Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

For other titles published in this series, go to
http://www.springer.com/series/6991

Giovanni Petris • Sonia Petrone • Patrizia Campagnoli

Dynamic Linear Models with R

Springer

Giovanni Petris
Department of Mathematical Sciences
University of Arkansas
Fayetteville, AR 72701
USA
gpetris@uark.edu

Sonia Petrone
Department of Decision Sciences
Bocconi University
Via Roentgen, 1
20136 Milano
Italy
sonia.petrone@unibocconi.it

Patrizia Campagnoli
Department of Decision Sciences
Bocconi University
Via Roentgen, 1
20136 Milano
Italy
patrizia.campagnoli@unibocconi.it

ISBN 978-0-387-77237-0 e-ISBN 978-0-387-77238-7
DOI 10.1007/b135794
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009926480

A Mary e Giulio
G.P.

A Francesca, ai miei nipoti
S.P.

A Andrea e Michele
P.C.

Preface

This book aims at introducing the reader to statistical time series analysis by dynamic linear models. We have tried to be precise and rigorous in discussing the main concepts and tools, yet keeping a simple and friendly style of presentation. The main methods and models are widely illustrated with examples based on real data, implemented in R. Together with the book, we developed an R package for inference and forecasting with dynamic linear models; the dlm package is available as a contributed package in the Comprehensive R Archive Network at `http://www.r-project.org/`.

In the recent years, there has been an enormous growth of interest for statistical applications of dynamic linear models and, more generally, state-space models, in a wide range of applied fields, such as biology, economics, finance, marketing, quality control, engineering, demography, climatology, to mention only a few. State space models provide a very flexible yet fairly simple tool for analyzing dynamic phenomena and evolving systems, and have significantly contributed to extend the classical domains of application of statistical time series analysis to non-stationary, irregular processes, to systems evolving in continuous-time, to multivariate, continuous and discrete data. An extremely wide range of applied problems can be treated inside the framework of dynamic linear models or, more generally, state-space models.

The book covers the basic notions of dynamic linear models and state space models, the celebrated Kalman filter for estimation and forecasting in a dynamic linear model with known parameters, and maximum likelihood estimation. It also presents a wide array of specific dynamic linear models particularly suited for time series analysis, both for univariate and multivariate data. But these topics are of course also covered in other very good books in the rich literature on dynamic linear models, and several statistical softwares include packages for time series analysis through maximum likelihood and Kalman filtering. What we felt was somehow missing was an up to date, rigorous yet friendly reference—and software—for applied Bayesian time series analysis through dynamic linear models and state space models. This seemed to be missing despite the fact that the Bayesian approach has become

more and more popular in applications, due to the availability of modern and efficient computational tools. So, while also covering maximum likelihood methods, our focus in the book is on *Bayesian* time series analysis based on dynamic linear models.

We do not expect the reader to be an expert in Bayesian inference, so we begin with a short introduction to the Bayesian approach in Chapter 1. Also for a Bayesian reader, this is useful to set the notation and to underline some basic concepts that are used in the following chapters: for example, in presenting the simplest notions, such as Bayesian conjugate inference for a Gaussian model, we underline the recursive structure of the estimates, that will be one of the basic aspects of inference for dynamic linear models. Chapter 2 introduces the general setting of state space models and dynamic linear models, including the fundamental algorithms to sequentially update estimates and forecasts and the Kalman filter. Chapters 3 and 4 are in a sense the core of the book. In Chapter 3 the reader will find a discussion of a broad spectrum of specific models suited for the analysis of many kinds of data showing different features. Thus, Chapter 3 should be considered as a toolbox, illustrating a set of models from which the user can select the most appropriate for the application at hand. Chapter 4 covers maximum likelihood and Bayesian inference for dynamic linear models containing unknown parameters—which is always the case in practice. Many of the models introduced in Chapter 3 are discussed again there in this perspective. For most of the covered models we provide detailed examples of their use, corredated with the relevant R code. When possible, Bayesian estimates are evaluated using closed form algorithms. But in more elaborate models, analytical computations become intractable and simulation techniques are used to approximate the Bayesian solutions. We describe Markov chain Monte Carlo methods for Bayesian inference in dynamic linear models. The R package dlm provides functions for one of the basic steps in Bayesian computations in dynamic linear models, the so-called *forward filtering-backward sampling* algorithm, and other computational tools, with many examples, are provided. In Chapter 5, we present modern sequential Monte Carlo and particle filter algorithms for on-line estimation and forecasting.

Of course we cannot cover all of the extremely rich variety of models, applications, and problems in Bayesian inference with dynamic linear models, and many things will be missing. However, we hope to give a solid background on the main concepts and notions, leading the reader to acquire the skills for specific, personal elaborations, for which the flexibility of R and the dlm package will provide convenient, helpful tools. On the web site of the book, `definetti.uark.edu/~gpetris/dlm`, the reader will find data sets not included in the package and the code to run all the examples in the book. In addition, we plan to post there an updated list of errata.

The motivation for this book came from the authors' teaching experience in courses on time series analysis. We wanted to teach a course including—besides the classical ARMA models, descriptive techniques, exponential smoothing, and so on—more modern approaches, in particular Bayesian inference for time series through dynamic linear models. Again, we felt that a textbook, and a friendly but flexible software, were missing. So we started working on this project. We hope students, researchers, and practitioners will find the book and the software that resulted from our effort of some help.

We would like to thank Springer-Verlag's referees for their encouragement and valuable suggestions. Our thanks go also to our editor, John Kimmel, for his patience and support.

The dlm package would not exist without R, for which we thank R-core. Several people on `r-help`, the general R mailing list, have contributed their suggestions and feedback during the development of the package: we thank all of them. In particular, we thank Spencer Graves and Michael Lavine for their comments and suggestions on earlier versions of the package. Michael Lavine taught a course at the University of Massachusets using R and dlm from an early draft of the book, and we thank him for the valuable feedback he gave us. One of the authors (GP) taught some short courses based on preliminary versions of the book at Bocconi University and the University of Roma 3 and would like to thank Pietro Muliere, Carlo Favero, Julia Mortera, and his coauthor, Sonia Petrone, for the kind invitations and the hospitality. SP used draft versions of the book in her graduate courses on time series analysis at Bocconi University: students' feedback has been precious. We thank all our students at the University of Arkansas, Bocconi University, and the University of Roma 3 who, with their comments, questions, suggestions, interest and enthusiasm, have contributed to the development of this book. Among them, a special thanks goes to Paolo Bonomolo and Guido Morandini.

