Università degli Studi di Padova

Corso di Laurea Magistrale
in Scienze Statistiche

TESI DI LAUREA

PREDIZIONE BAYESIANA APPROSSIMATA DI ORDINE SUPERIORE
Higher-order Approximate Bayesian Prediction

Relatore: Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando: Iurie Popa
Matricola N° 1036882

ANNO ACCADEMICO 2013/2014

Contents

INTRODUCTION **III**

1 STATISTICAL INFERENCE AND PREDICTION **1**

1.1 Introduction 1
1.2 Theory of statistical inference 1
1.3 Approaches to statistical inference 4
1.3.1 Frequentist inference 4
1.3.2 Bayesian inference 6
1.4 Prediction 9
1.4.1 Frequentist prediction 9
1.4.2 Bayesian prediction 10

2 APPROXIMATE BAYESIAN PREDICTION **13**

2.1 Introduction 13
2.2 Numerical integral approximation 14
2.3 Monte Carlo integral approximation 15
2.4 Analytical approximation 18
2.4.1 Laplace's method for univariate integrals 19
2.4.2 Laplace's method for multivariate integrals 20
2.5 Laplace's method for predictive densities 21

3 HIGH-ORDER PREDICTIVE AREA APPROXIMATION **27**

3.1 Introduction 27
3.2 Third order approximation to the tail area 28
3.3 HOPA method 29

3.4 Practical approach to HOPA method and HOPA simulation scheme 33

4 EXAMPLES **37**

4.1 Introduction 37

4.2 Normal model 38

4.2.1 Normal model with unknown mean 38

4.2.2 Normal model with unknown standard deviation . . . 42

4.2.3 Normal model with unknown mean and variance 47

4.3 Gamma model 53

4.3.1 Gamma model with unknown rate parameter 53

4.3.2 Gamma model with unknown parameters 61

4.4 Stationary first order autoregressive process AR(1) 68

CONCLUSIONS **75**

Bibliography **79**

INTRODUCTION

The objective of this thesis is to explore Bayesian predictive inference in statistical analysis using approximation and simulation techniques and to implement these techniques in practice. Prediction (Latin præ-, "before," and dicere, "to say") in statistics is a part of inference which aims to make statements about how things will happen in the future. It is usualy based on experience or knowledge and on the underlying statistical model. We can never be sure that the model is entirely appropriate. The discusion of the model building process and the question of adequacy of assumptions is outside the scope of the thesis. However, an introduction to statistical modelling will be presented and in particular to parametric statistical models.

The predictive inference also depends upon which approach is selected for this purpose. There are two main structured and distinct approaches to inference, frequentist inference and Bayesian inference. We are interested in the Bayesian view on the problem of prediction. This paradigm can be seen as an extra step in the modelling world just as parametric modelling is in the classical approach. Wishing to do predictive inference, Bayesian theory calls for the use of posterior predictive distribution, i.e., to predict the distribution of a new, unobserved data point. This approach uses the entire posterior distribution of the parameter. By comparison, prediction in frequentist statistics often means finding an optimal point estimate of the parameter and then plugging in this estimate into the formula for the distribution of the predictive variable. This has the disadvantage that it does not account for any uncertainty in the value of the parameter, and hence will underestimate the variance of the predictive distribution.

Even if the problem under consideration involves a conjugate prior, this

does not mean that the predictive distribution will be available in closed form. Hence analytical, stochastic or numerical approximations are usually needed to evaluate various characteristics of predictive distributions, especially their density, quantiles, or moments. Typically, Markov Chain Monte Carlo (MCMC) methods are used, but such methods have some disadvantages. They are computationaly intensive, time consuming and produce dependent samples. Moreover, MCMC methods typically require more attention from the practitioner, e.g. choice of proposal and convergence checks, and they may have poor tail behavior, especially when the number of parameters is large.

Wishing to avoid the difficulties related to MCMC methods, in this thesis we develop an easily computable approximate method for the predictive cumulative density function, which will be named Higher-Order Predictive Approximation (HOPA) method. This method will allow us to approximate the predictive distribution and related quantiles for univariate predictive random variables. Moreover, it can be used to obtain random samples from such variables, which can be used to approximate summary statistics, for example predictive moments. Its main advantages, compared to standard MCMC methods, are that it gives independent samples at a negligible computational cost.

The development of the HOPA method consists of two main stages. Firstly, the predictive density is approximated by Laplace's method for integrals. We will select two approximations, to which we will apply the third order approximation to the tail area. The results are approximations of the predictive cumulative distribution function, which can be used for inference and simulation.

The thesis incorporates two main parts: theoretical (Chapters 1-3) and practical (Chapter 4), and it will be organized as follows. In Chapter 1 the inference theory will be introduced, with the levels of model specification. Also, we will briefly review frequentist and Bayesian approaches, their differences and major concepts, especially in predictive inference, where the predictive density function will be introduced.

Chapter 2 deals with three major techniques for integral evaluation used

to approximate predictive density functions. We will mainly focus on the Laplace's method, which will be implemented to construct different approximations for the predictive density functions. We will also recall MCMC methods such Metropolis-Hastings algorithm and Gibbs sampling, as these methods are considered trustworthy and will allow us to make comparisons in the examples of Chapter 4.

The development of the HOPA method from the approximations of predictive densities provided in the previous part and conditions that allow us to implement the method, are discused in Chapter 3. Moreover, we will design the procedures to approximate the predictive cumulative distribution function from theoretical and practical point of view, which allows us to compute the related quantiles. Also the HOPA simulation scheme will be discussed.

As the focus of the thesis is both theoretical and practical, numerical examples are discussed in Chapter 4 to illustrate the accuracy of the HOPA method.

Chapter 1

STATISTICAL INFERENCE AND PREDICTION

1.1 Introduction

The aim of this chapter is to introduce the theoretical aspects and formalisation of the Baysian prediction procedure and to provide notation for all further developments. Wishing to grasp the difference between the Bayesian approach and the Fisherian's approach to prediction, the latter will be also recalled. But firstly, the field of statistical inference, with the two main approaches: classical and Bayesian ones, will be discussed.

The chapter is an analysis and a synthesis of Pace and Salvan (1997, Capters 1-4) for frequentist approach and Box and Tiao (1992, Capters 1-4), Barnett (1999, Chapters 1-2), Congdon (2001, Chapter 1) and Iversen (1984, Capters 1-4) for Bayesian approach.

1.2 Theory of statistical inference

The procedure that utilises collected information to obtain a description of a practical situation, through a probability model, is an inferential procedure. The study of such procedures will be termed statistical inference.

Often the collected information consists of data that are inherently vari-

able and from which we would like to highlight regularities or features about the phenomenon under study. The notation for the observed data is y, often of the form $y = (y_1, \ldots, y_n)$. In statistical inference, $y \in \mathcal{Y}$ can be thought as a realization of a random vector $Y \sim p^0(y)$ or, more generally, of a stochastic process, where $p^0(y)$ represents the unknown **probability distribution** o **probability model**, with respect to a suitable measure $\mathcal{P}$, and where $\mathcal{Y}$ is the sample space. Measure theory is not essential for the development of the arguments in the thesis; however, the reader is assumed to be familiar with the basic definitions and results (see Jacod and Protter (2003, Chapters 1-8)). The sample space $\mathcal{Y}$ is the set of possible outcomes of the experiment, i.e. the range of values of the random variable Y. We will assume that $\mathcal{Y} \subseteq \mathbb{R}^n$. In the following, depending on the nature of the random variable Y, discrete or continuous, $p(y)$ will denote either the probability mass function or the density function.

From a statistical point of view, the study of a process consists in reconstructing the unknown $p^0(y)$ on the basis of both suitable assumptions on the phenomenon and the observed data. It is commonly used to define a statistical model $\mathcal{F}$ as a collection of probability distribution functions $p(y)$ from which one assumes that a particular dataset is sampled. The assumptions on $\mathcal{F}$, which usually limitate the possible forms of $p(y)$, facilitate the reconstruction of the probability model. Obviously, the probability distributions $p(y)$ must be compatible with the observed data y, at least mathematically. The statistical model $\mathcal{F}$ is then said to be **correctly specified** if $p^0(y) \in \mathcal{F}$, otherwise the model is said to be **misspecified**.

One of the problems for statistical analysis is the problem of specification wich is very important, and usually impacts a lot on the inferencial conclusions. However, the theory of statistical inference, traditionally, lacks explicit indications on this aspects. The process of selecting on the available information an appropriate statistical model $\mathcal{F}$, where it can take a greater or lesser degree of extension, is called model specification. Widely speaking there are three levels of specification:

- **parametric specification**, where the elements of $p(y)$ can be indexed by a finite number d of parameters which form the vector of parameters

θ, that is

$$\mathcal{F} = \{p(y;\theta), \theta \in \Theta\}, \tag{1.1}$$

where Θ is the parameter space, i.e. the set of all possible combinations of values for all the different parameters which are allowable in the particular model;

- **semiparametric specification**, the elements of $\mathcal{F}$ can be identified through both a parametric and a nonparametric component and it is noted by

$$\mathcal{F} = \{p(y;\theta), \theta \in \Theta\},$$

where $\theta = (\tau, h(\cdot))$, with $\tau \in \mathrm{T} \subseteq \mathbb{R}^k$ whereas the set of possible specifications of the function $h(\cdot)$ cannot be indexed by a finite number of real parameters;

- **nonparametric specification**, the elements of $\mathcal{F}$ cannot be indexed by a finite number of parameters, nor is inference upon finite dimensional characteristics of the distribution of Y.

In this thesis we are concerned with parametric specification, which is the most restricted level of specification of a statistical model, where the probability distribution is a function of the parameters θ. A parametric statistical model $\mathcal{F}$, from the classical point of view, can be specified once the triplet

$$(\mathcal{Y}, p(y;\theta), \Theta),$$

has been assigned, where all the elements have just been defined. Often the parameter space is a subset of d-dimensional Euclidean space, i.e. $\Theta \subseteq \mathbb{R}^d$.

The statement about the probability model is the fundamental assumption made in this context. The knowledge of $p^0(y)$ will permit both interpretation and prediction.

1.3 Approaches to statistical inference

Seeking suitable techniques for the identification of $p^0(y)$, statistical investigation takes place in a context or in a paradigm, which depends on the approach that was taken to do inference. There are two main structured and distinct approaches to inference, namely *frequentist inference*, also known as classical, and a well-established alernative, *Bayesian inference*. There are other approaches, such as fiducial inference or pure likelihood inference that are not mentioned here. See Barnett (1999, Chapters 1-2) for a comparative review. The choice between the two main approaches of statistical inference depends on how we define the concept of probability and on what we consider as relevant information. These implications have a great impact on statistical modelling and inferences.

The main characteristics of the two approaches may be summarised in the manner indicated in Table 1.1 (see Barnett (1999, Section 1.5)), and are briefly described in Sections 1.3.1 and 1.3.2 .

Table 1.1: Comparison between classical and Bayesian approaches.

Approach	Probability concept	Relevant information
Classical	Frequency-based	Sample data
Bayesian	'Degree-of-belief'; subjective. Possible frequency interpretable components	Sample data. Prior information

1.3.1 Frequentist inference

Classical approach to inference originates in the work of R. A. Fisher, J. Neyman, E. S. Pearson, and others. Probabilities in this approach are seen as long-run relative frequencies or proportions and some people therefore call them frequentist probabilities or objective probabilities. In these respects classical statistics leans on a *frequency concept* of probability. This view was first formulated by Venn (1886) and later led to the Neymann-Person system of classical statistical inference.

In the context of a parametrical statistical model, if the model is correctly specified, we have $p^0(y) = p(y;\theta_0)$ for a value $\theta_0 \in \Theta$, called the **true parameter value** and $p^0(y)$ is named **true probability distribution**. The observables are the outcomes of the experiment Y and are determined by certain objective probabilities or joint probability distribution which can be viewed as function of a set of unknown parameters.

Wishing to throw light on the unknown parameter θ_0, the identificability condition and the information are needed. The identificability condition states that there is at least $y \in \mathcal{Y}$ such that $p(y;\theta) \neq p(y;\theta')$ if $\theta \neq \theta'$. And the only source of relevant information for all the procedures of this approach is sample data.

The main inferences about parameter are point and interval estimation, tests of significance and hypothesis testing. In this rispect, the terms **estimate** and **estimator**, or more generally **statistic**, play a special role. The estimator is a particular function $\tilde{\theta}(Y)$ of the random variable Y and the estimate is the actual value the estimator takes, $\tilde{\theta}(y)$. More formally, the estimator is viewed as the transformed random variable with its probability distribution that is called the **sampling distribution**. The sampling distribution of the estimator of θ is one of the keypoints for further steps of classical statistical inference.

There are many practical methods that are used for constructing estimators, such as maximum likelihood method, method of moments, method of estimation by order statistics, method of minimum chi-squared, least squares. The estimator (estimate) obtained by using maximum likelihood method is called **maximum likelihood estimator (estimate), MLE**. The method is based on a fundamental tool in statistical inference, namely the **likelihood function**. If we have a parametric statistical model $\mathcal{F}$ for data y, we write $p(y;\theta)$ to emphasize that the density is a function of both data and parameter. The likelihood function for θ based on y is defined to be the function $L : \Theta \to \mathbb{R}^+$

$$L(\theta) = L(\theta; y) = p(y;\theta) \tag{1.2}$$

regarded as a function of θ for fixed y. The likelihood function expresses

how the probability distribution, $p(y;\theta)$, for the particular data y varies as we consider different possible values for the parameter θ. It represents the information provided by the sample that does not involve its compression into a particular parameter-free statistic. So, the maximum likelihood estimator is the value of θ that maximizes $L(\theta;y)$ over Θ.

Consider an experiment that yields a sequence of data $y = (y_1, \ldots, y_n)$. Then the obvious form for the likelihood function is

$$L(\theta) = p(y_1;\theta)p(y_2|y_1;\theta)\ldots p(y_n|y_{n-1}, y_{n-2}, \ldots, y_1;\theta). \tag{1.3}$$

In the case of independent observations the likelihood function has the form

$$L(\theta) = \prod_{i=1}^{n} p(y_i;\theta). \tag{1.4}$$

In many cases, even in maximizing $L(\theta)$, it may prove more convenient to consider the **log-likelihood function**, $l : \Theta \to \mathbb{R}$, which is defined as

$$l(\theta) = \log L(\theta) = \log p(y;\theta), \tag{1.5}$$

where $\log(\cdot)$ denotes the natural logarithm, and $l(\theta) = -\infty$ if $L(\theta) = 0$.

