and Yu et al. (2009) discuss optimal design with heterogeneous respondents, both of which are relevant to the current application.

This problem has also received a lot of attention in the Machine Learning literature (Cohn et al., 1996; Roy and McCallum, 2001; Blum and Langley, 1997; Settles, 1994) under the name of *active learning*. In comparison, the machine learning literature puts more emphasis on incremental design of experiments and on the accompanying computational issues, both of which are also quite important for the current application.

In order to select informative items for user feedback, we must first decide on how to measure this information. In our approach to active learning / optimal design, we follow MacKay (1992) in defining the amount of information contained in a data point as the entropy reduction in our posterior distribution that we can expect upon conditioning on that data point. By maximizing the expected entropy reduction in our posterior we can then select the most informative items to present to the user for feedback. Kessels et al. (2006) discuss multiple other measures of information that can be used. The entropy of a

distribution is closely related to the *D-error* that Goos and Vandebroek (2004) propose to minimize, and we expect our results to be similar when using other popular measures of information.

Since the posterior distribution of the model given in Section 6.2 is not available in closed form, we cannot maximize the expected entropy reduction exactly. However, we can get an estimate of the amount of information in each possible observation by making use of posterior approximations. In doing so we will focus on the entropy reduction in the posterior distribution of the user parameters $q(\boldsymbol{u}_i)$, which – due to their greater number – are generally much more uncertain than the parameters of the items. To derive an expression for the approximate entropy reduction after obtaining a new observation, we assume a factorized posterior approximation over $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}$ and $\boldsymbol{S}$, optimized using a fictional Variational Bayes procedure similar to that of Lim and Teh (2007). Our aim here is not to develop another inference algorithm, but rather to develop a rough sense of the entropy in the true posterior distribution. The Variational Bayes EM algorithm uses the following update equation for the approximate posterior distribution on the user parameters:

$$q(\boldsymbol{u}_i) = N(\mu, \Sigma), \text{ with} \tag{6.13}$$

$$\Sigma = \left[\frac{1}{\pi} \mathrm{I}_K + \sum_j \mathbb{E}_q \left[\boldsymbol{v}_j \boldsymbol{v}_j' \right] \right]^{-1} \quad \mu = \Sigma \left[\sum_j \mathbb{E}_q \left[\boldsymbol{v}_j \left(s_{i,j} - b_j \right) \right] \right] \tag{6.14}$$

Note that these equations do not make any reference as to how the posterior approximation $q(\boldsymbol{V}, \boldsymbol{b})$ is determined, since our intent here is not to provide a complete inference algorithm. The entropy of this approximate posterior distribution $q(\boldsymbol{u}_i)$ is given by

$$H(q(\boldsymbol{u}_i)) \propto 0.5 \log |\Sigma| \,. \tag{6.15}$$

After adding a new item l to the ranking of the user we can use equation (6.14) to update the approximate posterior distribution $q(\boldsymbol{u}_i)$ to $q'(\boldsymbol{u}_i) = N(\mu', \Sigma')$, while keeping $q(\boldsymbol{V}, \boldsymbol{b})$

fixed. The new entropy of $q'(\boldsymbol{u}_i)$ is then given by

$$\begin{aligned} H(q'(\boldsymbol{u}_i)) &\propto 0.5 \log |\Sigma'| = -0.5 \log |\Sigma^{-1} + \mathbb{E}_q \boldsymbol{v}_l \boldsymbol{v}_l'| \\ &\propto H(q(\boldsymbol{u}_i)) - 0.5 \log(1 + \mathbb{E}_q \boldsymbol{v}_l' \Sigma \boldsymbol{v}_l) \end{aligned} \tag{6.16}$$

The most informative item is that item l, for which the difference in the entropy expressions derived here is largest. The approximation presented here is crude, but the conclusion from equation (6.16) is clear: in order to maximize the information gain, or entropy reduction, we should ask the user to rank that item for which the parameter vector $\boldsymbol{v}_l$ has the highest posterior expected Mahalanobis norm $\|\boldsymbol{v}_l\|_\Sigma$ with respect to the covariance matrix of the current posterior approximation. This has the effect of selecting items that are most informative for exactly those elements of the user vector $\boldsymbol{u}_i$ of which we are most uncertain. Comparing our model to a standard linear regression model, one could say that we should select those items that have high leverage with respect to our most uncertain regression coefficients. Note that for the approximate entropy (6.16) it does not matter what other item we compare the new item l to, or even whether we have a complete ranking with the new item or just a partial ranking. While this is obviously a very crude approximation, it still gives us a useful rule for actively selecting cases in our data set as is shown below.

We evaluate this active selection strategy using the sushi preference data, and we compare the resulting prediction accuracy with that obtained under random selection of cases from the data set. For each user the data set contains a ranking of 10 items of sushi to be used for model estimation, and a testing set of another 10 items to be used for evaluation. We actively select a subset from the estimation set for each user by starting out with an empty selection set and subsequently adding that sushi item that minimizes the expected entropy in Equation (6.16). We then use the resulting selection of data to predict the ranking of the test set. For comparison, we do the same while selecting randomly from the remaining sushi items at each iteration. We display the accuracy of the resulting predictions for different numbers of selected items from a minimum of 3 to the maximum of 10. As can be seen from Figure 6.3 the active selection method leads to faster learning of the correct preferences than random selection.

Note that the performance measures of the two selection methods in Figure 6.3 converge as the number of selected data points increases because both methods select from the same limited set of 10 data points: at the far right of the graph both methods use the same (full) data set of 10 data points. For a small number of data points the performance of the active selection method improves much faster than under random selection, indicating the practical value of such an active learning strategy for real life applications, where the user typically only provides feedback on a relatively small fraction of items.

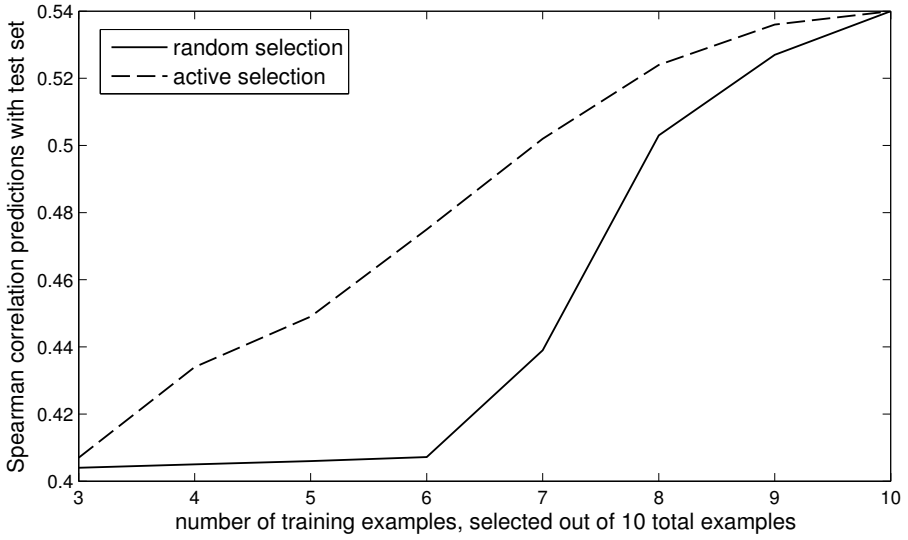


Figure 6.3: Sushi: Prediction accuracy obtained using active versus random selection of estimation data

6.6 Conclusion

We have proposed a Bayesian factor model to learn user preferences for the purpose of product recommendation. Learning rankings of preferred items with this model can be done quickly and efficiently at large scale, using the two inference algorithms we have developed. The accuracy of our model was demonstrated on a real world data set and was shown to improve upon existing methods. In addition, we have shown that the model can also be used effectively for active preference elicitation. By actively selecting product comparisons to present to the user, we can uncover the user's preferences without requiring large amounts of feedback. This makes the process of preference elicitation much less burdensome on the user, and it can dramatically improve prediction accuracy for real life

applications.