Needless to say, the responsibility for any remaining mistakes, obscurities, or omissions—in the book and in the package—lies solely with us.

Fayetteville, Arkansas
and
Milano, Italy
December 15, 2008

Giovanni Petris
Sonia Petrone
Patrizia Campagnoli

Contents

1

Introduction: basic notions about Bayesian inference

Dynamic linear models were developed in engineering in the early 1960's, to monitor and control dynamic systems, although pioneer results can be found in the statistical literature and go back to Thiele (1880). Early famous applications have been in the Apollo and Polaris aerospace programs (see, e.g., Hutchinson; 1984), but in the last decades dynamic linear models, and more generally state space models, have received an enormous impulse, with applications in an extremely vast range of fields, from biology to economics, from engineering and quality control to environmental studies, from geophysical science to genetics. This impressive growth of applications is largely due to the possibility of solving computational difficulties using modern Monte Carlo methods in a Bayesian framework. This book is an introduction to Bayesian modeling and forecasting of time series using dynamic linear models, presenting the basic concepts and techniques, and illustrating an R package for their practical implementation.

Statistical time series analysis using dynamic linear models was largely developed in the 1970-80's, and state space models are nowadays a focus of interest. In fact, the reader used to descriptive time series analysis or to ARMA models and Box–Jenkins model specification, may find the state space approach a bit difficult at first. But the powerful framework offered by dynamic linear models and state space models reveals to be a winning asset. ARMA models can be usefully regarded in terms of dynamic linear models. But dynamic linear models offer much more flexibility in treating non-stationary time series or modeling structural changes, and are often more easily interpretable; and the more general class of state space models extends the analysis to non-Gaussian and non-linear dynamic systems. There are, of course, different approaches to estimate dynamic linear models, via generalized least squares or maximum likelihood for example, but we believe that a Bayesian approach has several advantages, both methodological and computational. Kalman (1960) already underlines some basic concepts of dynamic linear models that we would say are proper to the Bayesian approach. A first step is moving from a deterministic to a stochastic system; the uncertainty,

G. Petris et al., *Dynamic Linear Models with R*, Use R, DOI: 10.1007/b135794_1,

which is always present due to forgotten variables, measurement errors, or imperfections, is described through probability. Consequently, the estimation of the quantities of interest (in particular, the state of the system at time t) is solved by computing their conditional distribution, given the available information. This is a general, basic concept in Bayesian inference. Dynamic linear models are based on the idea of describing the output of a dynamic system, for example a time series, as a function of a nonobservable state process (which has a simple, Markovian dynamics) affected by random errors. This way of modeling the temporal dependence in the data, by conditioning on latent variables, is simple and extremely powerful, and again it is quite natural in a Bayesian approach. Another crucial advantage of dynamic linear models is that computations can be done recursively: the conditional distributions of interest can be updated, incorporating the new data, without requiring the storage of all the past history. This is extremely advantageous when data arrive sequentially in time and on-line inference is required, and the reduction of the storage capacity needed becomes even more crucial for large data sets. The recursive nature of computations is a consequence of the Bayes formula in the framework of dynamic linear models.

However, analytical computations are often not manageable, but Markov chain Monte Carlo algorithms can be applied to state space models to overcome computational difficulties, and modern, sequential Monte Carlo methods, which have been enormously improved in the last years, are successfully used for on-line analysis.

We do not expect that the reader is already an expert in Bayesian statistics; therefore, before getting started, this chapter briefly reviews some basic notions, with a look to the concepts that are important in the study of dynamic linear models. Reference books on Bayesian statistics are Bernardo and Smith (1994), DeGroot (1970), Berger (1985), O'Hagan (1994), Robert (2001), Cifarelli and Muliere (1989), or Zellner (1971), Poirier (1995) and Geweke (2005) for a more econometric viewpoint.

1.1 Basic notions

In the analysis of real data, in economics, sociology, biology, engineering and in any field, we rarely have perfect information on the phenomenon of interest. Even when an accurate deterministic model describing the system under study is available, there is always something that is not under our control, such as effects of forgotten variables, measurement errors, or imperfections. We always have to deal with some uncertainty. A basic point in Bayesian statistics is that all the uncertainty that we might have on a phenomenon should be described by means of *probability*. In this perspective, probability has a *subjective* interpretation, being a way of formalizing the incomplete information that the researcher has about the events of interest. Probability

theory prescribes how to assign probabilities coherently, avoiding contradictions and undesirable consequences.

The Bayesian approach to the problem of "learning from experience" about a phenomenon moves from this crucial role played by probability. The learning process consists of the application of probability rules: one simply has to compute the *conditional probability* of the event of interest, given the experimental information. Bayes' theorem is the basic rule to be applied to this aim. Given two events A and B, probability rules say that the joint probability of A and B occurring is given by $\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A)$, where $\mathrm{P}(A|B)$ is the conditional probability of A given B and $\mathrm{P}(B)$ is the (marginal) probability of B. Bayes' theorem, or the theorem of inverse probability, is a simple consequence of the above equalities and says that

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}.$$

This is an elementary result that goes back to Thomas Bayes (who died in 1761). The importance of this theorem in Bayesian statistics is in the interpretation and scope of the inputs of the two sides of the equation, and in the role that, consequently, Bayes' theorem assumes for formalizing the inductive learning process. In Bayesian statistics, A represents the event of interest for the researcher and B an experimental result which she believes can provide information about A. Given $\mathrm{P}(A)$ and consequently $\mathrm{P}(\bar{A}) = 1 - \mathrm{P}(A)$, and having assigned the conditional probabilities $\mathrm{P}(B|A)$ and $\mathrm{P}(B|\bar{A})$ of the experimental fact B conditionally on A or $\bar{A}$, the problem of learning about A from the "experimental evidence" B is solved by computing the conditional probability $\mathrm{P}(A|B)$.