Quantities obtained from the likelihood function are called **likelihood quantities**. The most known among such quantities are the derivatives of the log-likelihood function up to the second order, named **score function** and, with the changed sign, the **information matrix**, which play a crucial role in classical inference. Basic reviews of the wide-ranging role of the likelihood are provided by for instance Barnard and Sprott (1983) and by Hills (1998).

1.3.2 Bayesian inference

The distinction between parameters and observables is not as clear as commonly supposed because each quantity has a range of possible values, and a single specific realized value. Bayesian inference avoids such problematic distinction by assuming that observables y and parameters θ are generic realisations of random variables Y and θ with joint distribution $p(y, \theta)$. The

parametric statistical model (1.1) gives the conditional distribution of Y given θ, and is only one of the ingredients in the Bayesian specifiation, merely extended by introduction of the *prior distribution* $\pi(\theta)$ for the parameters θ, also known as the distribution of θ *a priori*.

Bayesian inference needs the prior information as well as the observed sample. The prior information is modified by the sample data through the Bayes theorem to yield a combined assessment of the state of knowledge of the practical situation. Inferential statements are expressed through posterior probability distributions, and hence embody their own measure of accuracy. This approach cannot rest on a frequency interpretation of probability alone; a subjective interpretation is almost inevitable, and probabilities tend to be regarded as conditional on the observed data. The prior information represents a personal measure of uncertainty, based on the available evidence. Since our available evidence is mostly empirical both classical and Bayesian views often come up with very similar answers for a particular problem (Iversen (1984, pag. 8)).

The Bayes' theorem in the simplest form states as follows (see O'Hagan (1994, Chapter 1)). If we consider two events, A and B, then

$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}. \tag{1.6}$$

This formula can be interpreted in the following way. If we are concerned with event B, knowing or supposing *prior* probability $P(B)$ for its occurrence, and we observe the occurrence of event A then we can construct the *posterior* probability $P(B|A)$. The posterior probability describes how likely B is when A is known to have occurred. The probability $P(A|B)$ is known as the *likelihood* of A given B. The theorem can be understood as a formula for updating from prior to posterior probability multiplying the prior $P(B)$ by the ratio $P(A|B)/P(A)$. A more convenient form of Bayes' theorem for our purposes is

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

where $\pi(\theta)$ is a priori distribution for θ, $p(\theta|y)$ is the posterior distribution

of θ given y, or the distribution of θ *a posteriori*, and $p(y)$ is the marginal density of Y. From the point of view of Probability Theory, $\pi(\theta)$ is the marginal distribution of the parameter θ, $p(\theta|y)$ is the conditional probability distribution of θ given $Y = y$, and $p(y|\theta)$ is the conditional probability distribution of Y given θ. The latter, from a statistical point of view, is the likelihood function. Using Theorem of marginal distribution (see Jacod and Protter (2003, pag. 88)), to cover both cases, where θ can be continuous or discrete, we can write

$$p(y) = \sum_{\theta \in \Theta} p(y|\theta)\pi(\theta)$$

for discrete θ, where $\pi(\theta)$ is a probability mass function and

$$p(y) = \int_{\Theta} p(y|\theta)\pi(\theta)d\theta$$

for continuous θ, where $\pi(\theta)$ is a density function. In practice, it is exceptional to be concerned with a discrete parameter, then it is more useful to express the posterior density of θ in the form

$$f(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{\int_{\theta} p(y|\theta)\pi(\theta)d\theta}. \tag{1.7}$$

This formulation can be interpreted in the same way as the version of Bayes' theorem (1.6) for the simpler case of events. Posterior density of parameters combines two sources of information: the prior density of θ and the likelihood for θ given y, and it is proportional to their product. If there are particular values of θ that are well supported by both information sources, i.e. having high prior density and high likelihood then these values will also have high posterior density. And viceversa, posterior density for particular values of θ will be low if they have low prior density and low likelihood, so that they are essential discounted by both sources of information.

Notice that $p(y|\theta)$ appears in both numerator and denominator of (1.7). Therefore if we modify $p(y|\theta)$ by multiplying it by an arbitrary constant, then that constant will cancel and leave the same posterior disribution. Even more generally, we can multiply $p(y|\theta)$ by an arbitrary function of y, and still obtain

the same posterior distribution. The implication is that if we have y and x, data arising from two distinct experiments and the two likelihoods $L(\theta, y)$ and $L(\theta, x)$ are identical up to multiplication by an arbitrary function of y and x, then they contain identical information about θ and lead to identical posterior distributions. The experiments might be very different in other respects, but those differences are irrelevant for inference about θ. This principle is called *Likelihood Principle*. Broadly speaking, the Likelihood Principle implies that it matters only what was observed. The Likelihood Principle also represents a key difference between Bayesian and classical theory. Berger and Wolpert (1988) provide a careful and deep analysis, and show that classical inference does not follow the Likelihood Principle, as opposed to Bayesian inference.

1.4 Prediction

Statistical prediction consists of two experiments Y, called the informative experiment, and Z, called the future experiment. If there is some link between these two experiments then from the information gained from the informative experiment some reasoned statement concerning the future experiment could be made. Classical and Bayesian prediction inferences deal with problems where this link is through the parameters θ and the way in which the outcome of the two experiments y and z are related.

1.4.1 Frequentist prediction

In the classical approach, there is no such natural route to prediction nor interpretable formulation as in the Bayesian approach. Predictive inference from this viewpoint represents what happens in the long run as a particular prediction procedure for repeated sample data will be used. There are many proposals for prediction procedures in classical inference, and most of them effectively assume that θ just takes a specific value obtained as a point estimator.

Suppose that in the informative experiment the continuous random variable Y, with sample space $\mathcal{Y} \in \mathbb{R}^n$, has taken the value y. Consider for Y

the parametric statistical model with density functions that can be denoted by $\{f(y;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, where θ is unknown parameter. And we wish to make prediction for an unobserved outcome of continuous random variable Z (or vector) from a future experiment. Suppose, also, that the conditional distribution of Z given Y, denoted by $f(z|y;\theta)$, is known. The simplest procedure consists of using the estimative predictive density $\hat{f}(z|y) = f(z|y;\hat{\theta})$, obtained by substituting an asymptotically efficient estimator of θ, $\hat{\theta}$, such as the maximum likelihood estimator. This procedure however takes no account of the statistical variability of the estimator.

A number of papers aim to improve the estimative density, such as Harris (1995), Vidoni (1995), Vidoni (1998), Komaki (1996), Barndorff-Nielsen and Cox (1996) and Corcuera and Giummolè (1999). For instance, Vidoni (1998) obtains an upper α-prediction limit for z, that is a value $z_\alpha(y)$ such that, exactly or approximately,

$$Pr\{Z \leq z_\alpha(Y);\theta\} = 1 - \alpha \ ,$$

for all possible $\theta \in \Theta$, where the probability refers to joint distribution of (Y, Z).

Other related issues of classical prediction methods have been discussed by Guttman (1970), who compares with Bayesian methods, Butler (1986), Bjornstad (1990) and Geisser (1993).

1.4.2 Bayesian prediction

Assume the same setting of informative and future experiments as in Section 1.4.1, adding for the unknown parameter θ the prior density function $\pi(\theta)$, continuous in Θ, and the joint density of Y and Z denoted by $f(y, z|\theta)$, or the conditional distribution of Z given Y and θ, denoted by $f(z|y;\theta)$. Clearly, the plausibility of z given $\pi(\theta)$ and y is expressed by

$$f(z|y) = \int_\Theta f(z,\theta|y)d\theta = \int_\Theta f(z|y;\theta)f(\theta|y)d\theta \ , \tag{1.8}$$

and we term this function the predictive density function (see Aitchison and Dunsmore (1975, p.19)) for $Z = z$ given $Y = y$. The substitution in (1.8) of the posterior density (1.7) leads to the expresion

$$f(z|y) = \frac{\int_\Theta f(z, y|\theta)\pi(\theta)d\theta}{\int_\Theta f(y|\theta)\pi(\theta)d\theta} = \frac{\int_\Theta f(z|y;\theta)f(y|\theta)\pi(\theta)d\theta}{\int_\Theta f(y|\theta)\pi(\theta)d\theta}. \tag{1.9}$$

An extreme version of the predictive approach is to regard parameters as neither meaningful nor necessary. Then the predictive distribution for a future z is obtained by

$$f(z|y) = \frac{f(z, y)}{f(y)} \tag{1.10}$$

without reference to the parameter θ. We note that the ratio in the right-hand side includes only the joint density of observables and unobservables data. The method has the difficulty in defining these joint distributions. Introducing parameters is a natural way of representing the distribution of the data. All unobservable parameters are viewed as nuisance parameters, but they are still employed in constructing the basic model.

Formula (1.10) can yield the predictive density (1.9) if the marginal density of (Y, Z), $f(y, z)$, is obtained by joint density of (Y, Z, θ) and marginal density of Y, $f(y)$, is obtained by joint density of (Y, θ) through the marginalization on θ. When we have $f(y, z, \theta)$ and can not define $f(y, \theta)$, or for other convinience, then the marginal density of Y can be written as

$$f(y) = \int_\Theta \left[\int_{\mathbb{R}} f(y, z, \theta)dz \right] d\theta \ , \tag{1.11}$$

and the predictive density takes the following form

$$f(z|y) = \frac{\int_\Theta f(z|y;\theta)f(y|\theta)\pi(\theta)d\theta}{\int_\Theta \left[\int_{\mathbb{R}} f(y, z, \theta)dz\right] d\theta} = \frac{\int_\Theta f(z, y|\theta)\pi(\theta)d\theta}{\int_\Theta \left[\int_{\mathbb{R}} f(y, z, \theta)dz\right] d\theta}. \tag{1.12}$$

Formulas (1.9) and (1.12) will be employed further in Chapter 2.

We note that the probability distribution (1.8) is represented as a mixture distribution. To generate a random variable $Z|Y = y$ using such a representation, we can first generate a variable $\theta|Y = y$ from the mixing distribution

and then generate $Z|Y = y; \theta$ from the selected conditional distribution. For more details see Robert and Casella (2010, Section 2.2.3).

Chapter 2

APPROXIMATE BAYESIAN PREDICTION

2.1 Introduction

In practice, a user of Bayesian predictive inference needs to be able to evaluate various characteristics of the predictive distribution, such as its density, quantiles, mean and variance. In this chapter, the quantity of interest is the predictive density, defined as a ratio of two integrals (see Section 1.4.2). The results depend on the calculus of the nominator and denominator and this involves integration over the parameters. In some examples, prior distributions and likelihoods have convenient forms that enable the closed form computation of the predictive density. In general, however, closed form results are not available and then analytical, stochastic or numerical approximation methods are used.

In the following we are going to outline three major techniques for integral approximation. We will mainly focus on the Laplace's method and its implementation for predictive density. However, MCMC methods, such Metropolis-Hastings algorithm and Gibbs sampling, will be recalled, as we will use these methods for comparisons in some of the examples in Chapter 4. The chapter is based on Bruijn (1961, Chapter 4), Tierney and Kadane (1986), Robert and Casella (2010, Chapters 3-7), Pace and Salvan (1997,

Chapter 9) and Davison (1986).

2.2 Numerical integral approximation

In numerical analysis, numerical integration constitutes a broad family of algorithms for calculating the numerical value of a definite integral. Numerical integration for one-dimensional integrals of the form

$$I = \int_a^b f(x)dx,$$

also known as quadrature, essentially approximate I by calculating f at a number of points $x_1, x_2, \ldots, x_k \in [a, b]$ called **integration points** and applying some formula to the resulting values $f(x_1), f(x_2), \ldots, f(x_k)$. The simplest form for the one-dimensional integral is a weighted average

$$\hat{I} = \sum_{i=1}^{k} w_i f(x_i).$$

Different quadrature rules are distinguished by using different sets of design points $x_1, x_2, \ldots, x_k$, and/or different sets of weights $w_1, w_2, \ldots, w_k$. Also a large class of quadrature rules can be derived by constructing interpolating functions which are easy to integrate. The most known os such rules is the Gaussian quadrature.

Numerical integration over more than one dimension is sometimes described as cubature or Cartesian product quadrature. For the computation of integrals in multiple dimensions, one approach is to divide the multiple integral into repeated one-dimensional integrals by appealing to Fubini's theorem. This approach requires the function evaluations to grow exponentially as the number of dimensions increases. Numerical methods are known to suffer the so-called curse of dimensionality.

2.3 Monte Carlo integral approximation

A different approach for evaluating an integral is Monte Carlo integration. This method uses the evaluation of $f(x)$ at random points. Suppose that a series of points $x_1, x_2, \ldots, x_k$ are drawn independently from a distribution with density $s(x)$. Now we have

$$I = \int f(x)dx = \int \frac{f(x)}{s(x)} s(x)dx = E_s \left[\frac{f(x)}{s(x)} \right], \tag{2.1}$$

where E_s denotes expectation with respect to the distribution s. The estimation of (2.1) is given by the sample mean

$$\hat{I} = k^{-1} \sum_{i=1}^{k} \frac{f(x_i)}{s(x_i)}.$$

From a statistical point of view $\hat{I}$ is an unbiased estimator of the integral I with variance

$$\mathrm{var}(\hat{I}) = k^{-1} \mathrm{var}_s \left[\frac{f(x)}{s(x)} \right]. \tag{2.2}$$

For large k, $\hat{I}$ is asymptotically normally distributed, by the central limit theorem, with mean I and variance (2.2), which tends to zero as k increases. We can estimate the variance and give a confidence interval for the integral, based on the same sample. The possibility of assessing the accuracy of the integration is an advantage of Monte Carlo method over quadrature methods. Another advantage of Monte Carlo methods is that they are easy to apply to multi-dimensional integrals, and may yield greater accuracy for the same number of function evaluations than repeated integrations using one-dimensional methods.

Note that the key of Monte Carlo integration is Monte Carlo methods, wich are a broad class of computational algorithms that rely on repeated random sampling. A large class of useful Monte Carlo methods are the so-called Markov Chain Monte Carlo algorithms, which include the Metropolis-Hastings algorithm and Gibbs sampling. In the following we will describe both algorithms as they will be used in some of the examples of Chapter 4.