Chapter 7

Nederlandse samenvatting

> Kansrekening is slechts gezond verstand
> gereduceerd tot calculus.
>
> Pierre Simon Laplace, 1812

Het citaat hierboven is afkomstig uit het baanbrekende werk van Pierre Simon Laplace, *Théorie analytique des probabilités*, waarin hij de basis legt voor wat tegenwoordig bekend staat als Bayesiaanse analyse. In dit werk beschrijft hij kansrekening en statistiek als methodes die *men nauwkeurig laat maken wat weldenkende mensen instinctief aanvoelen, vaak zonder hier een reden voor te kunnen geven.* Deze beschrijving bevat een diep inzicht: Kansrekening biedt een strak en eenvoudig recept dat voorschrijft hoe we logisch kunnen redeneren onder onzekere omstandigheden. Dit inzicht is wat mij aantrok tot de econometrie toen ik hier voor het eerst mee in aanraking kwam. Naar mate mijn kennis van kansrekening toenam, kwam echter ook steeds meer het besef dat dit citaat op zichzelf de dingen te eenvoudig voorstelt: Gezond verstand vertalen naar de wiskunde is in de praktijk vaak erg lastig. Net als bij het maken van een goede vertaling uit het Nederlands naar een vreemde taal, kost ook het vertalen van ideeën in de taal van kansen veel oefening. Een bijkomend probleem is dat de wiskunde behorende bij een goede vertaling vaak te lastig is om exact op te lossen. Het is dan ook de taak van statistici en econometristen om praktische manieren te vinden voor het omzetten van gezond verstand naar kansrekening, en om slimme nieuwe manieren te verzinnen om vervolgens de resulterende wiskundige problemen op te lossen. Dit proefschrift bevat mijn bijdrage hiertoe.

7.1 Probabilistische modellering

Econometristen houden zich bezig met de vraag hoe data gebruikt kan worden om te leren over de economie. In de praktijk komt dit neer op het bestuderen van *data sets*, bestaand uit meerdere observaties van een bepaalde *afhankelijke variabele* y waarin we geïnteresseerd zijn, bijvoorbeeld het bruto binnenlands product van een land, en een aantal *verklarende variabelen* x, bijvoorbeeld de bevolkingsdichtheid en de beschikbaarheid van grondstoffen in dat land. Het doel van de econometrist is dan om iets te leren over de economische relatie tussen x en y. Op zichzelf is een lijst met een aantal x en y waarden niet erg nuttig: het geeft ons geen economische inzichten en het zegt niets over situaties waarvan we nog geen data hebben. Het leren uit data kan daarom pas plaats vinden als we eerst de context van deze data bepalen. In de econometrie bestaat deze context meestal uit een probabilistisch model.

Een model is een wiskunde beschrijving van een aantal (economische) hypotheses over de relatie tussen de variabelen in een data set. Zo'n model is vaak afkomstig uit de economische theorie. Een model beschrijft wat we van de wereld weten voordat we de data hebben gezien, en op deze manier maakt het model de data interpreteerbaar: In plaats van een grote lijst met getallen kunnen we onze data nu interpreteren in de context van een model. Een economisch model beschrijft op zijn best een ruwe benadering van de werkelijkheid. Economiën zijn zo complex dat we nooit over alle relevante informatie kunnen beschikken, of alle onderliggende processen kunnen begrijpen. Door onzekerheid toe te laten in onze modellen geven we toe dat ze niet perfect zijn. Kansrekening is een taal om deze onzekerheid precies te maken; het laat ons specificeren hoeveel en wat voor type onzekerheid we precies in onze modellen willen opnemen. Zulke modellen met onzekerheid noemen we probabilistische modellen.

Een econometrisch model bevat meestal twee vormen van onzekerheid: De eerste is onzekerheid over de economische relatie tussen x en y, wat we *parameter* of *model onzekerheid* noemen. Het kan bijvoorbeeld voorkomen dat we niet weten of deze relatie lineair of niet-lineair is, en zelfs als we aannemen dat een relatie lineair is dan weten we vaak nog niet met welke coëfficiënt. Dit type onzekerheid drukken we vaak uit in onze modellen door ze afhankelijk te maken van een aantal onbekende *parameters* die we

noteren als θ.

Het tweede type onzekerheid is een onderkenning dat een model fouten maakt: Zelfs een erg goed model zal nooit alle relevante factoren van een economisch proces bevatten, of de precieze relatie tussen deze variabelen en de afhankelijke variabele kunnen beschrijven. Door extra onzekerheid in het model op te nemen laten we de mogelijkheid bestaan dat iets de afhankelijke variabele y beïnvloedt dat we niet in het model hebben opgenomen. Dit type onzekerheid wordt doorgaans beschreven door een *foutenterm* waarvoor we meestal het symbool ϵ gebruiken. Ook bevat een model vaak een aantal *latente variabelen* die we s noemen. Zulke latente variabelen beschrijven vaak belangrijke maar verborgen aspecten van de landen of personen in onze data set, en worden bijvoorbeeld gebruikt in situaties waarin onze data incompleet is. De onzekerheid in ϵ en s maken we expliciet door aan hen een *kansverdeling* toe te kennen, die beschrijft hoe veel en wat voor type onzekerheid we aan deze variabelen verbinden.

Het is belangrijk om stil te staan bij het feit dat dit niet de enige manier is om uit data te leren. In veel situaties worden andere oplossingen gebruikt, die niet op probabilistische modellen zijn gebaseerd, zoals bijvoorbeeld in de klassieke nonparameterische econometrie. Het grote voordeel van probabilistische modellering is echter dat het een duidelijk en wijd toepasbaar recept oplevert voor het redeneren onder onzekerheid. Probabilistische modellen zijn inzichtelijk en hun structuur is modulair, wat onderzoekers de mogelijkheid geeft om elementen uit eerdere modellen te hergebruiken en de combineren voor het analyseren van verschillende vraagstukken. Een bijkomend voordeel is dat deze manier van redeneren een scheiding aanbrengt tussen de kennis en aannames die we hebben (het model) en het uiteindelijke algoritme dat we gebruiken om van de data te leren. Op deze manier blijft het duidelijk wat we uit de data kunnen concluderen, en welke van onze conclusies het resultaat zijn van het model.

7.2 Op aannemelijkheid gebaseerde econometrie

In de context van een probabilistisch model is ons doel om uit de data te leren welke van de mogelijke parameterwaarden aannemelijk zijn. Het is daarom belangrijk om na te denken over welke informatie de data eigenlijk bevat. Wat betreft de zienswijze van

dit proefschrift kunnen we zeggen dat alle informatie in een data set is beschreven door de *aannemelijkheidsfunctie*: Een probabilistisch model geeft ons een kansverdeling voor de afhankelijke variabele y, gegeven de verklarende variabelen x en de waarden van de parameters θ, wat we kunnen schrijven als $p(y|x;\theta)$. Als we de waarden van x en y in deze uitdrukking vastleggen op de waarden die we hebben waargenomen in onze data set, hebben we een functie over die alleen van de parameters θ afhankelijk is. Deze functie schrijven we meestal als $\mathrm{L}(\theta)$ en noemen we de aannemelijkheidsfunctie. Simpel gezegd is het de aannemelijkheidsfunctie die ons zegt hoe goed de data door het model wordt verklaard als de parameters gelijk zijn aan θ. De stelling dat de aannemelijkheidsfunctie alle informatie in de data beschrijft noemen we het *aannemelijkheidsprincipe*. In de praktijk betekent dit principe dat twee verschillende data sets met dezelfde aannemelijkheidsfunctie tot dezelfde statistische conclusies zouden moeten leiden (onder het zelfde model).