The event of interest and the experimental result depend on the problem. In statistical inference, the experimental fact is usually the result of a sampling procedure, and it is described by a random vector Y; it is common to use a parametric model to assign the probability law of Y, and the quantity of interest is the vector θ of the parameters of the model. Bayesian inference on θ consists of computing its conditional distribution given the sampling results. More specifically, suppose that, based on her knowledge of the problem, the researcher can assign a conditional distribution $\pi(y|\theta)$ for Y given θ, the *likelihood*, and a *prior distribution* $\pi(\theta)$ expressing her uncertainty on the parameter θ. Upon observing $Y = y$, we can use a generalization of the elementary Bayes' theorem, known as Bayes' formula, to compute the conditional density of θ given y:

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)},$$

where $\pi(y)$ is the marginal distribution of Y,

$$\pi(y) = \int \pi(y|\theta)\pi(\theta)\,\mathrm{d}\theta.$$

Thus, Bayesian statistics answers an inference problem by computing the relevant conditional distributions, and the Bayes formula is a basic tool to achieve this aim. It has an elegant, appealing coherence and simplicity. Differently from Bayesian procedures, frequentist statistical inference does not use a probability distribution for the unknown parameters, and inference on θ is based on the determination of estimators with good properties, confidence intervals, and hypothesis testing. The reason is that, since the value of the parameter θ does not "vary," θ is not interpretable as a random "variable" in a frequentist sense, neither can the probability that θ takes values in a certain interval have a frequentist interpretation. Adopting subjective probability instead, θ is a random quantity simply because its value is uncertain to the researcher, who should formalize the information she has about it by means of probability. This seems, indeed, quite natural. We refer the reader to the fundamental works by de Finetti (1970a,b) and Savage (1954) for a much deeper discussion.

In many applications, the main objective of a statistical analysis is *forecasting*; thus, the event of interest is the value of a future observation Y^*. Again, prediction of a future value Y^* given the data y is solved in the Bayesian approach simply by computing the conditional distribution of Y^* given $Y = y$, which is called *predictive distribution*. In parametric models it can be computed as

$$\pi(y^*|y) = \int \pi(y^*, \theta|y)\,\mathrm{d}\theta = \int \pi(y^*|y, \theta)\pi(\theta|y)\,\mathrm{d}\theta.$$

The last expression involves again the posterior distribution of θ. As a matter of fact, apart from controversies about frequentist or subjective probability, a difficulty with (prior or posterior) probability distributions on model parameters is that, in some problems, they do not have a clear physical interpretation, so that assigning to them a probability law is debatable, even from a subjective viewpoint. According to de Finetti, one can assign a probability only to "observable facts"; indeed, the ultimate goal of a statistical analysis is often forecasting the future observations rather than learning on unobservable parameters. Taking a *predictive approach*, the parametric model is to be regarded just as a tool to facilitate the task of specifying the probability law of the observable quantities and, eventually, of the predictive distribution. The choice of the prior should be suggested, in this approach, by predictive considerations, that is, by taking into account its implications on the probability law of Y. We discuss this point further in the next section.

Before moving on to the next, more technical, sections, let us introduce some notation and conventions that will be used throughout. Observable random variables or random vectors will be denoted by capital letters – most of the times by Y, possibly with a subscript. A possible value of the random variable or vector will be denoted by the corresponding lower-case letter. Note that we are not making any notational distinction between vectors and scalars, or between random variables and random vectors. This is true also when writing integrals. For example, $\int f(x)\,\mathrm{d}x$ denotes a univariate integral if

f is a function of one variable, but a multivariate integral if f is a function of a vector argument. The correct interpretation should be clear from the context. A univariate or multivariate *time series* is a sequence of random variables or vectors and will be denoted by $(Y_t : t = 1, 2, \dots)$, $(Y_t)_{t \geq 1}$, or just (Y_t) for short. When considering a finite sequence of consecutive observations, we will use the notation $Y_{r:s}$ for the observations between the rth and sth, both inclusive. Similarly, $y_{r:s}$ will denote a sequence of possible values for those observations. Probabiity densities will be generically denoted by $\pi(\cdot)$. We will adopt the sloppy but widespread convention of using the same symbol π for the distribution of different random variables: the argument will make clear what distribution we are referring to. For example, $\pi(\theta)$ may denote a prior distribution for the unknown parameter θ and $\pi(y)$ the marginal density of the data point Y. Appendix A contains the definitions of some common families of distributions. We are going to use the same symbol for a distribution and its density, in this case adding an extra argument. For example, $\mathcal{G}(a, b)$ denotes the gamma distribution with shape parameter a and rate parameter b, but $\mathcal{G}(y; a, b)$ denotes the density of that distribution at the point y. The k-dimensional normal distribution is $\mathcal{N}_k(m, C)$, but we will omit the subscript k whenever the dimension is clear from the context.

1.2 Simple dependence structures

Forecasting is one of the main tasks in time series analysis. A univariate or multivariate time series is described probabilistically by a sequence of random variables or vectors $(Y_t : t = 1, 2, \dots)$, where the index t denotes time. For simplicity, we will think of equally spaced time points (daily data, monthly data, and so on); for example, (Y_t) might describe the daily prices of m bonds, or monthly observations on the sales of a good. One basic problem is to make forecasts about the value of the next observation, Y_{n+1} say, having observed data up to time n, $Y_1 = y_1, \dots, Y_n = y_n$ or $Y_{1:n} = y_{1:n}$ for short. Clearly, the first step to this aim is to formulate reasonable assumptions about the dependence structure of the time series. If we are able to specify the probability law of the time series (Y_t), we know the joint densities $\pi(y_1, \dots, y_n)$ for any $n \geq 1$, and Bayesian forecasting would be solved by computing the predictive density

$$\pi(y_{n+1} | y_{1:n}) = \frac{\pi(y_{1:n+1})}{\pi(y_{1:n})}.$$

In practice, specifying the densities $\pi(y_1, \dots, y_n)$ directly is not easy, and one finds it convenient to make use of parametric models; that is, one often finds it simpler to express the probability law of $(Y_1, \dots, Y_n)$ conditionally on some characteristic θ of the data generating process. The relevant characteristic θ can be finite- or infinite-dimensional, that is, θ can be a random vector or, as is the case for state space models, a stochastic process itself. The researcher

often finds it simpler to specify the conditional density $\pi(y_{1:n}|\theta)$ of $Y_{1:n}$ given θ, and a density $\pi(\theta)$ on θ, then obtain $\pi(y_{1:n})$ as $\pi(y_{1:n}) = \int \pi(y_{1:n}|\theta)\pi(\theta)\,\mathrm{d}\theta$. We will proceed in this fashion when introducing dynamic linear models for time series analysis. But let's first study simpler dependence structures.

Conditional independence

The simplest dependence structure is conditional independence. In particular, in many applications it is reasonable to assume that $Y_1, \ldots, Y_n$ are conditionally independent and identically distributed (i.i.d.) given θ: $\pi(y_{1:n}|\theta) = \prod_{i=1}^{n} \pi(y_i|\theta)$. For example, if the Y_i's are repeated measurements affected by a random error, we are used to think of a model of the kind $Y_i = \theta + \epsilon_i$, where the ϵ_i's are independent Gaussian random errors, with mean zero and variance σ^2 depending on the precision of the measurement device. This means that, conditionally on θ, the Y_i's are i.i.d., with $Y_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$.