Metropolis-Hastings algorithm

Suppose that we have a univariate or multivariate probability density $f(x)$, where very little is known about it, and we want to generate a sample from such probability density, not necessarily i.i.d., but with property that the marginal distribution of this sample is $f(x)$. For this purpose the MCMC algorithm can be used to generate correlated samples from a Markov chain with stationary distribution $f(x)$. The working principle of Markov chain Monte Carlo methods is builded on a Markov kernel K, which generates such stationary distribution by the Ergodic Theorem (see Norris (1997, Chapter 3.8)). A method for deriving the kernel K, that is universal and theoretically valid for any density $f(x)$, is the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm associates to a given target density $f(x)$ an conditional density $q(y|x)$. In practice, the conditional density is chosen to be easy to simulate. From a theoretical point of view the only requirement is that the ratio $f(y)/q(y|x)$ should be known up to a constant independent of x and that $q(y|x)$ has enough dispersion to lead to an exploration of the entire support of $f(x)$. Than the algorithm produces a Markov chain $\{X_t\}$ trough the following steps:

1. given x_t, generate y_t from $q(y|x_t)$
2. calculate $\alpha_t = \min(\frac{f(y_t)}{f(x_t)} \frac{q(x_t|y_t)}{q(y_t|x_t)}, 1)$
3. set $X_{t+1} = y_t$ with probability α_{t+1} or $X_{t+1} = x_t$ with probability $1 - \alpha_{t+1}$.

The distribution $q(\cdot)$ is called the instrumental, or proposal, distribution and the probability α_t the Metropolis-Hastings acceptance probability. Another concept of this algorithm is acceptance rate, which is the average of the acceptance probabilities over iterations. This quantity allows an evaluation of the performance of the algorithm. Roberts *et al.* (1997) recommend the use of instrumental distributions with acceptance rates close to 1/4 for models with high-dimensional parameters and equal to 1/2 for the models with 1 or 2 parameters.

We can get special cases from the original algorithm such as symmetric, independent and random walk cases. In the symmetric case, when

$q(x|y) = q(y|x)$, the acceptance probablility α_t is driven by the objective ratio $f(y_t)/f(x_t)$ and thus even the acceptance probability is independent from $q(\cdot)$. If we require the candidate to be independent of the present state of the chain, i.e. $q(y|x) = g(y)$, we do get the independent Metropolis-Hastings algorithms.

The random walk Metropolis-Hastings uses the simulation of Y_t according to $Y_t = X_t + \epsilon_t$, where ϵ_t is a random perturbation with distribution g independent of X_t. The proposal density $q(y_t|x_t)$ is now of the form $g(y_t - x_t)$. If g is symmetric around zero then the Markov chain associated with q is a random walk. But, due to acceptance step, the Metropolis-Hastings Markov chain is not a random walk.

Gibs sampling algorithm

In statistics, Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm and, as Metropolis-Hastings algorithm, is used for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution $f(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_p)$. The method can be implemented when we can simulate from the corresponding conditional densities $f_i(x_i|\mathbf{x}_{-i})$, called full conditional, for $i = 1, \ldots, p$, where $\mathbf{x}_{-i}$ is the vector $\mathbf{x}$ whithout the element x_i. The main steps for general multistage Gibbs sampler algorithm is given by the following transition from X_t to X_{t+1}:

given $\mathbf{x}_t = (x_{1,t}, \ldots, x_{p,t})$, generate

1. $X_{1,t+1} \sim f_1(x_1|x_{2,t}, \ldots, x_{p,t})$;
2. $X_{2,t+1} \sim f_2(x_2|x_{1,t+1}, x_{3,t}, \ldots, x_{p,t})$;

$\vdots$

p. $X_{p,t+1} \sim f_p(x_p|x_{1,t+1}, x_{2,t+1}, \ldots, x_{p-1,t+1})$.

When some of the full conditionals cannot be simulated by standard random generators then the following Metropolis-within-Gibbs strategy can be adopted, where, instead of simulation from full conditional, we can run one single step of any MCMC scheme associated with the stationary distribution of full conditional. A simple solution is for instance to use a random walk

Metropolis algorithm.

Instead of using a joint Metropolis-Hastings algorithm, when the design of such an algorithm on a large-dimensional target is challenging or even impossible, we can iplement a Gibbs-like structure. The fundamental gain in using such a structure is that it breaks down a complex model into a large number of smaller and simpler targets, where local Metropolis-Hastings algorithms can be designed at little cost.

Metropolis-Hastings and Gibbs algorithms, as other MCMC algorithms, generate samples of correlated random variates. As a result, care must be taken if independent samples are desired, typically by thinning the resulting chain of samples by only taking every n-th value, e.g. 10-th value. In addition, samples from the beginning of the chain, the burn-in period, may not accurately represent the desired distribution, and should be discarded.

The samples generated with the algorithms discussed above, can be used to approximate the joint target distribution, the marginal distribution of one of the variables, or some subset of the variables, or to compute an integral, such as the expected value of one of the variables.

A large body of literature has been devoted to Monte Carlo methods. For exemple Ripley (1987) provide a general overview of Monte Carlo methods. Techniques particularly relevant for Bayesian applications are presented by Robert and Casella (2004, Chapter 6).

2.4 Analytical approximation

Numerical integration procedures such as Gaussian quadrature can sometimes be applied, but they are typically useful only for low-dimensional integrals. A powerful tool for approximate calculation of an integral is numerical integration by Monte Carlo simulation, but this method could be computationally intensive. A method that has minimal computational requirements, but with good accuracy is an analytical approach known as Laplace's method.

2.4.1 Laplace's method for univariate integrals

Consider the integral

$$I(n) = \int_{\mathbb{R}} \exp\{-ng(y)\}dy, \tag{2.3}$$

where $g(\cdot)$ is a smooth univariate real function with a unique absolute minimum at $\tilde{y}$, so that $g'(\tilde{y}) = 0$ and $g''(\tilde{y}) > 0$. Under these assumptions, the asymptotic behaviour of $I(n)$ is determined by the local behaviour of $g(\cdot)$ in a neighbourhood of $\tilde{y}$. A Taylor expansion of $g(y)$ around $\tilde{y}$ gives

$$g(y) = \tilde{g} + \tilde{g}'(y-\tilde{y}) + \frac{\tilde{g}''(y-\tilde{y})^2}{2} + \frac{\tilde{g}'''(y-\tilde{y})^3}{6} + \frac{\tilde{g}^{IV}(y-\tilde{y})^4}{24} + O((y-\tilde{y})^5) \tag{2.4}$$

with $\tilde{g} = g(\tilde{y})$, $\tilde{g}'' = g''(\tilde{y})$, $\tilde{g}''' = g'''(\tilde{y})$. The second summand in (2.4) under the assumptions above equals zero. Using expansion (2.4) into (2.3) we obtain

$$I(n) = e^{-n\tilde{g}} \int_{\mathbb{R}} \exp\{-\frac{n(y-\tilde{y})^2\tilde{g}''}{2}\}\exp\{-\frac{n\tilde{g}'''(y-\tilde{y})^3}{6} - \frac{n\tilde{g}^{IV}(y-\tilde{y})^4}{24} + \\ + nO((y-\tilde{y})^5)\}dy.$$

The first factor of the integrand is the density of $N(\tilde{y}, (n\tilde{g}'')^{-1})$ up to the normalizing constant $c = \sqrt{n\tilde{g}/(2\pi)}$. Changing the integration variable to $z = (y-\tilde{y})\sqrt{n\tilde{g}''}$ and highlighting the standard normal density function $\phi(\cdot)$ we have

$$I(n) = \frac{e^{-n\tilde{g}}}{c} \int_{\mathbb{R}} \exp\{-\frac{z^3\tilde{g}'''}{6\sqrt{n}(\tilde{g}'')^{3/2}} - \frac{z^4\tilde{g}^{IV}}{24n(\tilde{g}'')^2} + O(n^{-3/2})\}\phi(z)dz \tag{2.5}$$

We note that $dz = (\sqrt{n\tilde{g}'})^{-1}dy$ and $z(y)$ is a strictly increasing function of y and the extremes of integration do not change. If we use in (2.5) the expansion of the exponential function $e^x = 1 + x + x^2/2 + \ldots$ and the summands of order $O(n^{-3/2})$ and smaller are neglected, we find that

$$I(n) = e^{-n\tilde{g}}c^{-1} \int_{\mathbb{R}} \left[1 - \frac{z^3\tilde{g}'''}{6\sqrt{n}(\tilde{g}'')^{3/2}} - \frac{z^4\tilde{g}^{IV}}{24n(\tilde{g}'')^2} + \frac{z^6(\tilde{g}''')^2}{72n(\tilde{g}'')^3} + O(n^{-3/2})\right] \phi(z)dz.$$

Recall that, if $Z \sim N(0,1)$, then $E[Z^k]$ is zero if k is odd and $(k-2)!!$ if k is even. Note that the summands of order $O(n^{-1})$ is a product of z^4, i.e. the term of order $O(n^{-1})$ remains after integral. In the end, the approximation for integral (2.3) is

$$I(n) = \frac{\sqrt{2\pi}}{n^{1/2}} \frac{e^{-n\tilde{g}}}{\sqrt{\tilde{g}''}} + O(n^{-1})$$

2.4.2 Laplace's method for multivariate integrals

If we have a smooth multivariate function $g : \mathbb{R}^p \to \mathbb{R}$, with a unique absolute minimum at $\tilde{y}$, i.e. $\tilde{g}' = \partial g(y)/\partial y|_{y=\tilde{y}} = 0$ and Hessian matrix for $g(\cdot)$ function in $\tilde{y}$, $\tilde{H} = \partial^2 g(y)/\partial y \partial y^T|_{y=\tilde{y}}$ is positive define, $|\tilde{H}| > 0$ then we can use the same steps as in the univariate case. Hence,

$$I(n) = \int_{\mathbb{R}^p} \exp\{-ng(y)\} dy \tag{2.6}$$

admits the asymptotic approximation of first order

$$I(n) = \frac{(2\pi)^{p/2}}{n^{p/2}} \frac{e^{-n\tilde{g}}}{|\tilde{H}|^{1/2}} + O(n^{-1}) \ . \tag{2.7}$$

In statistical applications it is often useful to have asymptotic approximations of integrals of the form

$$I(n) = \int_{\mathbb{R}^p} b(y) \exp\{-ng(y)\} dy, \tag{2.8}$$

where $b(y)$ is a function of order $O(1)$ such that $\tilde{b} = b(\tilde{y}) \neq 0$ with $\tilde{y}$ as in the previous multivariate case. For this integral we can obtain the approximation

$$I(n) = \frac{(2\pi)^{p/2}}{n^{p/2}} \frac{e^{-n\tilde{g}}\tilde{b}}{|\tilde{H}|^{1/2}} + O(n^{-1}) \ . \tag{2.9}$$

by expanding both $g(\cdot)$ and $b(\cdot)$ around $\tilde{y}$ and taking the same line of reasoning as for the Laplace's approximation (2.7).

Note that the asymptotic expansions leading to (2.7) and (2.9) depend on the smoothness of the exponent functions near their modes. Thus in

particular the approximations does not apply in situations where the mode does not exist, i.e. the exponent functions are not limited from above. In the case we have an interval of integration, the method still applies provided that the mode is an inner point of the interval of integration. For more details see Bruijn (1961, Chapter 4) and Pace and Salvan (1997, Chapter 9). When $\tilde{y}$ is an end point of the interval of integration, the formula can be suitably modified.

2.5 Laplace's method for predictive densities

The aim of this section is the approximation of the predictive density using Laplace's approximations for the integrals in the numerator and denominator. We start from a parametric statistical model, $f(y, z; \theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^d$, where Z is a univariate or multivariate random vector with sample space $\mathcal{Z} \subseteq \mathbb{R}^m$ and Y is a random variable with sample space $\mathcal{Y} \subseteq \mathbb{R}^n$. Suppose $l(\theta)$ denote the log-likelihood function based on data y and $\pi(\theta) = \exp\{\rho(\theta)\}$ is a prior distribution for θ. Let

$$\tilde{l}(\theta) = \log\{f(y|\theta)\pi(\theta)\} = l(\theta) + \rho(\theta),$$

$$l_z(\theta) = \log f(y, z|\theta) = \log\{f(y|\theta)f(z|y;\theta)\} = l(\theta) + \log f(z|y;\theta)$$

and

$$\tilde{l}_z(\theta) = \log\{f(y, z|\theta)\pi(\theta)\} = \log f(y, z|\theta) + \rho(\theta) = l(\theta) + \log f(z|y;\theta) + \rho(\theta).$$

Then we can write the denominator of predictive density (1.9), the marginal density for $Y = y$, in one of the following forms

$$f(y) = \int_{\mathbb{R}^p} f(y|\theta)\pi(\theta)d\theta = \int_{\mathbb{R}^p} \exp\big\{l(\theta)\big\}\pi(\theta)d\theta$$

or

$$f(y) = \int_{\mathbb{R}^p} \exp\big\{l(\theta) + \rho(\theta)\big\}d\theta = \int_{\mathbb{R}^p} \exp\big\{\tilde{l}(\theta)\big\}d\theta.$$

The first expression can be viewed as formula (2.8) and the second expression can be viewed as formula (2.6). Suppose that $l(\theta)$ and $\tilde{l}(\theta)$ are unimodal and $O_p(n)$ and are twice continuously differentiable functions of θ. The measure of information in Y is n. Thus it is the size of simple random sample or a function of the length of time a stochastic process is observed. Then the Laplace's approximation may be applied to these expressions. If we use formula (2.9) then the denominator may be approximated as

$$f(y) = (2\pi)^{p/2}\frac{\exp\{l(\hat{\theta})\}h(\hat{\theta})}{|j(\hat{\theta})|^{1/2}} + O_p(n^{-1}) = (2\pi)^{p/2}\frac{\exp\{\tilde{l}(\hat{\theta})\}}{|j(\hat{\theta})|^{1/2}} + O_p(n^{-1}), \tag{2.10}$$

where $\hat{\theta}$ is the solution of the equation $\partial l(\theta)/\partial\theta = 0$, i.e. the maximum likelihood estimate, and $j(\theta)$ is minus the $d \times d$ matrix of second derivatives of $l(\theta)$ with respect to θ, i.e. the information matrix.

The second approximation for the denominator, using expression (2.7), may be written as follows

$$f(y) = (2\pi)^{p/2}\frac{\exp\left\{\tilde{l}(\tilde{\theta})\right\}}{|J(\tilde{\theta})|^{1/2}} + O_p(n^{-1}), \tag{2.11}$$

where $\tilde{\theta}$ is the solution of the equation $\partial\tilde{l}(\theta)/\partial\theta = 0$, and $J(\theta)$ is minus the $d \times d$ matrix of second derivatives of $\tilde{l}(\theta)$ with respect to θ. Thus $\tilde{\theta}$ is the mode of $\tilde{l}(\theta)$.