Uitgaande van het aannemelijkheidsprincipe is het logisch dat de aannemelijkheidsfunctie de basis vormt voor het leren uit data van de parameterwaarden θ. Hoewel ook andere methodes gebruikt worden, is veel van de moderne econometrie inderdaad op aannemelijkheid gebaseerd. De twee meest gebruikte methodes voor het werken met aannemelijkheid zijn de methode van de *maximale aannemelijkheid* en *Bayesiaanse analyse*.

Als we de methode van maximale aannemelijkheid gebruiken, dan wordt onze schatting van de parameters θ gegeven door die waarden die corresponderen met het maximum van de aannemelijkheidsfunctie:

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \mathrm{L}(\theta). \tag{7.1}$$

In andere woorden, we selecteren het model (beschreven door θ) waarvoor onze data het meest aannemelijk is.

In tegenstelling tot de meest aannemelijke schatter geeft Bayesiaanse analyse ons niet alleen een puntschatting van de parameters, maar een hele *posterior verdeling* of *nadichtheid*, die weergeeft hoe waarschijnlijk elke parameterwaarde is nadat we de data hebben gezien. Om een Bayesiaanse analyse van een probleem te maken moeten we eerst formuleren hoe waarschijnlijk we elke parameterwaarde vinden voordat we de data hebben

gezien. Dit doen we in de vorm van een *prior* of *voordichtheid*, geschreven als $p(\theta)$. De nadichtheid wordt dan verkregen met de volgende formule:

$$p(\theta|y) \propto p(y|x;\theta)p(\theta). \tag{7.2}$$

De aannemelijkheidsfunctie neemt hier de rol aan van een wegingsfunctie over the hypotheses die door θ worden beschreven. De voordichtheid $p(\theta)$ beschrijft het gewicht dat we aan elke parameterwaarde toekennen voordat we de data hebben gezien. De aannemelijkheidsfunctie werkt deze gewichten dan bij door ze te vermenigvuldigen met de aannemelijkheid. Het resultaat zijn de gewichten die we geven aan de parameterwaarden nadat we de data hebben gezien. Deze manier van het vermenigvuldigend bijwerken van de gewichten lijkt vrij willekeurig, maar het is de enige consistente manier om van de oorspronkelijke gewichten $p(\theta)$ naar nieuwe gewichten $p(\theta|y)$ te gaan; elke andere manier van het bijwerken van de gewichten is in strijd met zichzelf, zoals uitgelegd door onder anderen Jaynes en Bretthorst (2003).

7.3 Rekenkundige uitdagingen

Bayesiaanse analyse en de meest aannemelijke schatter volgen beiden een simpel concept; hun essentie wordt beschreven door slechts een enkele vergelijking (vergelijkingen 7.1 en 7.2). Echter, het daadwerkelijke rekenen met deze vergelijkingen is vaak erg lastig. Om de aannemelijkheidsfunctie $\mathrm{L}(\theta)$ uit te rekenen moeten we integreren over de fouten ϵ en de latente variabelen s:

$$\mathrm{L}(\theta) = \int\int p(y, \epsilon, s|x;\theta) d\epsilon ds. \tag{7.3}$$

Om de meest aannemelijke schatter te verkrijgen moeten we deze functie dan ook nog eens maximaliseren, wat zeer lastig kan zijn. Voor een Bayesiaanse analyse moeten we extra integraties doen over de parameters θ om te nadichtheid goed te beschrijven. Bijvoorbeeld, om de verwachting van een functie $f(\theta)$ onder de nadichtheid uit te rekenen

moeten we de volgende integraal oplossen:

$$\mathbb{E}[f(\theta)|y] = \int f(\theta)p(\theta|y)d\theta. \tag{7.4}$$

Beide integralen (7.3) en (7.4) kunnen meestal niet analytisch worden uitgerekend.

Ook het numeriek uitrekenen van de integralen is lastig als θ, ϵ en s van erg hoge dimensie zijn. Dit rekenkundige probleem heeft de NP-moeilijk complexiteit (Cooper, 1990; Bacchus et al., 2003; Dagum en Luby, 1993), wat betekent dat de hoeveelheid rekenwerk exponentieel toeneemt met de dimensie van θ, ϵ en s. In de praktijk betekent dit dat het exact uitrekenen van de aannemelijkheidsfunctie snel onmogelijk wordt voor algemene probabilistische modellen. Om in de praktijk met deze modellen te kunnen werken moeten we daarom benaderingen gebruiken.

Het meest gebruikte type benadering in de econometrie is de *Monte Carlo methode*, maar er bestaan ook meerdere *deterministische benaderingen*. Recentelijk worden deze twee types van benadering ook succesvol gecombineerd, zoals bijvoorbeeld in hoofdstuk 5 van dit proefschrift. De laatste jaren zijn door de ontwikkelingen in Monte Carlo methodes en andere benaderingsmethodes veel problemen oplosbaar geworden met Bayesiaanse analyse en de methode van maximale aannemelijkheid, maar de rekenkundige problemen zijn nog lang niet opgelost.

Dagum en Luby (1993) bewijzen dat zelfs het benaderen van (7.3) en (7.4) met een zekere nauwkeurigheid de NP-moeilijk complexiteit heeft in het algemene geval, wat betekent dat er niet één enkel algoritme kan bestaan dat snel en efficiënt schattingen kan maken voor alle probabilistische modellen. Het lijkt of Monte Carlo methodes aan deze val ontsnappen, aangezien de nauwkeurigheid van een Monte Carlo benadering niet direct van de dimensionaliteit afhangt, maar het simuleren van algemene kansverdelingen is helaas NP-moeilijk op zichzelf. Cooper (1990) zegt daarom dat dit 'suggereert dat onderzoek moet worden weggeleid van de zoektocht naar algemene, efficiënte probabilistische algoritmes, en moet worden gestuurd naar het ontwerp van efficiënte algoritmes voor speciale gevallen, gemiddelde gevallen, en benaderingsmethodes.' Dit is precies het doel van dit proefschrift: Het ontwikkelen van benaderingsalgoritmes voor het oplossen van interessant specifieke problemen in de econometrie, waarvoor op aannemelijkheid gebaseerde

econometrische analyse vooralsnog niet mogelijk was.

7.4 Overzicht

Het hierboven gestelde doel wordt gerealiseerd in vijf aparte hoofdstukken, die op zichzelf staan en apart gelezen kunnen worden. De eerste twee hoofdstukken gebruiken de methode van de maximale aannemelijkheid: Hier worden twee toepassingen besproken waarbij de aannemelijkheidsfunctie moeilijk uit te rekenen is. In deze hoofdstukken ontwikkelen we deterministische algoritmes om de aannemelijkheidsfunctie te benaderen. De overige hoofdstukken zijn geschreven vanuit het Bayesiaanse perspectief: De toepassingen die hier worden besproken leiden tot rekenkundige uitdagingen doordat er geavanceerde modellen worden gebruikt of doordat er grote hoeveelheden data worden geanalyseerd. In deze hoofdstukken gebruiken we een combinatie van Monte Carlo methodes en deterministische benaderingen.