Note that $Y_1, Y_2, \ldots$ are only conditionally independent: the observations $y_1, \ldots, y_n$ provide us information about the unknown value of θ and, through θ, on the value of the next observation Y_{n+1}. Thus, Y_{n+1} depends, in a probabilistic sense, on the past observations $Y_1, \ldots, Y_n$. The predictive density in this case can be computed as

$$\begin{aligned}\pi(y_{n+1}|y_{1:n}) &= \int \pi(y_{n+1}, \theta|y_{1:n})\,\mathrm{d}\theta \\ &= \int \pi(y_{n+1}|\theta, y_{1:n})\pi(\theta|y_{1:n})\,\mathrm{d}\theta \\ &= \int \pi(y_{n+1}|\theta)\pi(\theta|y_{1:n})\,\mathrm{d}\theta,\end{aligned}$$

the last equality following from the assumption of conditional independence, where $\pi(\theta|y_{1:n})$ is the posterior density of θ, conditionally on the data $(y_1, \ldots, y_n)$. As we have seen, the posterior density can be computed by the Bayes formula:

$$\pi(\theta|y_{1:n}) = \frac{\pi(y_{1:n}|\theta)\pi(\theta)}{\pi(y_{1:n})} \propto \prod_{t=1}^{n} \pi(y_t|\theta)\ \pi(\theta)\ . \tag{1.1}$$

Note that the marginal density $\pi(y_{1:n})$ does not depend on θ, having the role of normalizing constant, so that the posterior is proportional to the product of the likelihood and the prior[1].

It is interesting to note that, with the assumption of conditional independence, the posterior distribution can be computed *recursively*. This means that one does not need all the previous data to be kept in storage and reprocessed every time a new measurement is taken. In fact, at time $(n-1)$, the information available about θ is described by the conditional density

[1] The symbol $\propto$ means "proportional to".

$$\pi(\theta|y_{1:n-1}) \propto \prod_{t=1}^{n-1} \pi(y_t|\theta)\pi(\theta),$$

so that this density plays the role of prior at time n. Once the new observation y_n becomes available, we have just to compute the likelihood, which is $\pi(y_n|\theta, y_{1:n-1}) = \pi(y_n|\theta)$ by the assumption of conditional independence, and update the "prior" $\pi(\theta|y_{1:n-1})$ by the Bayes rule, obtaining

$$\pi(\theta|y_{1:n-1}, y_n) \propto \pi(\theta|y_{1:n-1})\pi(y_n|\theta) \propto \prod_{t=1}^{n-1} \pi(y_t|\theta)\pi(\theta)\pi(y_n|\theta),$$

which is (1.1). The recursive structure of the posterior will play a crucial role when we study dynamic linear models and the Kalman filter in the next chapters.

To illustrate the idea, let us use a simple example. Suppose that, after a wreck in the ocean, you landed on a small island, and let θ denote your position, the distance from the coast, say. When studying dynamic linear models, we will consider the case when θ is subject to change over time (you are on a life boat in the ocean and not on an island, so that you slowly move with the stream and the waves, being at distance θ_t from the coast at time t). However, for the moment let's consider θ as fixed. Luckily, you can see the coast at times; you have some initial idea of your position θ, but you are clearly interested in learning more about θ based on the measurements y_t that you can take. Let us formalize the learning process in the Bayesian approach.

The measurements Y_t can be modeled as

$$Y_t = \theta + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

where the ϵ_t's and θ are independent and, for simplicity, σ^2 is a known constant. In other words:

$$Y_1, Y_2, \ldots|\theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2).$$

Suppose you agree to express your prior idea about θ as

$$\theta \sim \mathcal{N}(m_0, C_0),$$

where the prior variance C_0 might be quite large if you are very uncertain about your guess m_0. Given the measurements $y_{1:n}$, you update your opinion about θ computing its posterior density, using the Bayes formula. We have

$$
\begin{aligned}
\pi(\theta|y_{1:n}) &\propto \text{ likelihood} \times \text{prior} \\
&= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_t-\theta)^2\right\} \frac{1}{\sqrt{2\pi C_0}} \exp\left\{-\frac{1}{2C_0}(\theta-m_0)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{n} y_t^2 - 2\theta\sum_{t=1}^{n} y_t + n\theta^2\right) - \frac{1}{2C_0}(\theta^2 - 2\theta m_0 + m_0^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2 C_0}\left((nC_0+\sigma^2)\theta^2 - 2(nC_0\bar{y} + \sigma^2 m_0)\theta\right)\right\}.
\end{aligned}
$$

The above expression might appear complicated, but in fact it is the kernel of a Normal density. Note that, if $\theta \sim \mathcal{N}(m, C)$, then $\pi(\theta) \propto \exp\{-(1/2C)(\theta^2 - 2m\theta)\}$; so, writing the above expression as

$$
\exp\left\{-\frac{1}{2\sigma^2 C_0/(nC_0+\sigma^2)}\left(\theta^2 - 2\frac{nC_0\bar{y}+\sigma^2 m_0}{(nC_0+\sigma^2)}\theta\right)\right\},
$$

we recognize that

$$
\theta|y_{1:n} \sim \mathcal{N}(m_n, C_n),
$$

where

$$
m_n = \mathrm{E}(\theta|y_{1:n}) = \frac{C_0}{C_0+\sigma^2/n}\bar{y} + \frac{\sigma^2/n}{C_0+\sigma^2/n}m_0 \tag{1.2a}
$$

and

$$
C_n = \mathrm{Var}(\theta|y_{1:n}) = \left(\frac{n}{\sigma^2} + \frac{1}{C_0}\right)^{-1} = \frac{\sigma^2 C_0}{\sigma^2 + nC_0}. \tag{1.2b}
$$

The posterior *precision* is $1/C_n = n/\sigma^2 + 1/C_0$, and it is the sum of the precision n/σ^2 of the sample mean and the initial precision $1/C_0$. The posterior precision is always larger than the initial precision: even data of poor quality provide some information. The posterior expectation $m_n = \mathrm{E}(\theta|y_{1:n})$ is a weighted average between the sample mean $\bar{y} = \sum_{i=1}^{n} y_i/n$ and the prior guess $m_0 = \mathrm{E}(\theta)$, with weights depending on C_0 and σ^2. If the prior uncertainty, represented by C_0, is small compared to σ^2, the prior guess receives more weight. If C_0 is very large, then $m_n \simeq \bar{y}$ and $C_n \simeq \sigma^2/n$.