If the prior information is flat in the sense that the first two derivatives of $\rho(\theta)$ are zero near the mode of the log-likelihood function $l(\theta)$ or the mode of the log-posterior function $\tilde{l}(\theta)$, then $\tilde{\theta}$ coincides with the maximum likelihood estimate, $\hat{\theta}$, and $J(\tilde{\theta}) = j(\hat{\theta})$.

The numerator, which is the marginal density for $(Z, Y) = (z, y)$, can be written as

$$f(z,y) = \int_{\mathbb{R}^p} f(z,y|\theta)\pi(\theta)d\theta = \int_{\mathbb{R}^p} \exp\{l_z(\theta)\}\pi(\theta)d\theta$$

or

$$f(z,y) = \int_{\mathbb{R}^p} \exp\{\tilde{l}_z(\theta)\}d\theta.$$

If we suppose that $l_z(\theta)$ and $\tilde{l}_z(\theta)$ are $O_p(n)$, unimodal and twice continuously differentiable functions of θ then we can proceed as we have done for the denominator. Using (2.9) we have

$$f(z,y) = (2\pi)^{p/2} \frac{\exp\left\{l_z\left(\hat{\theta}_z(z)\right)\right\} h\left(\hat{\theta}_z(z)\right)}{|j_z\left(\hat{\theta}_z(z)\right)|^{1/2}} + O_p(n^{-1})$$

$$= (2\pi)^{p/2} \frac{\exp\left\{\tilde{l}_z\left(\hat{\theta}_z(z)\right)\right\}}{|j_z\left(\hat{\theta}_z(z)\right)|^{1/2}} + O_p(n^{-1}), \tag{2.12}$$

where $\hat{\theta}_z(z)$, which depends on the unobserved value z of Z as well as the observed y of Y, is the solution of the equation $\partial l_z(\theta)/\partial\theta = 0$ and $j_z(\theta)$ is minus the $d \times d$ matrix of second derivatives of $l_z(\theta)$ with respect to θ, i.e. the information matrix based on $f(y, z|\theta)$.

The alternative approximation may be obtained by implementing expression (2.7) and the result is

$$f(z,y) = (2\pi)^{p/2} \frac{\exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right)\right\}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} + O_p(n^{-1}), \tag{2.13}$$

where $\tilde{\theta}_z(z)$ is the solution of the equation $\partial\tilde{l}_z(\theta)/\partial\theta = 0$ and also depends on both the unobserved value z of Z as well as the observed y of Y. And $J_z(\theta)$ is minus the $d \times d$ matrix of second derivatives of $\tilde{l}_z(\theta)$ with respect to θ, i.e. the information matrix based on $f(y, z|\theta)\pi(\theta)$.

If the prior information is constant such that the first two derivatives of $\rho(\theta)$ are zero for any θ then (2.12) and (2.13) are equal.

The substitution of expressions (2.10) and (2.12) into (1.10) yields an approximate posterior predictive density of $Z = z$ given $Y = y$ of the form

$$f(z|y) = \exp\left\{\tilde{l}_z\left(\hat{\theta}_z(z)\right) - \tilde{l}(\hat{\theta})\right\} \frac{|j(\hat{\theta})|^{1/2}}{|j_z\left(\hat{\theta}_z(z)\right)|^{1/2}} + O_p(n^{-1}). \tag{2.14}$$

We can prove that the error in the approximation of this expression is of

order $O_p(n^{-1})$ as follows. Our initial ratio is of form $(A + x_1)/(B + x_2)$, where A and B are constants, x_1 and x_2 are of order $O(n^{-1})$. We can expand $1/(B + x_2)$ with respect to x_2 in a Taylor series around zero and the result is

$$\frac{1}{B + x_2} = \frac{1}{B} - \frac{x_2}{(B + x_2)^2} + \frac{2x_2^2}{(B + x_2)^3} + \dots$$

Multiplying this series with $A + x_1$ and taking into consideration that x_2^2 is of order $O(n^{-2})$ we obtain that the initial ratio, as was previously stated, is of order $O_p(n^{-1})$.

The second expression of the predictive probability distribution function is given from the substitution of expressions (2.11) and (2.13) into (1.10)

$$f(z|y) \exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}(\tilde{\theta})\right\} \frac{|J(\tilde{\theta})|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} + O_p(n^{-1}). \tag{2.15}$$

As in the first formula, the error of the approximation is of order $O_p(n^{-1})$. In many important cases the error in (2.15) is not $O_p(n^{-1})$, but $O_p(n^{-2})$. See Davison (1986) and Tierney and Kadane (1986).

Formulas (2.14) and (2.15), are asymptotically equivalent for fixed prior information provided the first two derivatives of $\rho(\theta)$ are bounded in a neighbourhood of the mode, which is true in almost all cases. Therefore in the following we will use only the second approximation (2.15), since it is usually more accurate for finite samples.

A different way to approximate the predictive density uses formula (1.12). The denominator of this ratio, which is $f(y)$, can be written employing $\tilde{l}_z$ as a function of (z, θ) as follows

$$f(y) = \int_\Theta \left[\int_{\mathbb{R}} \exp\left\{\tilde{l}_z(z, \theta)\right\} dz\right] d\theta.$$

Using formula (2.7) the approximation of this integral becomes

$$f(y) = (2\pi)^{(p+1)/2} \frac{\exp\left\{\tilde{l}_z\left(\hat{z}, \tilde{\theta}_z(\hat{z})\right)\right\}}{|\tilde{J}_z\left(\hat{z}, \tilde{\theta}_z(\hat{z})\right)|^{1/2}} + O_p(n^{-1}),$$

where $(\hat{z}, \tilde{\theta}_z(\hat{z})) \in \mathcal{Z} \times \Theta$ is the solution of the equation $\partial \tilde{l}_z(z,\theta)/\partial(z,\theta^T)^T = 0$, and $\tilde{J}_z$ is minus the $(d+m)\times(d+m)$ symmetric matrix of second derivatives of $\tilde{l}_z(z,\theta)$ with respect to vector $(z,\theta^T)^T$. Note that the additional condition of the availability of this approximation respect to approximations (2.14) and (2.15) is that $\hat{z} \in \mathcal{Z}$. Using the formula above as an approximation of the marginal distribution of $Y = y$ and expression (2.13) for the marginal distribution for $(Z,Y) = (z,y)$, we can obtain the approximation of the predictive density as

$$f(z|y) = \frac{\exp\left\{\tilde{l}_z\left(z,\tilde{\theta}_z(z)\right) - \tilde{l}_z\left(\hat{z},\tilde{\theta}(\hat{z})\right)\right\}}{\sqrt{2\pi}} \frac{|\tilde{J}_z\left(\hat{z},\tilde{\theta}_z(\hat{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} + O_p(n^{-1}). \tag{2.16}$$

The order $O_p(n^{-1})$ of the error in the approximation could be proved by Taylor expansions as done previously. We note that the formula is based only on log-posterior of y and z, $\tilde{l}_z$, and its first and second order derivatives. The matrix J_z is a $d \times d$ submatrix of the $(d+m) \times (d+m)$ matrix $\tilde{J}_z$ that coresponds to the parameters θ.

Expressions (2.15) and (2.16) will be used in Chapter 3 as a basis for the development of the HOPA method.

Chapter 3

HIGH-ORDER PREDICTIVE AREA APPROXIMATION

3.1 Introduction

The previous chapter outlined some of the standard methods used to compute predictive distributions for both multivariate and univariate predictive random vectors. Here a specific method is developed, named Higher-Order Predictive Area (HOPA) method, in the case of a univariate predictive random vector. The HOPA method can be used to approximate predictive cumulative distribution functions with related quantities and to generate samples from such random variable, called HOPA simulation. Compared to standard Markov Chain Monte Carlo methods, its main advantages are that it gives independent samples at a negligible computational cost.

From the Laplace's approximations of predictive density formulas, that were discused in Section 2.5, we can approximate the predictive cumulative distribution function using third-order approximation to the tail area probabilities. The chapter is organized as follows. Firstly, Section 3.2 reviews the third-order approximation to the tail area (see Davison (2003, Section 11.3.1)). The HOPA method with its conditions will be discussed in Section 3.3. Finally Section 3.4 describes how HOPA method and HOPA sampling scheme can be implemented in practice.

3.2 Third order approximation to the tail area

We start from integral of form

$$I_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{u_0} a(u) e^{-ng(u)} \left\{1 + O(n^{-1})\right\} du, \tag{3.1}$$

where u is scalar, $a(u) > 0$. In addition, $\tilde{u}$ is such that $g(\tilde{u}) = 0$ and the second derivatives of $g(u)$, $\partial^2 g(u)/\partial u^2$, evaluated at $u = \tilde{u}$ is positive. The first step in approximating the integral I_n is to change the variable of integration from u to $r(u) = \text{sign}(u - \tilde{u}) \{2g(u)\}^{1/2}$, that is $r^2/2 = g(u)$. Then $g'(u) = dg(u)/du$ and $r(u)$ have the same sign, and $rdr/du = g'(u)$, so

$$I_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0} a(u) \frac{r}{r'(u)} e^{-nr^2/2} \left\{1 + O(n^{-1})\right\} dr$$

$$= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0} a(u) \frac{r}{g'(u)} e^{-nr^2/2 + \log b(r)} \left\{1 + O(n^{-1})\right\} dr,$$

where the positive quantity $b(r) = a(u)r/g'(u)$ is regarded as a function of r. Another change of variable is performed, from r to $r^* = r - (rn)^{-1} \log b(r)$, that yields

$$-nr^{*2} = -nr^2 + 2 \log b(r) - (nr)^{-1} \log^2 b(r)/r.$$

The Jacobian of the transformation and the third term in the right hand side above contribute only to the error of $I_n(u_0)$, so

$$I_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0^*} e^{-nr^{*2}/2} \left\{1 + O(n^{-1})\right\} dr^*,$$

where

$$r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right), \quad r_0 = \text{sign}(u_0 - \tilde{u}) \{2g(u_0)\}^{1/2}, \quad v_0 = \frac{g'(u)}{a(u_0)}.$$

Note that the expression under the integral is similar to the approximations of the predictive probability distribution function given in Section 2.5. Thus

this method can be implemeted to approximate the predictive cumulative distribution function as will be done in the following section.

3.3 HOPA method

Consider a parametric statistical model with density $f(y, z; \theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^d$, where Z is a univariate random variable with sample space $\mathcal{Z} \subseteq \mathbb{R}$ and Y is a random variable with sample space $\mathcal{Y} \subseteq \mathbb{R}^n$. Let us denote the functions $l(\theta)$, $\tilde{l}(\theta)$, $l_z(\theta)$, $\tilde{l}_z(\theta)$ and its related quantities as in Section 2.5. The basic requirements in order to develop HOPA method are the same as those of Section 2.5, with the additional condition of the existence of the unique posterior mode and the Hessian of $\tilde{l}_z(z, \theta)$ evaluated at the full mode is negative definite (see for instance Kass *et al.* (1990)). These assumptions are typically satisfied in many commonly used parametric models.

One of the approximations of the predictive density, given in Section 2.5, is

$$f(z|y) \approx \exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}(\tilde{\theta})\right\} \frac{|J(\tilde{\theta})|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} .$$

where the symbol $\approx$ indicates accuracy with relative error of order $O_p(n^{-1})$. Now we can multiply this formula and divide it by $\exp\left\{\tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)\right\} / |\tilde{J}_z(\tilde{z}, \tilde{\theta}_z(\tilde{z}))|^{1/2}$, where $\tilde{J}_z(z, \theta) = -\partial^2 \tilde{l}_z(z, \theta)/\partial(z, \theta^T)^T \partial(z, \theta^T)$ and the vector $(\tilde{z}, \tilde{\theta}_z(\tilde{z})^T)^T$ is the solution of equation $\partial \tilde{l}_z(z, \theta)/\partial(z, \theta^T)^T = 0$. This yields the following approximation

$$f(z|y) \approx \exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)\right\} \frac{|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} \frac{\exp\left\{\tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)\right\} |J(\tilde{\theta})|^{1/2}}{\exp\left\{\tilde{l}(\tilde{\theta})\right\} |\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)|^{1/2}}.$$

If k is the last fraction from the equation above multiplied by $\sqrt{2\pi}$ then we have

$$f(z|y) \approx \frac{k}{\sqrt{2\pi}} \exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)\right\} \frac{|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}}. \quad (3.2)$$

This expression has the form of the integrand from (3.1). So a third-order approximation to the predictive tail area can be obtained from formula (3.2). We start from

$$F(Z = z_0|Y = y) = \int_{-\infty}^{z_0} f(z|y)dz$$

and

$$\int_{-\infty}^{z_0} f(z|y)dz \approx \frac{k}{\sqrt{2\pi}} \int_{-\infty}^{z_0} \exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)\right\} \frac{|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} dz$$

Changing the variable of integration from z to $r(z)$, where $r(z) = \text{sign}(z - \hat{z})\left[2\left(l_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)\right]^{1/2}$, the Jacobian is $-\tilde{l'}_p(z)/r(z)$ with $\tilde{l'}_p(z) = \partial\tilde{l}_z\left(\tilde{\theta}_z(z)\right)/\partial z$, i.e. the first derivative of z's profile function, $\tilde{l}_p$, based on $\tilde{l}_z$. This gives

$$\int_{-\infty}^{z_0} f(z|y)dz \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r(z_0)} \exp\left\{-\frac{1}{2}r^2\right\} \frac{r(z)}{\tilde{l'}_p(z)} \frac{c|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} dr \quad (3.3)$$

The second step is another change of variable from $r(z)$ to

$$r^*(z) = r(z) + \frac{1}{r(z)} \log \frac{q^*(z)}{r(z)}$$

with

$$q^*(z) = \frac{\tilde{l'}_z(z)|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}}{k|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)|^{1/2}} .$$

The Jacobian of the transformation contributes only to the error term of (3.3), so

$$F(Z = z_0|Y = y) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r^*(z_0)} \exp\left\{-\frac{1}{2}t^2\right\} dt = \Phi\left(r^*(z_0)\right), \quad (3.4)$$

where $\Phi(\cdot)$ is the standard normal distribution function.