Hoofdstuk 2 is gebaseerd op het werk van Abbring en Salimans (2013). Hier presenteren we een nieuwe methode om de aannemelijkheidfunctie uit te rekenen voor *mixed hitting-time modellen*: duurmodellen gebaseerd op de tijd die het kost voor een latent Lévy proces om een heterogene drempelwaarde te overschrijden. De aannemelijkheidsfunctie voor dit soort modellen is niet beschikbaar in analytische vorm, maar de Laplace transformatie van de aannemelijkheidsfunctie is wel bekend. Door gebruik te maken van speciale eigenschappen van Lévy processen, ontwikkelen we een algoritme dat deze Laplace transformatie kan inverteren om zo de aannemelijkheidsfunctie te verkrijgen. Dit algoritme gebruiken we om de meest aannemelijke schatter uit te kunnen rekenen voor mixed hitting-time modellen. Deze schattingsmethode gebruiken we vervolgens voor een analyse van de stakingsdata van Kennan (1985).

Hoofdstuk 3 is gebaseerd op het werk van Salimans en Fok (2013). Hier ontwikkelen we een algoritme voor het benaderen van maximale aannemelijke schatters voor dynamische modellen van extreem hoge dimensie, toegepast op data die niet normaal verdeeld is. Onze methode is gebaseerd op het Expectation Maximization [EM] algoritme, waar we de Expectation stap benaderen met gebruik van het Expectation Propagation [EP] algoritme van Minka (2001). Met behulp van simulatiestudies laten we zien dat deze methode er

goed in slaagt om de parameters van een dynamisch model terug te schatten uit de data. Verder passen we onze methode succesvol toe op twee problemen uit de praktijk: (i) Het voorspellen van aantallen verkochte kranten over verschillende individuele winkels; en (ii) het voorspellen van de uitkomsten van schaakwedstrijden. Voor beide toepassingen gebruiken we dynamische modellen van extreem hoge dimensie, zonder normale verdeling.

Hoofdstuk 4 is gebaseerd op het werk van Salimans (2012). In dit hoofdstuk kijken we naar regressie analyses van economische groeidata in een cross-sectie van landen. Dit soort regressie analyses worden bemoeilijkt door twee soorten modelonzekerheid: De onzekerheid in het selecteren van verklarende variabelen en de onzekerheid over de vorm van de regressiefunctie. De meeste beschouwingen in de literatuur bekijken deze problemen los van elkaar, terwijl het juist essentieel is deze twee vormen van onzekerheid samen te behandelen. In hoofdstuk 4 ontwikkelen we een nieuwe methode die zo'n analyse mogelijk maakt, met gebruik van flexibele niet-lineaire modellen gespecificeerd door *Gaussian process prioren*, en met gebruik van *Bayesian model averaging* voor de selectie van de verklarende variabelen. Met deze methode breiden we het vaak gebruikte lineaire model uit, zodat het kan omgaan met de parameter heterogeniteit die wordt voorspeld door de *new growth theory* literatuur, terwijl tegelijkertijd de onzekerheid in de selectie van verklarende variabelen wordt behandeld. Met het in acht nemen van deze onzekerheid, onderbouwt onze analyse het bewijs voor parameter heterogeniteit zoals gepresenteerd in een aantal eerdere studies. Als we tegelijkertijd aandacht besteden aan de onzekerheid in de vorm van de regressiefunctie, vinden we echter dat een aantal van de effecten van verklarende variabelen uit de literatuur niet robuust zijn over verschillende landen en selecties van variabelen.

Hoofdstuk 5 is gebaseerd op het werk van Salimans en Knowles (2013). Hier ontwikkelen we een algemeen algoritme voor het benaderen van lastige Bayesiaanse nadichtheden. Het algoritme minimaliseert de Kullback-Leibler afstand tussen een benadering in de exponentiële familie van kansverdelingen en de echte nadichtheid. Onze methode kan worden gebruikt voor het benaderen van willekeurige nadichtheden, onder voorwaarde dat de nadichtheid in analytische vorm beschikbaar is (op de normalisatie na). Elke willekeurige kansverdeling uit de exponentiële familie kan worden gebruikt voor het vormen

van de benadering, en zelfs het gebruik van mengsels van zulke verdelingen is mogelijk. Dit betekent dat de benadering extreem nauwkeurig gemaakt kan worden. De snelheid en nauwkeurigheid van onze methode demonsteren we met behulp van diverse voorbeelden uit de praktijk.

Hoofdstuk 6 is gebaseerd op het werk van Salimans et al. (2012). In dit hoofdstuk ontwikkelen we een model voor het ontdekken van de voorkeuren van consumenten in de vorm van rangschikkingen over meerdere producten. Dit model kan vervolgens worden gebruikt voor het aanraden van nieuwe producten aan deze consumenten. Het model kan worden toegepast op door de consument uitgesproken voorkeuren in vergelijkingen tussen twee producten, of op (incomplete) lijsten van favoriete producten. We presenteren twee methodes voor het schatten van dit model. Beide methodes werken efficiënt voor toepassingen met veel producten en gebruikers. De nauwkeurigheid van de voorspellingen uit het model demonstreren we op de veel gebruikte Sushi dataset van Kamishima en Akaho (2006). Ten slotte laten we zien hoe het model gebruikt kan worden om actief data te verzamelen, en op deze manier goede aanbevelingen te geven in gevallen waar weinig data beschikbaar is.

Bibliography

Aalen, O. O., Gjessing, H. K., 2001. Understanding the shape of the hazard rate: A process point of view. Statistical Science 16 (1), 1–14.

Abate, J., Whitt, W., 1992. The Fourier-series method for inverting transforms of probability distributions. Queueing Systems 10, 5–88.

Abbring, J. H., 2002. Stayers versus defecting movers: A note on the identification of defective duration models. Economics Letters 74, 327–331.

Abbring, J. H., 2010. Identification of dynamic discrete choice models. Annual Review of Economics 2, 367–394.

Abbring, J. H., March 2012. Mixed hitting-time models. Econometrica 80 (2), 783–819.

Abbring, J. H., Salimans, T., 2013. The likelihood of mixed hitting times. Mimeo, Erasmus University Rotterdam, The Netherlands.

Aghion, P., Caroli, E., Garcia-Penalosa, C., 1999. Inequality and economic growth: The perspective of the new growth theories. Journal of Economic Literature 37 (4), 1615–1660.

Albert, J., 2009. Bayesian Computation with R. Springer Science, New York. Second edition.

Amari, S., 1997. Neural learning in structured parameter spaces - natural Riemannian gradient. In: Advances in Neural Information Processing Systems. MIT Press, pp. 127–133.

Andersen, P. K., Borgan, Ø., Gill, R. D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer-Verlag, New York.

Attias, H., 2000. A variational Bayesian framework for graphical models. In: Advances in Neural Information Processing Systems (NIPS) 12. pp. 209–215.

Azariadis, C., Drazen, A., 1990. Threshold externalities in economic development. Quarterly Journal of Economics 105(2), 501–526.

Bacchus, F., Dalmao, S., Pitassi, T., oct. 2003. Algorithms and complexity results for sat and bayesian inference. In: Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on. pp. 340 – 351.

Barnard, G., Jenkins, G., Winsten, C., 1962. Likelihood inference and time series. Journal of the Royal Statistical Society, A 3 (125), 321–372.