As we have seen, the posterior distribution can be computed recursively. At time n, the conditional density $\mathcal{N}(m_{n-1}, C_{n-1})$ of θ given the previous data $y_{1:n-1}$ plays the role of prior, and the likelihood for the current observation is

$$
\pi(y_n|\theta, y_{1:n-1}) = \pi(y_n|\theta) = \mathcal{N}(y_n; \theta, \sigma^2).
$$

We can update the prior $\mathcal{N}(m_{n-1}, C_{n-1})$ on the basis of the observation y_n using (1.2), with m_{n-1} and C_{n-1} in place of m_0 and C_0. We see that the resulting posterior density is Gaussian, with parameters

$$
\begin{aligned}
m_n &= \frac{C_{n-1}}{C_{n-1}+\sigma^2}\, y_n + \left(1 - \frac{C_{n-1}}{C_{n-1}+\sigma^2}\right)\, m_{n-1} \\
&= m_{n-1} + \frac{C_{n-1}}{C_{n-1}+\sigma^2}(y_n - m_{n-1})
\end{aligned} \tag{1.3a}
$$

and variance

$$C_n = \left(\frac{1}{\sigma^2} + \frac{1}{C_{n-1}}\right)^{-1} = \frac{\sigma^2 C_{n-1}}{\sigma^2 + C_{n-1}}. \tag{1.3b}$$

Since $Y_{n+1} = \theta + \epsilon_{n+1}$, the *predictive distribution* of $Y_{n+1}|y_{1:n}$ is Normal, with mean m_n and variance $C_n + \sigma^2$; thus, m_n is the posterior expected value of θ and also the one-step-ahead "point prediction" $\mathrm{E}(Y_{n+1}|y_{1:n})$. Expression (1.3a) shows that m_n is obtained by correcting the previous estimate m_{n-1} by a term that takes into account the forecast error $e_n = y_n - m_{n-1}$, weighted by

$$\frac{C_{n-1}}{C_{n-1} + \sigma^2} = \frac{C_0}{\sigma^2 + nC_0}. \tag{1.4}$$

As we shall see in Chapter 2, this "prediction-error correction" structure is typical, more generally, of the formulae of the Kalman filter for dynamic linear models.

Exchangeability

Exchangeability is the basic dependence structure in Bayesian analysis. Consider again an infinite sequence $(Y_t : t = 1, 2, \ldots)$ of random vectors. Suppose that the order in the sequence is not relevant, in the sense that, for any $n \geq 1$, the vector $(Y_1, \ldots, Y_n)$ and any permutation of its components, $(Y_{i_1}, \ldots, Y_{i_n})$, have the same distribution. In this case, we say that the sequence $(Y_t : t = 1, 2, \ldots)$ is *exchangeable*. This is a reasonable assumption when the Y_t's represent the results of experiments repeated under similar conditions. In the example of the previous paragraph, it is quite natural to consider that the order in which the measurements Y_t of the distance from the coast are taken is not relevant. There is an important result, known as de Finetti's representation theorem, that shows that the assumption of exchangeability is equivalent to the assumption of conditional independence and identical distribution that we have discussed in the previous paragraph. There is, however, an important difference. As you can see, here we move from a quite natural assumption on the dependence structure of the observables, that is, exchangeability; we have not introduced, up to now, parametric models or prior distributions on parameters. In fact, the hypothetical model, that is the pair likelihood and prior, arises from the assumption of exchangeability, as shown by the representation theorem.

Theorem 1.1. ***(de Finetti representation theorem).*** *Let* $(Y_t : t = 1, 2, \ldots)$ *be an infinite sequence of exchangeable random vectors. Then*

1. *With probability one, the sequence of empirical distribution functions*

$$F_n(y) = F_n(y; Y_1, \ldots, Y_n) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, y]}(Y_i)$$

converges weakly to a random distribution function F*, as* $n \to \infty$*;*

2. for any $n \geq 1$, the distribution function of $(Y_1, \ldots, Y_n)$ can be represented as

$$\mathrm{P}(Y_1 \leq y_1, \ldots, Y_n \leq y_n) = \int \prod_{i=1}^{n} \pi(y_i) \, \mathrm{d}\pi(F)$$

where π is the probability law of the weak limit F of the sequence of the empirical distribution functions.

The fascinating aspect of the representation theorem is that the hypothetical model results from the assumptions on the dependence structure of the observable variables (Y_t). If we assume that the sequence (Y_t) is exchangeable, then we can think of them as i.i.d. conditionally on the distribution function (d.f.) F, with common d.f. F. The random d.f. F is the weak limit of the empirical d.f.'s. The prior distribution π (also called, in this context, de Finetti measure) is a probability law on the space $\mathcal{F}$ of all the d.f.s on the sample space $\mathcal{Y}$ and expresses our beliefs on the limit of the empirical d.f.s. In many problems we can restrict the support of the prior to a parametric class $\mathcal{P}_\Theta = \{\pi(\cdot|\theta), \theta \in \Theta\} \subset \mathcal{F}$, where $\Theta \subseteq \mathbb{R}^p$; in this case the prior is said *parametric*. We see that, in the case of a parametric prior, the representation theorem implies that $Y_1, Y_2, \ldots$ are conditionally i.i.d., given θ, with common d.f. $\pi(\cdot|\theta)$, and θ has a prior distribution $\pi(\theta)$. This is the conditional i.i.d. dependence structure that we have discussed in the previous subsection.

Heterogeneous data

Exchangeability is the simplest dependence structure, which allows us to enlighten the basic aspects of Bayesian inference. It is appropriate when we believe that the data are homogeneous. However, in many problems the dependence structure is more complex. Often, it is appropriate to allow some heterogeneity among the data, assuming that

$$Y_1, \ldots, Y_n | \theta_1, \ldots, \theta_n \sim \prod_{t=1}^{n} f_t(y_t|\theta_t),$$

that is, $Y_1, \ldots, Y_n$ are conditionally independent given a vector $\theta = (\theta_1, \ldots, \theta_n)$, with Y_t depending only on the corresponding θ_t. For example, Y_t could be the expense of customer t for some service, and we might assume that each customer has a different average expense θ_t, introducing heterogeneity, or "random effects," among customers. In other applications, t might denote time; for example, each Y_t could represent the average sales in a sample of stores, at time t; and we might assume that $Y_t|\theta_t \sim \mathcal{N}(\theta_t, \sigma^2)$, with θ_t representing the expected sales at time t.