An alternative approximation of predictive cumulative distribution func-

tion uses formula (2.16) for approximating the predictive density function, and it represents the integral

$$F(Z = z_0|Y = y) \approx \int_{-\infty}^{z_0} \frac{\exp\left\{\tilde{l}_z\left(\tilde{\theta}_z(z)\right) - \tilde{l}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)\right\}}{\sqrt{2\pi}} \frac{|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\hat{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} dz \ .$$

Following the same line of reasoning, we change the variable of integration from z to the same $r(z)$ as above. It is then easily seen that

$$\int_{-\infty}^{z_0} f(z|y)dz \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r(z_0)} \exp\left\{-\frac{1}{2}r^2\right\} \frac{r(z)}{\tilde{l'}_p(z)} \frac{|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)|^{1/2}}{|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}} dr \ . \quad (3.5)$$

Another change of variable is needed from $r(z)$ to $r^{**}(z)$, where

$$r^{**}(z) = r(z) + \frac{1}{r(z)} \log \frac{q^{**}(z)}{r(z)}$$

with

$$q^{**}(z) = \frac{\tilde{l'}_p(z)|J_z\left(\tilde{\theta}_z(z)\right)|^{1/2}}{|\tilde{J}_z\left(\tilde{z}, \tilde{\theta}_z(\tilde{z})\right)|^{1/2}} \ .$$

As in the first development, the Jacobian of the transformation contributes only to the error term of (3.5) and this implies that

$$F(Z = z_0|Y = y) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r^{**}(z_0)} \exp\left\{-\frac{1}{2}t^2\right\} dt = \Phi\left(r^{**}(z_0)\right) \ . \quad (3.6)$$

Note that expressions (3.4) and (3.6) are computationally convenient since they rely entirely on simple posterior quantities. The difference between the two methods consists in $q^*(z)$ and $q^{**}(z)$, where $kq^*(z) = q^{**}(z)$. These quantities have part of $r^*(z)$ and $r^{**}(z)$ in the logarithm function. We can deduse that

$$r^*(z) = r^{**}(z) + \frac{\log(k)}{r(z)}.$$

We know that $r(z)$ is a continuous monoton increasing function, with $r(\tilde{z}) =$

0. If $k \neq 1$ then the distance between $r^*(z)$ and $r^{**}(z)$ goes to infinity as $|z - \tilde{z}| \to 0$, and functions converge as $|z - \tilde{z}| \to \infty$. The convergence is faster the closer the value of k is to 1. In the case $k = 1$ the functions $r^*(z)$ and $r^{**}(z)$ are equal.

As we see, the additional condition to implement HOPA method, as for predictive density approximation (2.16), is the existence of the mode for $\tilde{l}_z(z, \theta)$ as an inner point of the space $\mathcal{Z} \times \Theta$. With the notation for $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$ and its space Θ_{-i}, in the case the dimension of vector parameter d is greater than 1, we can say that the additional condition holds if: 1. the profile of θ_i from $\tilde{l}_z$, for $i = 1, \ldots, d$, is a smooth unimodal function, where the mode is an inner point of Θ_i, for every finite and fixed $(z, \theta_{-i}) \in \mathcal{Z} \times \Theta_{-i}$, and 2. the profile of z is a smooth unimodal function, where the mode is an inner point of $\mathcal{Z}$, for every finite and fixed $\theta \in \Theta$ and given y.

The first condition is formalized by the existence of a solution for the equation $\partial l_z / \partial \theta_i + \partial \rho(\theta) / \partial \theta_i = 0$. If we have a regular model (see, for example Azzalini (2001, pag. 76-77) or Pace and Salvan (1997, pag. 89)) and the $\rho(\theta)$ is a superior bounded function, then the equation has a solution, for $i = 1, \ldots, d$. If the prior distribution has not the required feature, then the solution exists under mild assumptions that depend on the sample y and its size n, because the weight of the information included in the log-likelihood function, l_z, relies on these quantities.

The second point is formalized by the existence of a solution for the equation $\partial \tilde{l}_z / \partial z = 0$ that coincides with $\partial \log f(z|y;\theta) / \partial z = 0$. Note that if the density $f(z|y;\theta)$ is a unimodal smooth function for every finite and fixed $(y, \theta) \in \mathcal{Y} \times \Theta$, where the mode is an inner point of $\mathcal{Z}$, then the condition holds.

As result we can say that if we have a regular model for (y, z) and the prior distribution of the parameter $\pi(\theta)$ is bounded, then the additional condition holds and we can implement the HOPA method. In the case $\pi(\theta)$ is unbounded, we have to consider the size of the sample y for every specific problem to check for a unique mode for $\tilde{l}_z$, which is an inner point of $\mathcal{Z} \times \Theta$, or to get another reparametrization, where the new parameters have a bounded

prior distribution.

There are a lot of situations when $Z|Y = y$ is a random variable with limited support and the mode, for a specific combination of y and θ, is the end point of this support. For example, if we have a gamma random variable with the shape parameter lower or equal than 1, then the mode of such distribution equal zero, that is the end point of gamma's support. In such problems we can transform the predictive random variable using a monotonic function. In the case of gamma distribution it is convinient to transform the random variable as $V = \log(Z)$. We observe that the new random variable V has a smooth unimodal probability distribution function and we can apply the HOPA method to predictive random variable V given $Y = y$. After implementation of HOPA method or HOPA simulation for $V|Y = y$, we can obtain all the quantities we need for $Z|Y = y$. For instance, to calculate the quantile α for the original predictive random variable we use this formula

$$q_{z,\alpha} = \exp\left\{F_V^{-1}(\alpha)\right\} \ .$$

And the cumulative distribution function of Z given $Y = y$ can be determined as

$$F_Z = F_V(\log z_0) \ .$$

Clearly, the simulated sample of $V|Y = y$ could be transformed into a simulated sample for $Z|Y = y$, using the exponential function, i.e. the inverse transformation.

3.4 Practical approach to HOPA method and HOPA simulation scheme

From a practical point of view in order to construct an approximation of the predictive cumulative distribution function for $Z|Y = y$ or its inverse, we should be able to calculate $r^*(z)$ and $r^{**}(z)$ for every z. These functions are monotonical increasing in z and, typically, have a numerical discontinuity at $\tilde{z}$. This problem is not a concern for practical purposes and there

are techniques to avoid it. We will use numerical spline interpolation (see Brazzale *et al.* (2007, Section 9.3)). This technique may exclude values of z in a ϵ-neighborhood of $\tilde{z}$, for some small ϵ. As the algorithms for $r^*(z)$ and $r^{**}(z)$ are equal, we will show only one of them.

Firstly, we have to fix a grid of equally spaced values $z_1 < z_2 < \ldots < z_N$, for a moderate value of N (e.g. 50-100). The extremes of the grid can be found by solving numerically the equations $r^*(z_1) = -4$ and $r^*(z_N) = 4$ because $\Phi(-4) \approx 0$ and $\Phi(4) \approx 1$. Then exclude $z_i \in [\tilde{z} \pm \epsilon j_p(\tilde{z})^{-1/2}]$, where $j_p = -\partial^2 \tilde{l}_z(z, \tilde{\theta}_z(z))/\partial z \partial z$, i.e. the first order derivative of the profile of z, and evaluate $r^*(z)$ over the grid of remaining z values. Finally, a spline interpolator to $(z_i, r^*(z_i))$, for $i = 1, \ldots, N$ and $z_i \notin [\tilde{z} \pm \epsilon j_p(\tilde{z})^{-1/2}]$, is applied. The result from the interpolation, function $r_s^*(z)$, is an approximation of $r^*(z)$ and it will permit us to obtain an approximation of the cumulative distribution function for $Z = z_0$ given $Y = y$ by the formula

$$F(Z = z_0 | Y = y) \approx \Phi\left(r_s^*(z_0)\right) \ . \tag{3.7}$$

The inverse spline, denoted by $r_s^*(z)^{-1}$, is the spline interpolator applied to $(r^*(z_i), z_i)$, for $j = 1, \ldots, N$ and $z_i \notin [\tilde{z} \pm \epsilon j_p(\tilde{z})^{-1/2}]$, and it will allow us to get quantiles, in particular the median. If we are concerned with quantile of level α_0, then it can be approximated by the following formula

$$q_{\alpha_0} \approx r_s^*\left(\Phi^{-1}(\alpha_0)\right)^{-1} \ , \tag{3.8}$$

with all the components defined above.

The given expressions provide accurate approximation of the predictive distribution function and related quantiles, but it is not possible to use them to obtain certain posterior summaries, such as posterior moments or highest posterior density (HPD) regions. These summaries, as any other one, can be obtained from a simulated sample of the predictive random variable. In the following we will introduce the HOPA simulation scheme, which has an inverse simulation approach, similar to what done in Ruli *et al.* (2013) for marginal posterior distribution.

Firstly, we should generate $x = (x_1, \ldots, x_T)$ from standard normal distribution, where T is the desired numerosity, and find the extremes of the grid from the equations $r^*(z_1) = x_{(1)}$ and $r^*(z_N) = x_{(T)}$. The spline interpolator of $(r^*(z_i), z_i)$, for $i = 1, \ldots, N$ and $z_i \notin [\tilde{z} \pm \epsilon j_p(\tilde{z})^{-1/2}]$, will permit us to obtain the predicted value of every z_t, corresponding to the value of x_t. Independently of the number of simulations T, the method requires only a limited number N of function evaluations.

Also, we note that the simulation method is based on the transformation of the normal sample $x = (x_1, \ldots, x_T)$ into $z = (z_1, \ldots, z_T)$ by the spline interpolator. Since x is drawn independently, z will be a random sample as well. The HOPA simulation scheme, as happens in every simulation method, is subject to Monte Carlo error of order $O_p(T^{1/2})$, where T is the number of Monte Carlo trials. However, the HOPA simulation method will permit to control the Monte Carlo error by taking T large enough, given the independence of the simulated random variates.

Chapter 4

EXAMPLES

4.1 Introduction

The aim of this chapter is to illustrate in some examples the performance of the HOPA method based on both $r^*(z)$ and $r^{**}(z)$, which for convenience we will refer to as HOPA1 and HOPA2. Also we will show the performance of the approximations of predictive distribution function, which were used in the development of the two mentioned HOPA methods, i.e. formulas (2.15) and (2.16), and which for convenience will be noted $\tilde{f}^*(z|y)$ and $\tilde{f}^{**}(z|y)$. In every example we will use a simulated sample of sizes ranging from 10 to 30.

Where it will be possible, we will write the exact form for predictive density or/and predictive cumulative distribution functions, which will be used for graphical comparisons with approximations of predictive density and HOPA approximation of predictive cumulative distribution, respectively. Moreover, we will also compare numerical summaries for the most important quantities, such as quantiles of 5% and 95%, the median, the mean, the variance and 90% HPD interval. In the examples, where it is not possible to get the closed form of these quantities, the comparison will be done using a MCMC simulated sample from the predictive random variable.

In Section 1.4.2 it was mentioned that the predictive probability distribution function can be represented as a mixture distribution and the simulation from such representation could be performed in two stages: generation of a

sample from the posterior of θ and generation of $Z|Y = y; \theta$ from the selected conditional distributions.

Wishing to simulate a sample from posterior distribution of the parameters, we will use the trustworthy MCMC methods most widely implemented in practice, the random walk Metropolis-Hastings algorithm and Gibbs within Metropolis algorithm (see Chapter 2.3). The proposal of every parameter is a uniform random variable of the form $Unif(-\delta, \delta)$, where $\delta > 0$ is suitably scaled in order to have an acceptance rate of 30-40%. To rise the quality of MCMC sample we will set the number of iterations to be 10^5, from which 10^3 initial observations will be discarded, and will check the convergence of the chain. For every element of the remaining sample values from the first step, θ_t, one observation for Z, given $Y = y$ and θ_t, will be simulated. The final step gives us a sample from predictive random variable $Z|Y = y$.

The required derivatives and maximizations in HOPA method will be obtained in closed form, wherever possible, or will be computed numerically. There are a lot of routines which perform accurate numerical derivatives. For our purpose the *numDeriv R package* will be used (see Gilbert and Varadhan (2012)). The R function, *optim()*, will be used to find the maximums of the log-prior functions. The spline interpolation will be applied to a grid of 70 values evenly spaced with $0.1 < \delta < 0.3$, where δ depends on the level of irregularity of $r^*(z)$ and $r^{**}(z)$ functions arround the mode $\tilde{z}$. All these numerical approximations are a source of error that is probably negligible but is however difficult to quantify.

4.2 Normal model

4.2.1 Normal model with unknown mean

Suppose that $y = (y_1, \ldots, y_n)$ is an i.i.d. sample from $Y \sim N(\mu, \sigma_0^2)$, where $\sigma_0 > 0$ is known and μ is an unknown parameter. Let the prior distribution of the parameter μ be $N(a, b^2)$, where a and b are hyperparameters.

Then the likelihood function is

$$L(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}.$$

Multiplying the prior density of the parameter μ,

$$\pi(\mu) = \frac{1}{\sqrt{2\pi b^2}} \exp\left\{-\frac{1}{2b^2}(\mu-a)^2\right\} \ ,$$

with the likelihood function, we can obtain the posterior density of μ, which is

$$\pi(\mu|y) \propto \exp\left\{-\frac{1}{2B^2}(\mu-A)^2\right\},$$

where $A = (b^2\sum_{i=1}^{n} y_i + a\sigma_0^2)/(nb^2+\sigma_0^2)$ and $B = \sqrt{\sigma^2b^2/(nb^2+\sigma_0^2)}$.