Barro, R., 1991. Economic growth in a cross section of countries. Quarterly Journal of Economics 106(2), 407–43.

Barro, R. J., Sala-i-Martin, X. I., 2003. Economic Growth. The MIT Press.

Basturk, N., Paap, R., Dijk, D. v., January 2012. Structural differences in economic growth: an endogenous clustering approach. Applied Economics 44 (1), 119–134.
URL http://dx.doi.org/10.1080/00036846.2010.500274

Beal, M. J., Ghahramani, Z., 2002. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In: Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting. p. 453.

Beal, M. J., Ghahramani, Z., 2006. Variational Bayesian learning of directed graphical models with hidden variables. Bayesian Analysis 1 (4), 793–832.

Bengoa, M., Sanchez-Robles, B., 2003. Foreign direct investment, economic freedom and growth: new evidence from latin america. European Journal of Political Economy 19 (3), 529 – 545, economic Freedom.
URL http://www.sciencedirect.com/science/article/pii/S0176268003000119

Bengtsson, T., Bickel, P., Li, B., 2008. Curse-of-dimensionality revisited: Collapse of the particle filter in very large systems. In: Nolan, D., Speed, T. (Eds.), Probability and Statistics: Essays in Honor of David A. Freedman. Institute of Mathematical Statistics, Beachwoord, Ohio, USA, pp. 316–334.

Bertoin, J., 1996. Lévy Processes. No. 121 in Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge.

Beskos, A., Crisan, D., Jasra, A., Mar. 2011. On the Stability of Sequential Monte Carlo Methods in High Dimensions. ArXiv e-prints.

Bickel, P., Li, B., Bengtsson, T., 2008. Sharp failure rates for the bootstrap particle filter in high dimensions. In: Clarke, B., Ghosal, S. (Eds.), Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh. Institute of Mathematical Statistics, Beachwoord, Ohio, USA, pp. 318–329.

Binder, M., Pesaran, M. H., 1999. Stochastic growth models and their econometric implications. Journal of Economic Growth 4, 139–183.

Birnbaum, A., 1962. On the foundations of statistical inference. Journal of the American Statistal Association 298 (57), 269–326.

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial intelligence 97 (1), 245–271.

Borensztein, E., Gregorio, J. D., Lee, J.-W., 1998. How does foreign direct investment affect economic growth? Journal of International Economics 45 (1), 115 – 135.
URL http://www.sciencedirect.com/science/article/pii/S0022199697000330

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010). Springer, pp. 177–187.

Boyarchenko, S., Levendorskiĭ, S., 2007. Irreversible Decisions under Uncertainty: Optimal Stopping Made Easy. Springer-Verlag, Berlin.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Cameron, A. C., Trivedi, P. K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge.

Carlin, B. P., Chib, S., 1995. Bayesian model choice via markov chain monte carlo methods. Journal of the Royal Statistical Society, Ser. B 57, 473–484.

Carlin, B. P., Polson, N. G., Stoffer, D. S., 1992. A Monte Carlo approach to nonnormal and nonlinear state-space modeling. Journal of the American Statistical Association 87 (418), 493–500.

Carter, C. K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81 (3), 541–553.

Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis-Hastings output. Journal of the American Statistical Association 96, 270–281.

Ciccone, A., Jarocinski, M., 2010. Determinants of economic growth: Will data tell? American Economic Journal: Macroeconomics, forthcoming.

Cohn, D., Ghahramani, Z., Jordan, M., 1996. Active learning with statistical models. Journal of Artificial Intelligence Research 4, 129–145.

Cooper, G. F., 1990. The computational complexity of probabilistic inference using bayesian belief networks. Artificial Intelligence 42 (23), 393 – 405.
URL http://www.sciencedirect.com/science/article/pii/000437029090060D

Cox, D. R., 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society Series B 34, 187–202.

Cox, D. R., Miller, H. D., 1965. The Theory of Stochastic Processes. Methuen, London.

Cseke, B., Heskes, T., Feb. 2011. Approximate marginals in latent gaussian models. J. Mach. Learn. Res. 12, 417–454.
URL http://dl.acm.org/citation.cfm?id=1953048.1953061

Cuaresma, C., Doppelhofer, G., 2007. Nonlinearities in cross-country growth regressions: A Bayesian averaging of thresholds (BAT) approach. Journal of Macroeconomics 29, 541–554.

Dagum, P., Luby, M., 1993. Approximating probabilistic inference in bayesian belief networks is np-hard. Artificial Intelligence 60 (1), 141 – 153.
URL http://www.sciencedirect.com/science/article/pii/000437029390036B

Dale, A. I., 1995. Pierre-Simon Laplace, Philoshophical Essays on Probabilities. Springer-Verlag Berlin.

Dangauthier, P., Herbrich, R., Minka, T., Graepel, T., 2008. Trueskill through time: Revisiting the history of chess. Advances in Neural Information Processing Systems 20, 931–938.

Davies, B., 2002. Integral transforms and their applications. Springer-Verlag.

de Freitas, N., Pedro, Russell, S. J., 2001. Variational MCMC. In: UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 120–127.
URL http://portal.acm.org/citation.cfm?id=647235.720242

Dickey, J. M., 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. The Annals of Mathematical Statistics 42 (1), 204–223.

Dixit, A. K., 1989. Entry and exit decisions under uncertainty. Journal of Political Economy 97 (3), 620–638.

Dixit, A. K., Pindyck, R. S., 1994. Investment under Uncertainty. Princeton University Press.

Durbin, J., Koopman, S., 2001. Time Series Analysis by State Space Methods. Oxford University Press.

Durbin, J., Koopman, S. J., 1997. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84 (3), 669–684.

Durlauf, S. N., 1993. Nonergodic economic growth. Review of Economic Studies 60, 349–366.

Durlauf, S. N., Johnson, P. A., 1995. Multiple regimes and cross-country growth behaviour. Journal of Applied Econometrics 10(4), 365–384.

Fernandez, C., Ley, E., Steel, M. F., 2001a. Benchmark priors for Bayesian model averaging. Journal of Econometrics 100(2), 381–427.

Fernandez, C., Ley, E., Steel, M. F., 2001b. Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16(5), 563–576.

Frühwirth-Schnatter, 2004. Efficient Bayesian parameter estimation. In: Harvey, A. C., Koopman, S. J., Shephard, N. (Eds.), State Space and Unobserved Component Models: Theory and Applications. Cambridge University Press, Ch. 7, pp. 123–151.

George, E. I., McCulloch, R. E., 1997. Approaches for Bayesian variable selection. Statistica Sinica 7, 339–373.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. Wiley-Interscience.

Geweke, J., 2012. Nonparametric bayesian modelling of monotone preferences for discrete choice experiments. Journal of Econometrics 171 (2), 185 – 204.
URL http://www.sciencedirect.com/science/article/pii/S0304407612001509

Gilks, W., Thomas, A., Spiegelhalter, D., 1994. A language and program for complex bayesian modelling. The Statistician, 169–177.

Giordani, P., Kohn, R., Villani, M., 2009. Regression density estimation using smooth adaptive Gaussian mixtures. Journal of Econometrics 153 (2), 155–173.

Girolami, M., Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (2), 123–214.
URL http://dx.doi.org/10.1111/j.1467-9868.2010.00765.x

Goos, P., Vandebroek, M., 2004. Outperforming completely randomized designs. Journal of Quality Technology 36 (1), 12–26.

Green, P., 1995. Reversible jump markov chain monte carlo computation and Bayesian model choice. Biometrika 82, 711–732.