In these cases, the model specification is completed by assigning the probability law of the vector $(\theta_1, \ldots, \theta_n)$. For modeling random effects, a common assumption is that $\theta_1, \ldots, \theta_n$ are i.i.d. according to a distribution G. If there

is uncertainty about G, we can model $\theta_1, \ldots, \theta_n$ as conditionally i.i.d. given G, with common distribution function G, and assign a prior on G.

If $(Y_t : t = 1, 2, \ldots)$ is a sequence of observations over time, then the assumption that the θ_t's are i.i.d., or conditionally i.i.d., is generally not appropriate, since we want to introduce a temporal dependence among them. As we shall see in Chapter 2, in state space models we assume a Markovian dependence structure among the θ_t's.

We will return to this problem in the next section.

1.3 Synthesis of conditional distributions

We have seen that Bayesian inference is simply solved, in principle, by computing the conditional probability distributions of the quantities of interest: the posterior distribution of the parameters of the model, or the predictive distribution. However, especially when the quantity of interest is multivariate, one might want to present a summary of the posterior or predictive distribution. Consider the case of inference on a multivariate parameter $\theta = (\theta_1, \ldots, \theta_p)$. After computing the joint posterior distribution of θ, if some elements of θ are regarded as nuisance parameters, one can integrate them out to obtain the (marginal) posterior of the parameters of interest. For example, if $p = 2$, we can marginalize the joint posterior $\pi(\theta_1, \theta_2|y)$ and compute the marginal posterior density of θ_1:

$$\pi(\theta_1|y) = \int \pi(\theta_1, \theta_2|y)\, \mathrm{d}\theta_2.$$

We can provide a graphical representation of the marginal posterior distributions, or some summary values, such as the posterior expectations $\mathrm{E}(\theta_i|y)$ or the posterior variances $\mathrm{Var}(\theta_i|y)$, and so on. We can also naturally show intervals (usually centered on $\mathrm{E}(\theta_i|y)$) or bands with high posterior probability.

The choice of a summary of the posterior distribution (or of the predictive distribution) can be more formally regarded as a decision problem. In a statistical decision problem we want to choose an *action* in a set $\mathcal{A}$, called the action space, on the basis of the sample y. The consequences of action a are expressed through a *loss function* $L(\theta, a)$. Given the data y, a *Bayesian decision rule* selects an action in $\mathcal{A}$ that minimizes the conditional expected loss, $\mathrm{E}(L(\theta, a)|y) = \int L(\theta, a)\pi(\theta|y)\, \mathrm{d}\theta$. Bayesian point estimation can be seen as a decision problem in which the action space coincides with the parameter space. The choice of the loss function depends on the problem at hand, and, of course, different loss functions give rise to different Bayes estimates of θ. Some commonly used loss functions are briefly discussed below.

Quadratic loss. Let θ be a scalar. A common choice is a quadratic loss function $L(\theta, a) = (\theta - a)^2$. Then the posterior expected loss is $\mathrm{E}((\theta - a)^2|y)$, which is minimized at $a = \mathrm{E}(\theta|y)$. So, the Bayes estimate of θ with

quadratic loss is the posterior expected value of θ. If θ is p-dimensional, a quadratic loss function is expressed as $L(\theta, a) = (\theta - a)'H(\theta - a)$, for a symmetric positive definite matrix H. Then the Bayes estimate of θ is the vector of posterior expectations $\mathrm{E}(\theta|y)$.

Linear loss. If θ is scalar and

$$L(\theta, a) = \begin{cases} c_1 \mid a - \theta \mid & \text{if } a \leq \theta, \\ c_2 \mid a - \theta \mid & \text{if } a > \theta, \end{cases}$$

where c_1 and c_2 are positive constants, then the Bayes estimate is the $c_1/(c_1 + c_2)$ quantile of the posterior distribution. As a special case, if $c_1 = c_2$, the Bayes estimate is a posterior median.

Zero-one loss. If θ is a discrete random variable and

$$L(\theta, a) = \begin{cases} c & \text{if } a \neq \theta, \\ 0 & \text{if } a = \theta, \end{cases}$$

then the Bayes estimate is a mode of the posterior distribution.

For example, if $Y_1, \ldots, Y_n|\theta$ are i.i.d. with $Y_t|\theta \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(m_0, C_0)$, the posterior density is $\mathcal{N}(m_n, C_n)$, where m_n and C_n are given by (1.2). The Bayes estimate of θ, adopting a quadratic loss, is $\mathrm{E}(\theta|y_{1:n}) = m_n$, a weighted average between the prior guess m_0 and the sample mean $\bar{y}$. Note that, if the sample size is large, then the weight of the prior guess decreases to zero, and the posterior density concentrates around $\bar{y}$, which is the maximum likelihood estimate (MLE) of θ.

This asymptotic behavior of the posterior density holds more generally. Let $(Y_t : t = 1, 2, \ldots)$ be a sequence of conditionally i.i.d. random vectors, given θ, with $Y_t|\theta \sim \pi(y|\theta)$ and $\theta \in \mathbb{R}^p$ having prior distribution $\pi(\theta)$. Under general assumptions, it can be proved that the posterior distribution $\pi(\theta|y_1, \ldots, y_n)$, for n large, can be approximated by a Normal density centered at the MLE $\hat{\theta}_n$. This implies that, in these cases, Bayesian and frequentist estimates tend to agree for a sufficiently large sample size. For a more rigorous discussion of asymptotic normality of the posterior distribution, see Bernardo and Smith (1994, Section 5.3), or Schervish (1995, Section 7.4).

As a second example, linking Bayes estimators and classical decision theory, consider the problem of estimating the mean of a multivariate Normal distribution. In its simplest formulation, the problem is as follows. Suppose that $Y_1, \ldots, Y_n$ are independent r.v.s, with $Y_t \sim \mathcal{N}(\theta_t, \sigma^2), t = 1, \ldots, n$, where σ^2 is a known constant. This is the case of heterogeneous data, discussed in Section 1.2. For instance, the Y_t's could be sample means, in n independent experiments; however, note that here $\theta = (\theta_1, \ldots, \theta_n)$ is regarded as a vector of unknown constants. Thus we have

$$Y = (Y_1, \ldots, Y_n) \sim \mathcal{N}_n(\theta, \sigma^2 I_n),$$

where I_n denotes the n-dimensional identity matrix, and the problem is estimating the mean vector θ. The MLE of θ, which is also the uniform minimum variance unbiased estimator, is given by the vector of sample means:

$\hat{\theta} = \hat{\theta}(Y) = Y$. However, an important result, which had a great impact when Stein proved it in 1956, shows that the MLE is not optimal with respect to the quadratic loss function $L(\theta, a) = (\theta - a)'(\theta - a)$ if $n \geq 3$. The overall expected loss, or mean square error, of $\hat{\theta}$ is

$$\mathrm{E}\left((\theta - \hat{\theta}(Y))'(\theta - \hat{\theta}(Y))\right) = \mathrm{E}\left(\sum_{t=1}^{n}(\theta_t - \hat{\theta}_t(Y))^2\right)$$

where the expectation is with respect to the density $\pi_\theta(y)$, i.e., the $\mathcal{N}_n(\theta, \sigma^2 I_n)$ distribution of the data. Stein (1956) proved that, if $n \geq 3$, there exists another estimator $\theta^* = \theta^*(Y)$, which is more efficient than the MLE $\hat{\theta}$ in the sense that

$$\mathrm{E}\big((\theta - \theta^*(Y))'(\theta - \theta^*(Y))\big) < \mathrm{E}\big((\theta - \hat{\theta}(Y))'(\theta - \hat{\theta}(Y))\big)$$

for every θ. For $\sigma^2 = 1$, the Stein estimator is given by $\theta^*(Y) = (1 - (n - 2)/Y'Y)Y$; it shrinks the sample means $Y = (Y_1, \ldots, Y_n)$ towards zero. More generally, *shrinkage estimators* shrink the sample means towards the overall mean $\bar{y}$, or towards different values. Note that the MLE of θ_t, that is $\hat{\theta}_t = Y_t$, does not make use of the data Y_j, for $j \neq t$, which come from the other independent experiments. Thus, Stein's result seems quite surprising, showing that a more efficient estimator of θ_t can be obtained using the information from "independent" experiments. Borrowing strength from different experiments is in fact quite natural in a Bayesian approach. The vector θ is regarded as a random vector, and the Y_t's are *conditionally* independent given $\theta = (\theta_1, \ldots, \theta_n)$, with $Y_t|\theta_t \sim \mathcal{N}(\theta_t, \sigma^2)$, that is

$$Y|\theta \sim \mathcal{N}_n(\theta, \sigma^2 I_n).$$

Assuming a $\mathcal{N}_n(m_0, C_0)$ prior density for θ, the posterior density is $\mathcal{N}_n(m_n, C_n)$ where

$$m_n = (C_0^{-1} + \sigma^{-2} I_n)^{-1}(C_0^{-1} m_0 + \sigma^{-2} I_n y)$$

and $C_n = (C_0^{-1} + \sigma^{-2} I_n)^{-1}$. Thus the posterior expectation m_n provides a shrinkage estimate, shrinking the sample means towards the value m_0. Clearly, the shrinkage depends on the choice of the prior; see Lindley and Smith (1972).

Similarly to a Bayes point estimate, a Bayes point forecast of Y_{n+1} given $y_{1:n}$ is a synthesis of the predictive density with respect to a loss function, which expresses consequences of the forecast error of predicting Y_{n+1} with a value $\hat{y}$, say. With the quadratic loss function, $L(y_{n+1}, \hat{y}) = (y_{n+1} - \hat{y})^2$, the Bayes forecast is the expected value $\mathrm{E}(Y_{n+1}|y_{1:n})$.

Again, point estimation or forecasting is coherently treated in the Bayesian approach on the basis of statistical decision theory. However, in practice the computation of Bayes estimates or forecasts can be difficult. If θ is multivariate and the model structure complex, posterior expectations or, more generally,

integrals of the kind $\int g(\theta)\pi(\theta|y)d\theta$, can be analytically untractable. In fact, despite its attractive theoretical and conceptual coherence, the diffusion of Bayesian statistics in applied fields has been hindered, in the past, by computational difficulties, which had restricted the availability of Bayesian solutions to rather simple problems. As we shall see in Section 1.6, these difficulties can be overcome by the use of modern simulation techniques.

1.4 Choice of the prior distribution

The explicit use of prior information, besides the information from the data, is a basic aspect of Bayesian inference. Indeed, some prior knowledge of the phenomenon under study is always needed: data never speak entirely by themselves. The Bayesian approach allows us to explicitly introduce all the information we have (from experts' opinions, from previous studies, from the theory, and from the data) in the inferential process. However, the choice of the prior can be a delicate point in practical applications. Here we briefly summarize some basic notions, but first let us underline a fundamental point, which is clearly enlightened in the case of exchangeable data: the choice of a prior is in fact the choice of the *pair* $\pi(y|\theta)$ and $\pi(\theta)$. Often, the choice of $\pi(y|\theta)$ is called *model specification*, but in fact it is part, with the specification of $\pi(\theta)$, of the subjective choices that we have to do in order to study a phenomenon, based of our prior knowledge. At any rate, given $\pi(y|\theta)$, the prior $\pi(\theta)$ should be an honest expression of our beliefs about θ, with no mathematical restrictions on its form.

That said, there are some practical aspects that deserve some consideration. For computational convenience, it is common practice to use *conjugate priors*. A family of densities on θ is said to be conjugate to the model $\pi(y|\theta)$ if, whenever the prior belongs to that family, so does the posterior. In the example in Section 1.2, we used a Gaussian prior density $\mathcal{N}(m_0, C_0)$ on θ, and the posterior resulted still Gaussian, with updated parameters, $\mathcal{N}(m_n, C_n)$; thus, the Gaussian family is conjugate to the model $\pi(y|\theta) = \mathcal{N}(y; \theta, \sigma^2)$ (with σ^2 known). In general, a prior will be conjugate when it has the same analytic form of the likelihood, regarded as a function of θ. Clearly this definition does not determine uniquely the conjugate prior for a model $\pi(y|\theta)$. For the exponential family, we have a more precise notion of *natural conjugate prior*, which is defined from the density of the sufficient statistics; see for example Bernardo and Smith (1994, Section 5.2). Natural conjugate priors for the exponential family can be quite rigid in the multivariate case, and *enriched* conjugate priors have been proposed (Brown et al.; 1994; Consonni and Veronese; 2001). Furthermore, it can be proved that any prior for an exponential family parameter can be approximated by a mixture of conjugate priors (Dalal and Hall; 1983; Diaconis and Ylvisaker; 1985). We provide some examples below and in the next section. Anyway, computational ease has become less strin-

gent in recent years, due to the availability of simulation-based approximation techniques.