Hence we can see that the posterior density function of the parameter μ is $N(A, B^2)$. Assume a new observation Z, which is independent from Y. The predictive density function of z given y is

$$f(z|y) = \int_0^{+\infty} f(z|\mu)f(\mu|y)d\mu.$$

Using the normality of posterior distribution of mean parameter, we can show that $Z|Y = y \sim N(A, B^2+\sigma_0^2)$ and its density is

$$f(z|y) = \frac{1}{\sqrt{2\pi(B^2+\sigma_0^2)}} \exp\left\{-\frac{(z-A)^2}{2(B^2+\sigma_0^2)}\right\}$$

or

$$f(z|y) = \frac{1}{\sqrt{(B^2+\sigma_0^2)}}\phi\left(\frac{z-A}{\sqrt{B^2+\sigma_0^2}}\right),$$

where $\phi(\cdot)$ is the standard normal density function. Similarly we can write the cumulative distribution function

$$F(z|y) = \Phi\left(\frac{z-A}{\sqrt{B^2+\sigma_0^2}}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Laplace's approximation of predictive probability distribution function

With the same notation for $\tilde{l}(\mu)$ and $\tilde{l}_z(\mu)$, of Section 2.5, we have

$$\tilde{l}(\mu) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(\sum_{i=1}^{n} y_i^2 - 2\mu n\bar{y} + n\mu^2) - \frac{1}{2}\log(2\pi b^2) - \frac{1}{2b^2}(\mu^2 - 2a\mu + a^2)$$

and

$$\tilde{l}_z(\mu) = -\frac{n+1}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n} y_i^2 + z^2 - 2\mu(n\bar{y} + z) + (n+1)\mu^2\right) -$$

$$-\frac{1}{2}\log(2\pi b^2) - \frac{1}{2b^2}(\mu^2 - 2a\mu + a^2).$$

From first order condition of $\tilde{l}(\mu)$ and $\tilde{l}_z(\mu)$ we obtain

$$\tilde{\mu} = \frac{n\bar{y}b^2 + a\sigma_0^2}{nb^2 + \sigma_0^2}$$

and

$$\tilde{\mu}_z(z) = \frac{(n\bar{y} + z)b^2 + a\sigma_0^2}{(n+1)b^2 + \sigma_0^2}.$$

The derivatives of second order for $\tilde{l}(\mu)$ and $\tilde{l}_z(\mu)$ with changed sign are

$$J(\mu) = \frac{n}{\sigma_0^2} + \frac{1}{b^2}$$

and

$$J_z(\mu) = \frac{n+1}{\sigma_0^2} + \frac{1}{b^2}.$$

All we need to implement the Laplace's approximation for predictive density for z given y is to put the results into the formula

$$\tilde{f}^*(z|y) = \exp\left\{\tilde{l}_z\left(\tilde{\mu}_z(z)\right) - \tilde{l}(\tilde{\mu})\right\} \frac{|J(\tilde{\mu})|^{1/2}}{|J_z\left(\tilde{\mu}_z(z)\right)|^{1/2}}\ . \tag{4.1}$$

HOPA method for approximation of predictive cumulative distribution function

To implement HOPA method we have to find $\tilde{z}$, an inner point of Z's support, such that $(\tilde{z}, \tilde{\mu}_z(\tilde{z}))$ is the mode for $\tilde{l}_z(z, \mu)$. First order condition is

$$\begin{cases} \partial\tilde{l}_z/\partial\mu = -\frac{1}{\sigma_0^2}(-(n\bar{y}+z)+(n+1)\mu) - \frac{1}{b^2}(\mu - a) = 0 \\ \partial\tilde{l}_z/\partial z = -\frac{1}{\sigma_0^2}(z-\mu) = 0 \end{cases}$$

and the solution of this system is

$$\begin{cases} \tilde{z} = \frac{b^2\sum_{i=1}^n y_i + a\sigma_0^2}{nb^2+\sigma_0^2} = A \\ \tilde{\mu}_z(\tilde{z}) = \tilde{z} = A \end{cases}$$

Note that the conditions to implement HOPA method hold as A is finite, which holds for finite $\sum_{i=1}^n y_i$. Also approximation $\tilde{f}^{**}$ can be implemented. Knowing the function $\tilde{\mu}_z(z)$ we can write the profile function of z from $\tilde{l}_z$ that is

$$\tilde{l}_p(z) = -\frac{1}{2\sigma_0^2}\left[z^2 - 2\tilde{\mu}_z(z)(n\bar{y}+z) + (n+1)\tilde{\mu}_z(z)^2\right] - \frac{1}{2b^2}\left[\tilde{\mu}_z(z)^2 - 2a\tilde{\mu}_z(z)\right] + \text{ ,}$$

which has the first derivative equal to

$$\tilde{l}'_p(z) = -\frac{1}{\sigma_0^2}\left[z - \tilde{\mu}_z(z) - (n\bar{y}+z)\tilde{\mu}'_z(z) + (n+1)\tilde{\mu}'_z(z)\tilde{\mu}'(z)\right] - \frac{1}{b^2}\left[\tilde{\mu}_z(z)\tilde{\mu}'_z(z) - a\tilde{\mu}'_z(z)\right],$$

where $\tilde{\mu}'_z(z) = b^2/\left((n+1)b^2+\sigma_0^2\right)$. The second derivative of $\tilde{l}_z(z,\mu)$ with respect to (z,μ) with changed sign, $\tilde{J}_z(z,\mu)$, is the symmetric matrix

$$\begin{pmatrix} \frac{1}{\sigma_0^2} & -\frac{1}{\sigma_0^2} \\ -\frac{1}{\sigma_0^2} & \frac{n+1}{\sigma_0^2} + \frac{1}{b^2} \end{pmatrix}$$

Then we can define the quantities to implement the HOPA method (see Section 3.3).

Numerical illustration

The dataset is an i.i.d sample of size 10 simulated from a normal distribution, with mean equal 1 and standard deviation equal 1. We set the hyperparameters of the prior distribution to $a = 0$ and $b = 1$. Table 4.1 contains the main summaries of predictive random variable $Z|Y = y$ calculated from exact distribution, and approximated by HOPA method and HOPA simulation based on $r^*(z)$ and $r^{**}(z)$. The plots in Figure 4.1 show the intermediate and final results of the implementation of HOPA methods and the approximations of predictive distribution function.

Table 4.1: Predictive summaries of normal model using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
Exact	-1.28	0.44	2.16	0.44	1.09	[-1.28 , 2.16]
HOPA1	-1.28	0.44	2.16	-	-	-
HOPA1 sim.(10^6)	-1.28	0.44	2.16	0.44	1.09	[-1.28 , 2.16]
HOPA2	-1.28	0.44	2.16	-	-	-
HOPA2 sim.(10^6)	-1.28	0.44	2.16	0.44	1.09	[-1.27 , 2.16]

4.2.2 Normal model with unknown standard deviation

Consider the same setting as in the previous example, where $y = (y_1, \ldots, y_n)$ is an independent and identically normally distributed sample, but now from $Y_i \sim N(\mu_0, \sigma^2)$, where μ_0 is known and σ^2 an unknown parameter. Let the prior distribution of the parameter σ^2 be $Inv - Gamma(a, b)$, where a is the shape hyperparameter and b is the rate hyperparameter. We can write the likelihood function for σ^2 as

$$L(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_0)^2\right\}.$$

Using the prior density of the parameter σ^2,

$$\pi(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-b/\sigma^2}$$

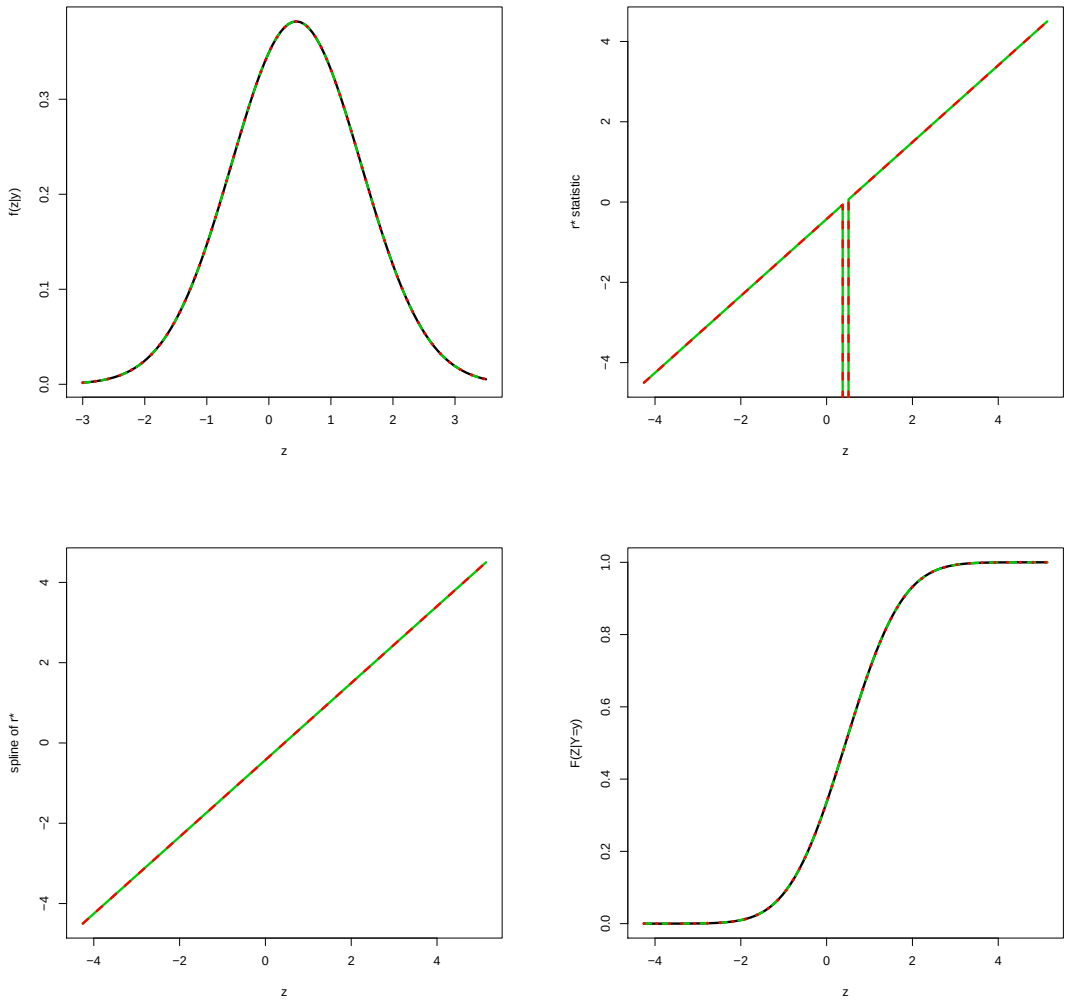

Figure 4.1: Normal model with unknown mean. *Top-left:* The predictive density functions: exact (black-solid line), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = \epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: exact (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

and the likelihood function above, we obtain that the posterior density of σ^2 is

$$\pi(\sigma^2|y) \propto (\sigma^2)^{-(\frac{n}{2}+a)-1} \exp\left\{-\frac{b+\sum_{i=1}^{n}(y_i-\mu_0)^2/2}{\sigma^2}\right\}.$$

We note that the posterior density function of the parameter σ^2 is $Inv-Gamma(n/2+a, b+\sum_{i=1}^{n}(y_i-\mu_0)^2/2)$. Assume a new observation Z independent from Y. Solving the integral

$$f(z|y) = \int_0^{+\infty} f(z|\sigma^2) f(\sigma^2|y) d\sigma^2 \ .$$

we can show that

$$f(z|y) = \frac{\Gamma(\frac{n+1}{2}+a)}{\sqrt{2\pi}\Gamma(\frac{n}{2}+a)} \frac{(S/2+b)^{n/2+a}}{\left(\frac{S+(z-\mu_0)^2}{2}+b\right)^{\frac{n+1}{2}+a}},$$

where $S = \sum_{i=1}^{n}(y_i-\mu_0)^2$. Taking in exam only the factor which contains z, the predictive density has the following feature

$$f(z|y) \propto \left(1+\frac{\frac{(z-\mu_0)^2}{\frac{2b+S}{n+2a}}}{n+2a}\right)^{\frac{(n+2a)+1}{2}} .$$

So we can say that $(Z-\mu_0)/\sqrt{\frac{2b+S}{n+2a}}$ has t-distribution with $(n+2a)$ degrees of freedom and the cumulative distribution function of Z given $Y=y$ is

$$F(z|y) = F_{T_{(n+2a)}}\left(\frac{z-\mu_0}{\sqrt{\frac{2b+S}{n+2a}}}\right),$$

where $F_{T_{(n+2a)}}(\cdot)$ is the cumulative distribution function of t-distribution with $(n+2a)$ degrees of freedom. The random variable $Z|Y=y$ is a generalized Student's t-distribution with location parameter μ_0 and scale parameter $\sqrt{\frac{2b+S}{n+2a}}$.

Laplace's approximation of predictive probability distribution function

With the same notation for $\tilde{l}(\sigma^2)$ and $\tilde{l}_z(\sigma^2)$ as in previous example, we can write these quantities as follows

$$\tilde{l}(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{S}{2\sigma^2} - (a+1)\log(\sigma^2) - \frac{b}{\sigma^2} - \log\left(\Gamma(a)\right) + a\log(b)$$

and

$$\tilde{l}_z(\sigma^2) = -\frac{n+1}{2}\log(2\pi\sigma^2) - \frac{S+(z-\mu_0)^2}{2\sigma^2} - (a+1)\log(\sigma^2) - \frac{b}{\sigma^2} - \log\left(\Gamma(a)\right) + a\log(b).$$

The modes of functions above can be finded by the first order conditions of $\tilde{l}(\sigma^2)$ and $\tilde{l}_z(\sigma^2)$ that give

$$\tilde{\sigma}^2 = \frac{S+2b}{n+2a+2}$$

and

$$\tilde{\sigma}_z^2(z) = \frac{S+(z-\mu_0)^2+2b}{n+2a+3}.$$

The second derivatives of $\tilde{l}(\sigma^2)$ and $\tilde{l}_z(\sigma^2)$ with respect to σ^2 with changed sign are

$$J(\sigma^2) = \frac{n}{2(\sigma^2)^2} - \frac{S}{(\sigma^2)^3} + \frac{a+1}{(\sigma^2)^2} - \frac{2b}{(\sigma^2)^3}$$

and

$$J_z(\sigma^2) = \frac{n+1}{2(\sigma^2)^2} - \frac{S+(z-\mu_0)^2}{(\sigma^2)^3} + \frac{a+1}{(\sigma^2)^2} - \frac{2b}{(\sigma^2)^3}.$$

The formulas above can be plugged in Laplace's approximation of predictive density

$$\tilde{f}^*(z|y) = \exp\left\{\tilde{l}_z\left(\tilde{\sigma}_z^2(z)\right) - \tilde{l}(\tilde{\sigma}^2)\right\}\frac{|J(\tilde{\sigma}^2)|^{1/2}}{|J_z\left(\tilde{\sigma}_z^2(z)\right)|^{1/2}}. \tag{4.2}$$

HOPA method for approximation of predictive cumulative distribution function

We have to find if the additional condition holds, solving the system

$$\begin{cases} \partial\tilde{l}_z/\partial\sigma^2 = -\frac{n+1}{2\sigma^2} + \frac{S+(z-\mu_0)^2}{2(\sigma^2)^2} - \frac{a+1}{\sigma^2} + \frac{b}{(\sigma^2)^2} = 0 \\ \partial\tilde{l}_z/\partial z = -\frac{z-\mu_0}{(\sigma^2)^2} = 0 \end{cases}$$

The solution of this system for z is $\tilde{z} = \mu_0$. We can implement both HOPA methods and approximation $\tilde{f}^{**}$ for predictive density. The profile function of z from $\tilde{l}_z$ is

$$\tilde{l}_p(z) = \tilde{l}_z\left(\tilde{\sigma}_z^2(z)\right) \ ,$$

which has the first derivative equal with

$$\tilde{l}_p'(z) = \left[-\frac{n+1}{2\tilde{\sigma}_z^2(z)} - \frac{a+1}{\tilde{\sigma}_z^2(z)} + \frac{b}{\left(\tilde{\sigma}_z^2(z)\right)^2}\right]\tilde{\sigma}_z^{2'}(z) +$$

$$+\frac{2(z-\mu_0)\tilde{\sigma}_z^2(z) - (S+(z-\mu_0)^2)\tilde{\sigma}_z^{2'}(z)}{2\tilde{\sigma}_z^2(z)},$$

where $\tilde{\sigma}_z^{2'}(z)$ is the first derivative of $\tilde{\sigma}_z^2(z)$ with respect to z and it is equal with

$$\tilde{\sigma}_z^{2'}(z) = \frac{z-\mu_0}{\frac{n+1}{2}+a+1}.$$

The second derivative of $\tilde{l}_z(z,\sigma^2)$ with respect to (z,σ^2) with changed sign, $\tilde{J}_z(z,\sigma^2)$, is the symmetric matrix 2×2

$$\begin{pmatrix} \frac{1}{\sigma^2} & -\frac{z-\mu_0}{(\sigma^2)^2} \\ -\frac{z-\mu_0}{(\sigma^2)^2} & J_z(\sigma^2) \end{pmatrix}$$

Now we can calculate the needed quantities to implement HOPA methods based on $r^*(z)$ and $r^{**}(z)$ such as $r(z)$, $q(z)$ and $q(z)$, as given in Section 3.3.