Greenberg, E., 2008. Introduction to Bayesian Econometrics. Cambridge University Press.

Grier, K. B., Tullock, G., 1989. An emperical analysis of cross-national economic growth. Journal of Monetary Economics 24 (2), 259–276.

Grossman, G. M., Helpman, E., 1991. Quality ladders in the theory of growth. The Review of Economic Studies 58 (1), 43–61.
URL http://restud.oxfordjournals.org/content/58/1/43.abstract

Harvey, A. C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Heckman, J. J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.

Herbrich, R., Minka, T., Graepel, T., 2007. Trueskill: A Bayesian skill rating system. Advances in Neural Information Processing Systems 20.

Hoeting, J. A., Raftery, A. E., Madigan, D., 2002. Bayesian variable and transformation selection in linear regression. Journal of Computational and Graphical Statistics 11 (3), 485–507.

Hoffman, M., Blei, D., Bach, F., 2010. Online learning for latent Dirichlet allocation. Advances in Neural Information Processing Systems 23.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., Karhunen, J., 2010. Approximate Riemannian conjugate gradient learning for fixed-form variational bayes. Journal of Machine Learning Research, 3235–3268.

Jaeger, S. R., Jørgensen, A. S., Aaslyng, M. D., Bredie, W. L., 2008. Best-worst scaling: An introduction and initial comparison with monadic rating for preference elicitation with food products. Food Quality and Preference 19 (6), 579 – 588.

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G., 2010. Recommender Systems: An Introduction. Cambridge University Press.

Jaynes, E., Bretthorst, G. L., 2003. Probability Theory: The Logic of Science. Cambridge University Press.

Jong, P. D., Shephard, N., 1995. The simulation smoother for time series models. Biometrika 82, 339–350.

Jungbacker, B., Koopman, S. J., 2007. Monte Carlo estimation for nonlinear non-Gaussian state space models. Biometrika 94 (4), 827–839.

Kalaitzidakis, P., Mamuneas, T. P., Stengos, T., 2000. A non-linear sensitivity analsyis of cross-country growth regressions. Canadian Journal of Economics 33(3), 604–17.

Kamishima, T., Akaho, S., 2006. Nantonac collaborative filtering. In: Proceedings of The International Workshop on Data-Mining and Statistical Science. pp. 117–124.

Kass, R. E., Wasserman, L., 1995. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. Journal of the American Statistical Association 90 (431), 928–934.

Kennan, J., 1985. The duration of contract strikes in U.S. manufacturing. Journal of Econometrics 28, 5–28.

Kessels, R., Goos, P., Vandebroek, M., 2006. A comparison of criteria to design efficient choice experiments. Journal of Marketing Research 43 (3), pp. 409–419.
URL http://www.jstor.org/stable/30162415

Kim, H.-C., Ghahramani, Z., 2006. Bayesian Gaussian process classification with the EM-EP algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12), 1948–1959.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. The Review of Economic Studies 65 (3), pp. 361–393.
URL http://www.jstor.org/stable/2566931

Kitagawa, G., 1987. Non-Gaussian state-space modeling of nonstationary time series. Journal of the American Statistical Association 82 (400), 1032–1041.

Kloek, T., van Dijk, H. K., 1978. Bayesian estimates of equation system parameters: An application of intergration by Monte Carlo. Econometrica 46, 1–20.

Knowles, D. A., Minka, T. P., 2011. Non-conjugate variational message passing for multinomial and binary regression. In: Advances in Neural Information Processing Systems (NIPS). No. 25.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. The MIT Press.

Koop, G., Poirier, D. J., 2004. Bayesian variants of some classical semiparametric regression techniques. Journal of Econometrics 123 (2), 259–282.

Koopman, S. J., 1993. Disturbance smoother for state space models. Biometrika 80, 117–126.

Koopman, S. J., Lucas, A., Scharth, M., 2011. Numerically accelerated importance sampling for nonlinear non-Gaussian state space models. Tech. Rep. TI 2011-057/4, Tinbergen Institute.

Koopman, S. J., Nguyen, T. M., 2012. Fast efficient importance sampling by state space methods. Tech. Rep. TI 2012-008/4, Tinbergen Institute.

Koopman, S. J., Shephard, N., Doornik, J. A., 1999. Statistical algorithms for models in state space using SsfPack 2.2. Econometrics Journal 2, 113–166.

Koren, Y., Bell, R., Volinsky, C., aug. 2009. Matrix factorization techniques for recommender systems. Computer 42 (8), 30 –37.

Kormendi, R., Meguire, P., 1985. Macroeconomic determinant of growth, cross-country evidence. Journal of Monetary Economic 16, 141–163.

Kuss, M., Rasmussen, C. E., 2005. Assesing approximations for gaussian process classification. Journal of Machine Learning Research 6, 1679–1704.

Kyprianou, A. E., 2006. Introductionary Lectures on Fluctuations of Lévy Processes with Applications. Springer-Verlag, Berlin.

Lancaster, T., 1972. A stochastic model for the duration of a strike. Journal of the Royal Statistical Society Series A 135 (2), 257–271.

Lancaster, T., 1979. Econometric methods for the duration of unemployment. Econometrica 47, 939–956.

Leamer, E. E., 1983. Let's take the con out of econometrics. American Economic Review 73(1), 31–43.

Lee, J., McVinish, R., Mengersen, K., 2011. Population monte carlo algorithm in high dimensions. Methodology and Computing in Applied Probability 13, 369–389, 10.1007/s11009-009-9154-2.
URL http://dx.doi.org/10.1007/s11009-009-9154-2

Lee, M.-L. T., Whitmore, G. A., 2006. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. Statistical Science 21 (4), 501–513.

Lee, Y., Nelder, J., 2006. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. Chapman and Hall, London.

Levine, R., Renelt, D., 1992. Sensitity analysis of cross-country growth regressions. American Economic Review 82(4), 942–963.

Ley, E., Steel, M. F., 2009. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Journal of Applied Econometrics 24, 651–674.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., Berger, J. O., 2008. Mixtures of g priors for bayesian variable selection. Journal of the American Statistical Association 103 (481), 410–423.

Liesenfeld, R., Richard, J.-F., 2008. Improving MCMC, using efficient importance sampling. Computational Statistics and Data Analysis 53 (2), 272 – 288.
URL http://www.sciencedirect.com/science/article/pii/S0167947308003770

Lim, Y. J., Teh, Y. W., 2007. Variational Bayesian Approach to Movie Rating Prediction. In: KDD Cup and Workshop 2007.

Liu, Z., Stengos, T., 1999. Non-linearities in cross country growth regressions: A semiparametric approach. Journal of Applied Econometrics 14(5), 527–538.

Lovell, M., 2008. A simple proof of the fwl theorem. The Journal of Economic Education 39 (1), 88–91.

Maasoumi, E., Racine, J., Stengos, T., 2007. Growth and convergence: A profile of distribution dynamics and mobility. Journal of Econometrics 136 (2), 483 – 508.

MacKay, D., 1998. Introduction to gaussian processes. In: Bishop, C. (Ed.), Neural Networks and Machine Learning. Springer-Verlag.

MacKay, D. J., 1992. Information-based objective functions for active data selection. Neural Computation 4, 590–604.

Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. International Statistical Review 63, 215–232.

Mankiw, N., Romer, D., Weil, D., 1992. A contribution to the empirics of economic growth. Quarterly Journal of Economics 107(2), 407–437.

McDonald, R., Siegel, D., 1986. The value of waiting to invest. Quarterly Journal of Economics 101 (4), 707–728.