In practice, people quite often use *default priors* or *non-informative priors*, for expressing a situation of "prior ignorance" or vague prior information. The problem of appropriately defining the idea of "prior ignorance," or of a prior with "minimal effect" relative to the data on the inferential results, has a long history and is quite delicate; see Bernardo and Smith (1994, Section 5.6.2) for a detailed treatment; or also O'Hagan (1994) or Robert (2001). If the parameter θ takes values in a finite set, $\{\theta_1^*, \ldots, \theta_k^*\}$ say, then the classical notion of a non-informative prior, since Bayes (1763) and Laplace (1814), is that of a uniform distribution, $\pi(\theta_j^*) = 1/k$. However, even in this simple case it can be shown that care is needed in defining the quantity of interest (see Bernardo and Smith; 1994). Anyway, extending the notion of a uniform prior when the parameter space is infinite clearly leads to *improper* distributions, which cannot be regarded as probability distributions. For example, if $\theta \in (-\infty, +\infty)$, a uniform prior would be a constant, and its integral on the real line would be infinite. Furthermore, a uniform distribution for θ implies a nonuniform distribution for any nonlinear monotone transformation of θ, and thus the Bayes–Laplace postulate is inconsistent in the sense that, intuitively, "ignorance about θ" should also imply "ignorance" about one-to-one transformations of it. Priors based on invariance considerations are Jeffreys priors (Jeffreys; 1998). Widely used are also *reference priors*, suggested by Bernardo (1979a,b) on an information-decisional theoretical base (see for example Bernardo and Smith; 1994, Section 5.4). The use of improper priors is debatable, but often the posterior density from an improper prior turns out to be proper, and improper priors are anyway widely used, also for reconstructing frequentist results in a Bayesian framework. For example, if $Y_t|\theta$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$, using an improper uniform prior $\pi(\theta) = c$ and formally applying Bayes' formula gives

$$\pi(\theta|y_{1:n}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t-\theta)^2\right\} \propto \exp\left\{-\frac{n}{2\sigma^2}(\theta^2-2\theta\bar{y})^2\right\},$$

that is, the posterior is $\mathcal{N}(\bar{y}, \sigma^2/n)$. In this case, the Bayes point estimate under quadratic loss is $\bar{y}$, which is also the MLE of θ. As we noted before, starting with a proper Gaussian prior would give a posterior density centered around the sample mean only if the prior variance C_0 is very large compared to σ^2, or if the sample size n is large.

Another common practice is to have a hierarchical specification of the prior density. This means assuming that θ has density $\pi(\theta|\lambda)$ conditionally on some hyperparameter λ, and then a prior $\pi(\lambda)$ is assigned to λ. This is often a way for expressing a kind of uncertainty in the choice of the prior density. Clearly, this is equivalent to the prior $\pi(\theta) = \int \pi(\theta|\lambda)\pi(\lambda)\,\mathrm{d}\lambda$.

In order to avoid theoretical and computational difficulties related to the use of improper priors, in this book we will use only proper priors. It is im-

portant, however, to be aware of the effect of the prior on the analysis. This can be assessed using sensitivity analysis, which, in one of its basic forms, may simply consist in comparing the inferences resulting from different prior hyperparameters.

We conlude this section with an important example of conjugate prior. In Section 1.2 we considered conjugate Bayesian analysis for the mean of a Gaussian population, with known variance. Let now $Y_1, \ldots, Y_n|\theta, \sigma^2$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$, where both θ and σ^2 are unknown. It is convenient to work with the *precision* $\phi = 1/\sigma^2$ rather than with the variance σ^2. A conjugate prior for (θ, ϕ) can be obtained noting that the likelihood can be written as

$$\pi(y_{1:n}|\theta, \phi) \propto \phi^{(n-1)/2} \exp\left\{-\frac{1}{2}\phi\, n s^2\right\}\ \phi^{1/2} \exp\left\{-\frac{n}{2}\phi(\mu - \bar{y})^2\right\}$$

where $\bar{y}$ is the sample mean and $s^2 = \sum_{t=1}^{n}(y_i - \bar{y})^2/n$ is the sample variance (add and subtract $\bar{y}$ in the squared term and note that the cross product is zero). We see that, as a function of (θ, ϕ), the likelihood is proportional to the kernel of a Gamma density in ϕ, with parameters $(n/2 + 1, ns^2/2)$ times the kernel of a Normal density in θ, with parameters $(\bar{y}, (n\phi)^{-1})$. Therefore, a conjugate prior for (θ, σ^2) is such that ϕ has a Gamma density with parameters (a, b) and, conditionally on ϕ, θ has a Normal density with parameters $(m_0, (n_0\phi)^{-1})$. The joint prior density is

$$\begin{aligned}\pi(\theta, \phi) &= \pi(\phi)\ \pi(\theta|\phi) = \mathcal{G}(\phi; a, b)\ \mathcal{N}(\theta; m_0, (n_0\phi)^{-1}) \\ &\propto \phi^{a-1} \exp\{-b\phi\}\ \phi^{1/2} \exp\left\{-\frac{n_0}{2}\phi(\theta - m_0)^2\right\},\end{aligned}$$

which is a Normal-Gamma, with parameters $(m_0, (n_0)^{-1}, a, b)$ (see Appendix A). In particular, $\mathrm{E}(\theta|\phi) = m_0$ and $\mathrm{Var}(\theta|\phi) = (n_0\phi)^{-1} = \sigma^2/n_0$, that is, the variance of θ, given σ^2, is expressed as a proportion $1/n_0$ of σ^2. Marginally, the variance $\sigma^2 = \phi^{-1}$ has an Inverse Gamma density, with $\mathrm{E}(\sigma^2) = b/(a-1)$, and it can be shown that

$$\theta \sim \mathcal{T}(m_0, (n_0\, a/b)^{-1}, 2a),$$

a Student-t with parameters $m_0, (n_0\, a/b)^{-1}$ and $2a$ degrees of freedom, with $\mathrm{E}(\theta) = \mathrm{E}(\mathrm{E}(\theta|\phi)) = m_0$ and $\mathrm{Var}(\theta) = \mathrm{E}(\sigma^2)/n_0 = (b/(a-1))/n_0$.

With a conjugate Normal-Gamma prior, the posterior of (θ, ϕ) is still Normal-Gamma, with updated parameters. In order to show this, we have to do some calculations. Start with

$$\begin{aligned}&\pi(\theta, \phi|y_{1:n}) \propto \\ &\quad \phi^{\frac{n}{2}+a-1}\ \exp\left\{-\frac{1}{2}\phi(ns^2 + 2b)\right\}\ \phi^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\phi n\left((\theta - \bar{y})^2 + n_0(\theta_0)^2\right)\right\}.\end{aligned}$$

After some algebra and completing the square that appears in it, the last exponential term can be written as