Numerical illustration

The dataset is an i.i.d sample of size 20 simulated from a normal distribution, with mean equal 0 and standard deviation equal 1. Setting the prior distribution of σ^2 to be $Inv-Gamma(1,1)$, we obtain the summaries in the Table 4.2. The plots in Figure 4.2 show the intermediate and final results of the implementation of approximation of probability distribution function and HOPA methods.

Table 4.2: Predictive summaries of normal model with unknown standard deviation using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
Exact	-1.63	0	1.63	0	0.9	[-1.63 , 1.63]
HOPA1	-1.61	0	1.61	-	-	-
HOPA1 sim.(10^6)	-1.61	0	1.61	0	0.96	[-1.6 , 1.61]
HOPA2	-1.61	0	1.61	-	-	-
HOPA2 sim.(10^6)	-1.61	0	1.61	0	0.97	[-1.61 , 1.61]

4.2.3 Normal model with unknown mean and variance

Let $y = (y_1, \ldots, y_n)$ be an independent and identically distributed sample from $Y \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters. Let the prior distribution of the vector of parameters (μ, σ^2) be Normal-inverse-Gamma, denoted as $(\mu, \sigma^2) \sim N - \Gamma^{-1}(\alpha, \lambda, a, b)$, where if σ^2 is $Inv - Gamma(a, b)$ then $\mu|\sigma^2 \sim N(\alpha, \sigma^2/\lambda)$. The likelihood function for observables Y is

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.$$

Multiplying the prior density of the parameters (μ, σ^2),

$$\pi(\mu, \sigma^2) = \frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma^2}}\frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}exp\left\{-\frac{2b + \lambda(\mu - \alpha)^2}{2\sigma^2}\right\}$$

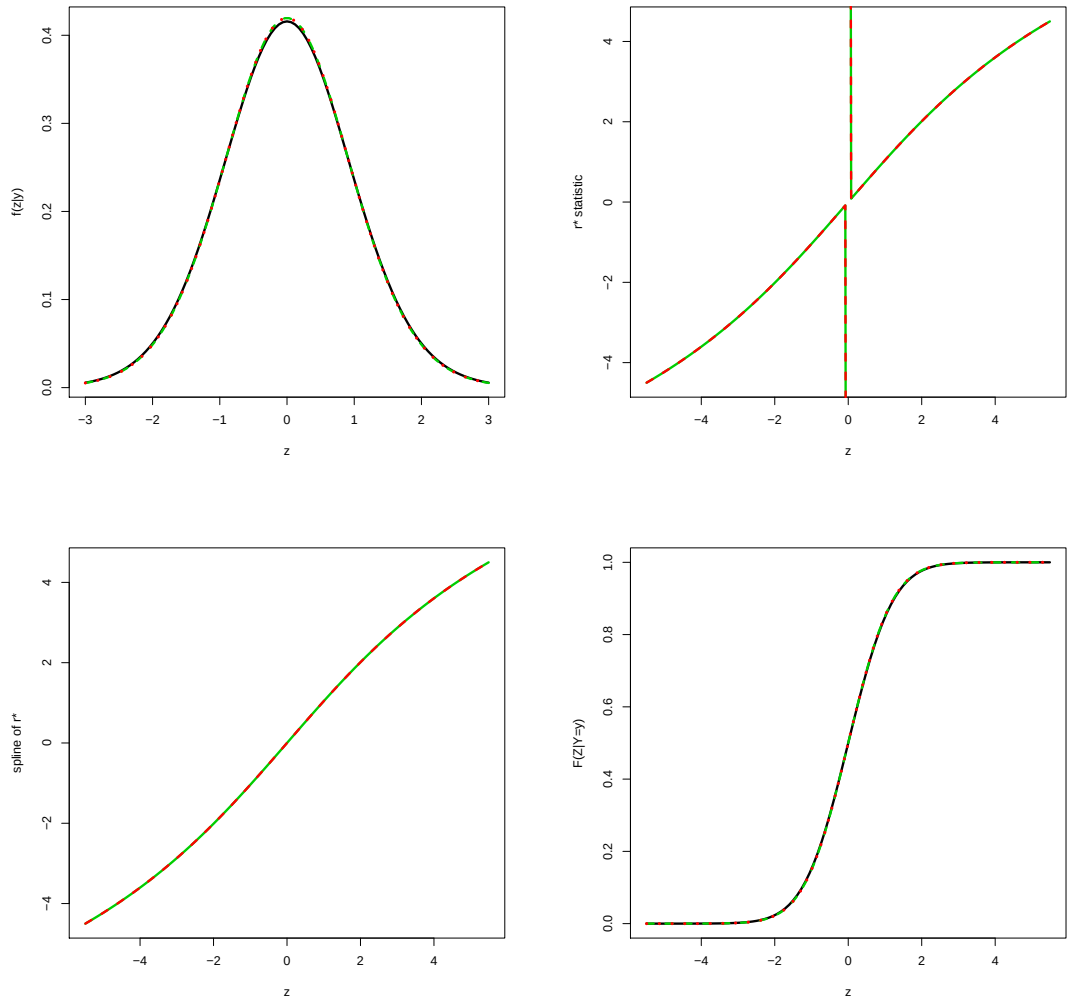

Figure 4.2: Normal model with unknown standard deviation. *Top-left:* The predictive density functions: exact (black-solid line), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = \epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: exact (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

with the likelihood function, we can obtain the posterior density of the vector (μ, σ^2), which is

$$\pi(\mu, \sigma^2|y) \propto \frac{(\sigma^2)^{-a-n/2-1}}{\sqrt{\sigma^2}} \exp\left\{ -\frac{2\left(b + \frac{ns^2}{2} + \frac{n\lambda(\bar{y}-\alpha)^2}{2(n+\lambda)}\right) + (n+\lambda)(\mu - \frac{n\bar{y}+\lambda\alpha}{n+\lambda})^2}{2\sigma^2} \right\},$$

where s^2 is the distort sample variance of our observables y. Can be noted that the posterior density function of the parameters (μ, σ^2) is again Normal-inverse-Gamma with updated parameters

$$\alpha^* = \frac{n\bar{y} + \alpha\lambda}{n+\lambda}; \lambda^* = n + \lambda; a^* = a + n/2; b^* = b + \frac{ns^2}{2} + \frac{n\lambda(\bar{y}-\alpha)^2}{2(n+\lambda)}.$$

Assume a new observation Z that is independent from Y. The predictive density function of z given y can be calculated from the next integral

$$f(z|y) = \int_0^{+\infty} \int_{-\infty}^{+\infty} f(z|\mu, \sigma^2) f(\mu, \sigma^2|y) d\mu d\sigma^2 \ ,$$

and can be written in closed form as following

$$f(z|y) = \frac{b^{*a^*}}{\sqrt{2\pi}} \frac{\Gamma(a^* + \frac{1}{2})}{\Gamma(a^*)} \sqrt{\frac{\lambda^*}{1+\lambda^*}} \left(b + \frac{\lambda^*}{1+\lambda^*} \frac{(z-\alpha^*)^2}{2} \right)^{-\frac{2a^*+1}{2}} \ .$$

Note that the predictive density has the next feature

$$f(z|y) \propto \left(1 + \frac{(z-\alpha^*)^2}{\frac{b^*(1+\lambda^*)}{\lambda^* a^*}} \frac{1}{2a^*} \right)^{-\frac{2a^*+1}{2}} \ .$$

We can deduce that the transformation $(Z - \alpha^*)/\sqrt{\frac{b^*(1+\lambda^*)}{\lambda^* a^*}}$ is t-distributed random variable with $2a^*$ degrees of freedom and the cumulative distribution function of Z given $Y = y$ is

$$F(z|y) = F_{T_{2a^*}} \left(\frac{(z-\alpha^*)^2}{\sqrt{\frac{b^*(1+\lambda^*)}{\lambda^* a^*}}} \right),$$

where $F_{T_{2a^*}}(\cdot)$ is the cumulative distribution function of t distribution with $2a^*$ degrees of freedom. The result is that the random variable $Z|Y = y$ is a generalized Student's t-distribution with the location parameter α^* and the scale parameter $\sqrt{\frac{b^*(1+\lambda^*)}{\lambda^* a^*}}$.

Laplace's approximation of predictive probability distribution function

With the same notation for $\tilde{l}(\mu, \sigma^2)$ and $\tilde{l}_z(\mu, \sigma^2)$ as in previous examples, we write

$$\tilde{l}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2} + \frac{\log(\lambda)}{2} -$$

$$-(a+1)\log(\sigma^2) - \frac{2b + \lambda(\mu - \alpha)^2}{2\sigma^2} - \log\left(\Gamma(a)\right) + a\log(b)$$

and

$$\tilde{l}_z(\mu, \sigma^2) = -\frac{n+1}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (y_i - \mu)^2 + (z-\mu)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2} +$$

$$+\frac{\log(\lambda)}{2} - (a+1)\log(\sigma^2) - \frac{2b + \lambda(\mu - \alpha)^2}{2\sigma^2} - \log\left(\Gamma(a)\right) + a\log(b).$$

The first order condition of $\tilde{l}(\mu, \sigma^2)$ is a system of two equation, from which we find the elements of the mode

$$\begin{cases} \tilde{\mu} = \frac{n\bar{y} + \lambda\alpha}{n+\lambda} \\ \tilde{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \tilde{\mu})^2 + 2b + \lambda(\tilde{\mu} - \alpha)^2}{n + 2a + 3} \end{cases}$$

From other hand, the first order condition of $\tilde{l}_z(\mu, \sigma^2)$ gives the solutions

$$\begin{cases} \tilde{\mu}_z(z) = \frac{n\bar{y} + z + \lambda\alpha}{n+1+\lambda} \\ \tilde{\sigma}_z^2(z) = \frac{\sum_{i=1}^n (y_i - \tilde{\mu}_z(z))^2 + (z - \tilde{\mu}_z(z))^2 + 2b + \lambda(\tilde{\mu}_z(z) - \alpha)^2}{n + 2a + 4} \end{cases}$$

The second derivative of $\tilde{l}(\mu,\sigma^2)$ with respect to (μ,σ^2) with changed sign, $J(\mu,\sigma^2)=\partial^2\tilde{l}(\mu,\sigma^2)/\partial(\mu,\sigma^2)\partial(\mu,\sigma^2)^T$, is the symmetric matrix

$$\begin{pmatrix} \frac{n+\lambda}{\sigma^2} & \frac{\sum_{i=1}^n(y_i-\mu)-\lambda(\mu-\alpha)}{(\sigma^2)^2} \\ - & -\frac{n+2a+3}{2(\sigma^2)^2}+\frac{\sum_{i=1}^n(y_i-\mu)^2+2b+\lambda(\mu-\alpha)^2}{(\sigma^2)^3} \end{pmatrix}$$

The second derivative of $\tilde{l}_z(\mu,\sigma^2)$ with changed sign, $J_z(\mu,\sigma^2)=\partial^2\tilde{l}_z(\mu,\sigma^2)/\partial(\mu,\sigma^2)\partial(\mu,\sigma^2)^T$, is the symmetric matrix

$$\begin{pmatrix} \frac{n+1+\lambda}{\sigma^2} & \frac{\sum_{i=1}^n(y_i-\mu)+(z-\mu)-\lambda(\mu-\alpha)}{(\sigma^2)^2} \\ - & -\frac{n+2a+4}{2(\sigma^2)^2}+\frac{\sum_{i=1}^n(y_i-\mu)^2+(z-\mu)^2+2b+\lambda(\mu-\alpha)^2}{(\sigma^2)^3} \end{pmatrix}$$

We have all we need to implement the Laplace's approximation for predictive density, where for the problem under consideration is

$$\tilde{f}^*(z|y)=\exp\left\{\tilde{l}_z\left(\tilde{\mu}_z(z),\tilde{\sigma}_z^2(z)\right)-\tilde{l}(\tilde{\mu},\tilde{\sigma}^2)\right\}\frac{|J(\tilde{\mu},\tilde{\sigma}^2)|^{1/2}}{|J_z\left(\tilde{\mu}_z(z),\tilde{\sigma}_z^2(z)\right)|^{1/2}}.$$