McLachlan, G., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley.

Meng, X.-L., Rubin, D. B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80 (2), 267–278.
URL http://biomet.oxfordjournals.org/content/80/2/267.abstract

Minier, J., 2007. Nonlinearities and robustness in growth regressions. American Economic Review 97(2), 388–392.

Minka, T., 1998. Expectation-maximization as lower bound maximization. http://research.microsoft.com/en-us/um/people/minka/papers/em.html.

Minka, T., 2005. Divergence measures and message passing. Tech. Rep. MSR-TR-2005-173, Microsoft Research.

Minka, T., Lafferty, J., 2002. Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. pp. 352–359.

Minka, T. P., 2001a. Expectation propagation for approximate Bayesian inference. In: UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 362–369.
URL http://portal.acm.org/citation.cfm?id=720257

Minka, T. P., 2001b. A family of algorithms for approximate Bayesian inference. Ph.D. thesis, MIT.

Minka, T. P., Winn, J. M., Guiver, J. P., Knowles, D. A., 2010. Infer.NET 2.4.

Mitchell, T., Beauchamp, J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1046.

Moreno, R., Trehan, B., 1997. Location and the growth of nations. Journal of Economic Growth 2, 399–418, 10.1023/A:1009741426524.
URL http://dx.doi.org/10.1023/A:1009741426524

Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A., 2009. Robust stochastic approximation approach to stochastic programming. SIAM J. Optim 19 (4), 1574–1609.

Nickisch, H., Rasmussen, C. E., Oct. 2008. Approximations for binary Gaussian process classification. Journal of Machine Learning Research 9, 2035–2078.

Nocedal, J., Wright, S. J., 2006. Numerical Optimization. Springer-Verlag.

Nott, D., Tan, S., Villani, M., Kohn, R., 2012. Regression density estimation with variational methods and stochastic approximation. Journal of Computational and Graphical Statistics 21, 820.

O'Hagan, A., 1995. Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society 57, 99–138.

Opper, M., Archambeau, C., 2009. The variational gaussian approximation revisited. Neural Computation 21 (3), 786–792.

Ormerod, J. T., Wand, M. P., 2010. Explaining variational approximations. The American Statistician 64 (2), 140–153.
URL http://ideas.repec.org/a/bes/amstat/v64i2y2010p140-153.html

Ormerod, J. T., Wand, M. P., 2011. Gaussian variational approximate inference for generalized linear mixed models. Journal of Computational and Graphical Statistics forthcoming.

Paap, R., Franses, P., van Dijk, D., 2005. Does Africa grow slower than Asia, Latin America and the Middle East? evidence from a new data-based classification method. Journal of Development Economics 77, 553–570.

Paisley, J., Blei, D., Jordan, M., 2012. Variational Bayesian inference with stochastic search. In: ICML 2012.

Papyrakis, E., Gerlagh, R., 2004. The resource curse hypothesis and its transmission channels. Journal of Comparative Economics 32 (1), 181 – 193.

Paquet, U., Thomson, B., Winther, O., 2011. A hierarchical model for ordinal matrix factorization. Statistics and Computing 21 (3), 1–13.

Paterek, A., 2012. Predicting movie ratings and recommender systems - a monograph. http://arek-paterek.com/book/.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M., 2008. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining. pp. 569–577.

Qi, Y., Guo, Y., 2012. Message passing with relaxed moment matching. arXiv preprint arXiv:1204.4166.

Qi, Y. A., Minka, T. P., Picard, R. W., Ghahramani, Z., 2004. Predictive automatic relevance determination by expectation propagation. In: Proceedings of the twenty-first international conference on Machine learning. ICML '04. ACM, New York, NY, USA, pp. 85–.
URL http://doi.acm.org/10.1145/1015330.1015418

Quah, D., 1993. Galton's fallacy and tests of the convergence hypothesis. Scandinavian Journal of Economics 95 (4), 427–43.

Raftery, A., Madigan, D., Hoeting, J., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 179–191.

Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. MIT Press.

Raudenbush, S. W., Yang, M.-L., Yosef, M., 2000. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. Journal of Computational and Graphical Statistics 9 (1), 141–157.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics 80 (4), 647–657.

Richard, J.-F., Zhang, W., 2007. Efficient high-dimensional importance sampling. Journal of Econometrics 141 (2), 1385 – 1411.

URL http://www.sciencedirect.com/science/article/pii/S0304407607000486

Rivera-Batiz, L. A., Xie, D., 1993. Integration among unequals. Regional Science and Urban Economics 23 (3), 337 – 354.
URL http://www.sciencedirect.com/science/article/pii/016604629390051F

Robbins, H., Monro, S., 1951. A stochastic approximation method. The Annals of Mathematical Statistics 22 (3), 400–407.

Rogers, L. C. G., 2000. Evaluating first-passage probabilities for spectrally one-sided Lévy processes. Journal of Applied Probability 37 (4), 1173–1180.

Roy, N., McCallum, A., 2001. Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 441–448.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society, series B 71, 319–392.

Sachs, J. D., Warner, A. M., 2001. The curse of natural resources. European Economic Review 45 (4-6), 827–838.

Sala-i-Martin, X., 1997. I just ran two million regressions. American Economic Review 87(2), 178–183.

Sala-i-Martin, X., Doppelhofer, G., Miller, R. I., 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. American Economic Review 94(4), 813–835.

Salimans, T., 2012. Variable selection and functional form uncertainty in cross-country growth regressions. Journal of Econometrics 171 (2), 267 – 280.
URL http://www.sciencedirect.com/science/article/pii/S0304407612001546

Salimans, T., Fok, D., 2013. Approximate expectation-maximization for large non-gaussian state space models. Mimeo, Erasmus University Rotterdam, The Netherlands.

Salimans, T., Knowles, D., 2013. Fixed-form variational posterior approximation through stochastic linear regression. To appear in *Bayesian Analysis*.

Salimans, T., Paquet, U., Graepel, T., 2012. Collaborative learning of preference rankings. In: Proceedings of the sixth ACM conference on Recommender systems. RecSys '12. ACM, New York, NY, USA, pp. 261–264.
URL http://doi.acm.org/10.1145/2365952.2366009

Sandor, Z., Wedel, M., 2005. Heterogeneous conjoint choice designs. Journal of Marketing Research 42 (2), pp. 210–218.
URL http://www.jstor.org/stable/30164018

Saul, L., Jordan, M., 1996. Exploiting tractable substructures in intractable networks. Advances in Neural Information Processing Systems, 486–492.

Savage, L. J., 1962. The Foundations of Statistical Inference. London: Methuen.

Schweppe, F., 1965. Evaluation of likelihood functions for Gaussian signals. IEEE Transactions on Information Theory 11, 61–70.

Scott, J. G., Berger, J. O., 2010. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. Annals of Statistics 38 (5), 2587–2619.

Seeger, M. W., Nickisch, H., 2011. Fast convergent algorithms for expectation propagation approximate bayesian inference. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Ft. Lauderdale, FL, USA, pp. 652–660.

Settles, B., 1994. Active learning literature survey. Machine Learning 15 (2), 201–221.

Shephard, N., Pitt, M. K., 1997. Likelihood analysis of non-Gaussian measurement time series. Biometrika 84 (3), 653–667.

Singpurwalla, N. D., 1995. Survival in dynamic environments. Statistical Science 10 (1), 86–103.

Snyder, C., Bengtsson, T., Bickel, P., Anderson, J., 2008. Obstacles to High-dimensional Particle Filtering. Monthly Weather Review, 4629–4640.

Stern, D. H., Herbrich, R., Graepel, T., 2009. Matchbox: large scale online Bayesian recommendations. In: Proceedings of the 18th international conference on World wide web. pp. 111–120.

Stokey, N. L., 2009. The Economics of Inaction: Stochastic Control Models with Fixed Costs. Princeton University Press, Princeton, NJ.

Storkey, A. J., 2000. Dynamic trees: A structured variational method giving efficient propagation rules. In: Conference on Uncertainty in Artificial Intelligence (UAI).

Teh, Y., Newman, D., Welling, M., 2007. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Advances in Neural Information Processing Systems 19, 1353.

Tierney, L., Kadane, J. B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81 (393), 82–86.

Train, K. E., 2003. Discrete choice methods with simulation. Cambridge university press.

Turner, R. E., Berkes, P., Sahani, M., 2008. Two problems with variational expectation maximisation for time-series models. Inference and Estimation in Probabilistic Time-Series Models.

Vaupel, J. W., Manton, K. G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–454.

Verdinelli, I., Wasserman, L., 1995. Computing Bayes Factors using a generalization of the Savage-Dickey density ratio. Journal of the American Statistical Association 90, 614–618.

Vermeulen, B., Goos, P., Vandebroek, M., 2007. Rank-order conjoint experiments: Efficiency and design. Mimeo.
URL http://ssrn.com/abstract=1085999

Wainwright, M. J., Jordan, M. I., 2003. Graphical models, exponential families, and variational inference. Vol. 649. Now Publishers Inc. Hanover, MA, USA.

Wei, G., Tanner, M., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85, 699–704.

Winn, J., Bishop, C. M., 2006. Variational message passing. Journal of Machine Learning Research 6 (1), 661.

Wood, A., Berge, K., 1997. Exporting manufactures: Human resources, natural resources, and trade policy. Journal of Development Studies 34 (1), 35–59.

Woodbury, M. A., 1950. Inverting modified matrices. Memorandum Rept. 42, Princeton University.

Yai, T., Iwakura, S., Morichi, S., 1997. Multinomial probit with structured covariance for route choice behavior. Transportation Research Part B: Methodological 31 (3), 195–207.

Yanikkaya, H., 2003. Trade openness and economic growth: a cross-country empirical investigation. Journal of Development Economics 72 (1), 57 – 89.
URL http://www.sciencedirect.com/science/article/pii/S0304387803000683

Yashin, A., Manton, K., 1997. Effects of unobserved and partially observed covariate processes on system failure: A review of models and estimation strategies. Statistical Science 12, 20–34.

Yu, J., Goos, P., Vandebroek, M., 2009. Efficient conjoint choice designs in the presence of respondent heterogeneity. Marketing Science 28 (1), 122–135.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In: Goel, P., Zellner, A. (Eds.), Bayesian Inference and Decision Making: Essays in Honour of Bruno de Finetti. North-Holland, Amsterdam, pp. 233–243.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

511. S.J.J. KONIJN, *Empirical Studies on Credit Risk*
512. H. VRIJBURG, *Enhanced Cooperation in Corporate Taxation*
513. P. ZEPPINI, *Behavioural Models of Technological Change*
514. P.H. STEFFENS, *It's Communication, Stupid! Essays on Communication, Reputation and (Committee) Decision-Making*
515. K.C. YU, *Essays on Executive Compensation - Managerial Incentives and Disincentives*
516. P. EXTERKATE, *Of Needles and Haystacks: Novel Techniques for Data-Rich Economic Forecasting*
517. M. TYSZLER, *Political Economics in the Laboratory*
518. Z. WOLF, *Aggregate Productivity Growth under the Microscope*
519. M.K. KIRCHNER, *Fiscal Policy and the Business Cycle — The Impact of Government Expenditures, Public Debt, and Sovereign Risk on Macroeconomic Fluctuations*
520. P.R. KOSTER, *The cost of travel time variability for air and car travelers*
521. Y. ZU, *Essays of nonparametric econometrics of stochastic volatility*
522. B. KAYNAR, *Rare Event Simulation Techniques for Stochastic Design Problems in Markovian Setting*
523. P. JANUS, *Developments in Measuring and Modeling Financial Volatility*
524. F.P.W. SCHILDER, *Essays on the Economics of Housing Subsidies*
525. S.M. MOGHAYER, *Bifurcations of Indifference Points in Discrete Time Optimal Control Problems*
526. C. ÇAKMAKLI, *Exploiting Common Features in Macroeconomic and Financial Data*

527. J. LINDE, *Experimenting with new combinations of old ideas*
528. D. MASSARO, *Bounded rationality and heterogeneous expectations in macroeconomics*
529. J. GILLET, *Groups in Economics*
530. R. LEGERSTEE, *Evaluating Econometric Models and Expert Intuition*
531. M.R.C. BERSEM, *Essays on the Political Economy of Finance*
532. T. WILLEMS, *Essays on Optimal Experimentation*
533. Z. GAO, *Essays on Empirical Likelihood in Economics*
534. J. SWART, *Natural Resources and the Environment: Implications for Economic Development and International Relations*
535. A. KOTHIYAL, *Subjective Probability and Ambiguity*
536. B. VOOGT, *Essays on Consumer Search and Dynamic Committees*
537. T. DE HAAN, *Strategic Communication: Theory and Experiment*
538. T. BUSER, *Essays in Behavioural Economics*
539. J.A. ROSERO MONCAYO, *On the importance of families and public policies for child development outcomes*
540. E. ERDOGAN CIFTCI, *Health Perceptions and Labor Force Participation of Older Workers*
541. T.WANG, *Essays on Empirical Market Microstructure*
542. T. BAO, *Experiments on Heterogeneous Expectations and Switching Behavior*
543. S.D. LANSDORP, *On Risks and Opportunities in Financial Markets*
544. N. MOES, *Cooperative decision making in river water allocation problems*
545. P. STAKENAS, *Fractional integration and cointegration in financial time series*
546. M. SCHARTH, *Essays on Monte Carlo Methods for State Space Models*
547. J. ZENHORST, *Macroeconomic Perspectives on the Equity Premium Puzzle*
548. B. PELLOUX, *The Role of Emotions and Social Ties in Public On Good Games: Behavioral and Neuroeconomic Studies*
549. N. YANG, *Markov-Perfect Industry Dynamics: Theory, Computation, and Applications*
550. R.R. VAN VELDHUIZEN, *Essays in Experimental Economics*
551. X. ZHANG, *Modeling Time Variation in Systemic Risk*

552. H.R.A. KOSTER, *The internal structure of cities: the economics of agglomeration, amenities and accessibility.*
553. S.P.T. GROOT, *Agglomeration, globalization and regional labor markets: micro evidence for the Netherlands.*
554. J.L. MÖHLMANN, *Globalization and Productivity Micro-Evidence on Heterogeneous Firms, Workers and Products*
555. S.M. HOOGENDOORN, *Diversity and Team Performance: A Series of Field Experiments*
556. C.L. BEHRENS, *Product differentiation in aviation passenger markets: The impact of demand heterogeneity on competition*
557. G. SMRKOLJ, *Dynamic Models of Research and Development*
558. S. PEER, *The economics of trip scheduling, travel time variability and traffic information*
559. V. SPINU, *Nonadditive Beliefs: From Measurement to Extensions*
560. S.P. KASTORYANO, *Essays in Applied Dynamic Microeconometrics*
561. M. VAN DUIJN, *Location, choice, cultural heritage and house prices*