HOPA method for approximation of predictive cumulative distribution function

The system for the first order condition of $\tilde{l}_z(z,\mu,\sigma^2)$ is

$$\begin{cases} \partial\tilde{l}_z/\partial z=-\frac{z-\mu}{\sigma^2}=0 \\ \partial\tilde{l}_z/\partial\mu=\frac{\sum_{i=1}^n(y_i-\mu)+(z-\mu)}{\sigma^2}-\frac{\lambda(\mu-\alpha}{\sigma^2}=0 \\ \partial\tilde{l}_z/\partial\sigma^2=-\frac{n+1}{2\sigma^2}+\frac{\sum_{i=1}^n(y_i-\mu)+(z-\mu)^2}{2(\sigma^2)^2}-\frac{a+1}{\sigma^2}+\frac{2b+\lambda(\mu-\alpha)^2}{2(\sigma^2)^2}=0 \end{cases}$$

and from which we have to find $\tilde{z}$, an inner point of Z's support, such that $(\tilde{z},\tilde{\mu}_z(\tilde{z}),\tilde{\sigma}_z^2(\tilde{z}))$ is the mode for our function. The solution of the system for z is

$$\tilde{z}=\tilde{\mu}=\frac{n\bar{y}+\lambda\alpha}{n+\lambda}\ .$$

Note that $\tilde{z}\in\mathcal{Z}$ and the additional condition to implement HOPA method holds. Also we can calculate the approximation of the predictive density

function by the formula (2.16). We can get the profile function of z from $\tilde{l}_z$

$$\tilde{l}_p(z) = \tilde{l}_z\left(\tilde{\mu}_z(z), \tilde{\sigma}_z^2(z)\right)$$

which has the first derivative equal with

$$\tilde{l}_p'(z) = -\frac{n+2a+4}{2\tilde{\sigma}_z^2(z)}\tilde{\sigma}_z^{2'}(z) -$$

$$-\frac{2\left[-\sum_{i=1}^n\left(y_i-\tilde{\mu}_z(z)\right)\tilde{\mu}_z'(z)+\left(z-\tilde{\mu}_z(z)\right)\left(1-\tilde{\mu}_z'(z)\right)+\lambda\left(\tilde{\mu}_z(z)-\alpha\right)\tilde{\mu}_z'(z)\right]\tilde{\sigma}_z^2(z)}{2\left(\tilde{\sigma}_z^2(z)\right)^2} -$$

$$-\frac{\left[\sum_{i=1}^n\left(y_i-\tilde{\mu}_z(z)\right)^2+\left(z-\tilde{\mu}_z(z)\right)^2+2b+\lambda\left(\tilde{\mu}_z(z)-\alpha\right)^2\right]\tilde{\sigma}_z^{2'}(z)}{2\left(\tilde{\sigma}_z^2(z)\right)^2},$$

where $\tilde{\sigma}_z^{2'}(z)$ and $\tilde{\mu}_z'(z)$ are the first derivatives of $\tilde{\sigma}_z^2(z)$ and $\tilde{\mu}_z(z)$ with respect to z and they are equal with

$$\tilde{\mu}_z'(z) = \frac{1}{n+1+\lambda}$$

$$\tilde{\sigma}_z^{2'}(z) = 2\frac{-\sum_{i=1}^n\left(y_i-\tilde{\mu}_z(z)\right)\tilde{\mu}_z'(z)+\left(z-\tilde{\mu}_z(z)\right)\left(1-\tilde{\mu}_z'(z)\right)+\lambda(\tilde{\mu}_z(z)-\alpha)\tilde{\mu}_z'(z)}{n+2a+4}.$$

The second derivative of $\tilde{l}_z(\mu,\sigma^2)$ with respect to (z,μ,σ^2) with changed sign, $\tilde{J}_z(z,\mu,\sigma^2)$, is the symmetric matrix 3×3

$$\begin{pmatrix} \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} & -\frac{z-\mu}{(\sigma^2)^2} \\ - & \frac{n+\lambda+1}{\sigma^2} & \frac{\sum_{i=1}^n(y_i-\mu)+(z-\mu)-\lambda(\mu-\alpha)}{(\sigma^2)^2} \\ - & - & -\frac{n+2a+4}{2(\sigma^2)^2}+\frac{\sum_{i=1}^n(y_i-\mu)^2+(z-\mu)^2+2b+\lambda(\mu-\alpha)^2}{(\sigma^2)^3} \end{pmatrix}$$

With the quantities above we can implement the HOPA method (see Section 3.3).

Numerical illustration

We simulate an i.i.d sample of size 30 from a normal distribution with mean equal 0 and standard deviation equal 1. We set the joint prior distribution of the mean and variance $N-\Gamma^{-1}(1,1,1,1)$. Table 4.3 provides the

main summaries of predictive random variable $Z|Y = y$ calculated from exact distribution and approximated by HOPA method and HOPA simulation based on $r^*(z)$ and $r^{**}(z)$. The Figure 4.3 contains the plots to comparison exact and approximation results after implementation of Laplace's and HOPA methods.

Table 4.3: Predictive summaries of normal model with unknown mean and variance using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
Exact	-1.55	-0.01	1.54	-0.01	0.83	[-1.55 , 1.54]
HOPA1	-1.55	-0.01	1.53	-	-	-
HOPA1 sim.(10^6)	-1.54	0	1.54	0	0.88	[-1.55 , 1.53]
HOPA2	-1.55	-0.01	1.53	-	-	-
HOPA2 sim.(10^6)	-1.54	0	1.54	0	0.88	[-1.54 , 1.54]

4.3 Gamma model

4.3.1 Gamma model with unknown rate parameter

Suppose that $y = (y_1, \ldots, y_n)$ is an i.i.d. sample from $Y \sim Gamma(\alpha, \lambda)$, where α is known shape parameter and λ is an unknown rate parameter. Let the prior distribution of the parameter λ be $Gamma(a, b)$, where a is the shape hyperparameter and b is the rate hyperparameter. The likelihood function based on sample y is

$$L(\lambda) = \frac{1}{\Gamma(\alpha)^n}\lambda^{n\alpha}\left(\prod y_i\right)^{\alpha-1} e^{-\lambda\sum_{i=1}^n y_i}.$$

Using the prior density of the parameter λ,

$$\pi(\lambda) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}e^{-b\lambda} \ ,$$

and the likelihood function, we can obtain the posterior density of λ, which is

$$\pi(\lambda|y) \propto \lambda^{n\alpha+a-1}e^{-(b+n\bar{y})\lambda}.$$

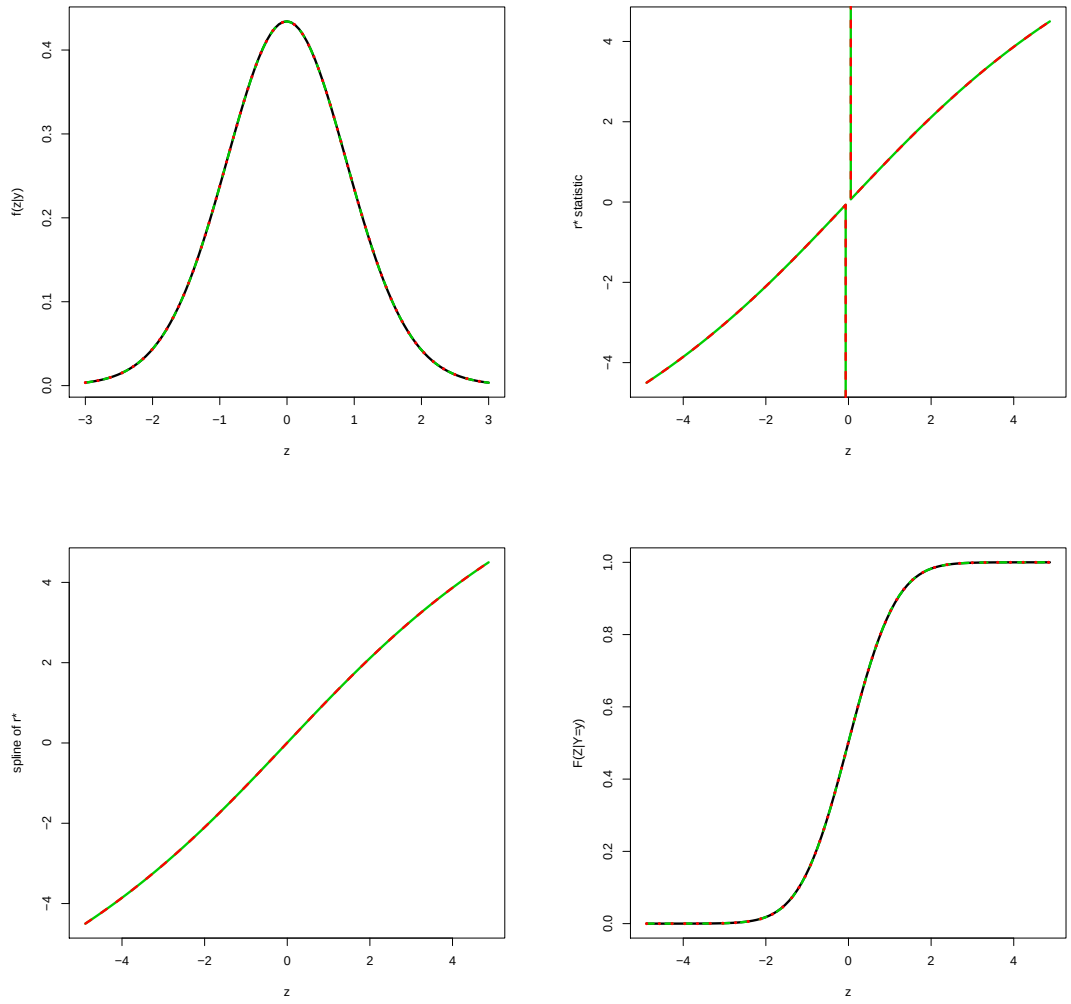

Figure 4.3: Normal model with unknown mean and variance. *Top-left:* The predictive density functions: exact (black-solid line), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = \epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: exact (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

Observe that the posterior density function of the parameter λ is $Gamma(n\alpha+a, b+n\bar{y})$. Assume a new observation Z independent from Y. We can deduce the exact form for predictive distribution function $Z|Y=y$ as follows

$$f(z|y) = \frac{z^{\alpha-1}}{(b+n\bar{y}+z)^{(n+1)\alpha+a}} \frac{(b+n\bar{y})^{n\alpha+a}\Gamma((n+1)\alpha+a)}{\Gamma(\alpha)\Gamma(n\alpha+a)}.$$

Only for specific α we can obtain the cumulative distribution function of z given y in closed form. One of these cases is $\alpha = 2$. Under this assumption the density function becomes

$$f(z|y) = \frac{z}{(b+n\bar{y}+z)^{2n+a+2}} \frac{(b+n\bar{y})^{2n+a}\Gamma(2n+a+2)}{\Gamma(2n+a)}.$$

$$= \frac{z}{(b+n\bar{y}+z)^{2n+a+2}} (b+n\bar{y})^{2n+a}(2n+a+1)(2n+a)$$

and the cumulative distribution function of z given y is

$$F(z|y) = 1 - c\left[\frac{z}{(2n+a+1)(b+n\bar{y}+z)^{2n+a+1}} + \frac{1}{(2n+a+1)(2n+a)(b+n\bar{y}+z)^{2n+a}}\right],$$

where $c = (b+n\bar{y})^{2n+a}(2n+a+1)(2n+a)$.

Laplace's approximation of predictive probability distribution function

As in Section (2.5) we define $\tilde{l}(\lambda)$ and $\tilde{l}_z(\lambda)$ and we can write

$$\tilde{l}(\lambda) = (\alpha n+a-1)\log(\lambda) - (b+n\bar{y})\lambda + (\alpha-1)\sum_{i=1}^{n}\log(y_i) - n\log\left(\Gamma(\alpha)\right) + a\log(b) - \log\left(\Gamma(a)\right)$$

and

$$\tilde{l}_z(\lambda) = (\alpha(n+1)+a-1)\log(\lambda) - (b+n\bar{y}+z)\lambda + (\alpha-1)\left(\log(z) + \sum_{i=1}^{n}\log(y_i)\right) -$$

$$-(n+1)\log\left(\Gamma(\alpha)\right) + a\log(b) - \log\left(\Gamma(a)\right).$$

The derivatives of first order for $\tilde{l}(\lambda)$ and $\tilde{l}_z(\lambda)$ are equal zero for

$$\tilde{\lambda} = \frac{\alpha n + a - 1}{b + n\bar{y}}$$

for first function, and

$$\tilde{\lambda}_z(z) = \frac{\alpha(n+1) + a - 1}{b + n\bar{y} + z}$$

for second function. The derivatives of second order for $\tilde{l}(\lambda)$ and $\tilde{l}_z(\lambda)$ with changed sign are

$$J(\lambda) = \frac{\alpha n + a - 1}{\lambda^2}$$

and

$$J_z(\lambda) = \frac{\alpha(n+1) + a - 1}{\lambda^2}.$$

We can to implement the Laplace's approximation for predictive density for z given y (2.15), which in our specific case becomes

$$\tilde{f}^*(z|y) = \exp\left\{\tilde{l}_z\left(\tilde{\lambda}_z(z)\right) - \tilde{l}(\tilde{\lambda})\right\} \frac{\left|J(\tilde{\lambda})\right|^{1/2}}{\left|J_z\left(\tilde{\lambda}_z(z)\right)\right|^{1/2}}. \tag{4.3}$$

HOPA method for approximation of predictive cumulative distribution function

To implement HOPA we have to find $\tilde{z}$, an inner point of Z's support, such that $\left(\tilde{z}, \tilde{\lambda}_z(\tilde{z})\right)$ is the mode for $\tilde{l}_z(z, \lambda)$. First order condition is

$$\begin{cases} \partial\tilde{l}_z/\partial\lambda = (\alpha(n+1) + a - 1)/\lambda - (b + n\bar{y} + z) = 0 \\ \partial\tilde{l}_z/\partial z = -\lambda + \frac{\alpha-1}{z} = 0 \end{cases}$$

and the solution of this system is

$$\begin{cases} \tilde{z} = \frac{(\alpha-1)(b+n\bar{y})}{\alpha n+a} \\ \lambda_z(\tilde{z}) = \frac{\alpha(n+1)+a-1}{b+n\bar{y}+\tilde{z}} = \frac{\alpha n+a}{b+n\bar{y}} \end{cases}$$

Note that $\tilde{z} \in \mathcal{Z} = (0, +\infty)$ if only if $\alpha > 1$. In particular, for $\alpha = 1$ the gamma distribution becomes an exponential distribution and the additional condition to implement HOPA method does not hold. So the HOPA method, with the settings of the problem above, can be implemented only in the case we have $\alpha > 1$. Knowing the function $\tilde{\lambda}_z(z)$, we can write the profile function of z from $\tilde{l}_z$

$$\tilde{l}_p(z) = (\alpha(n+1) + a - 1) \log\left(\frac{\alpha(n+1) + a - 1}{b + n\bar{y} + z}\right) + (\alpha - 1)\log(z) + \text{const} \ ,$$

which has the first derivative equal with

$$\tilde{l}'_p(z) = -\frac{\alpha(n+1) + a - 1}{b + n\bar{y} + z} + \frac{\alpha - 1}{z}.$$

The second derivative of $\tilde{l}_z(\lambda)$ with respect to (z, λ) with changed sign, $\tilde{J}_z(z, \lambda)$, is the symmetric matrix 2×2

$$\begin{pmatrix} \frac{\alpha(n+1)+a-1}{\lambda^2} & 1 \\ 1 & \frac{\alpha-1}{z^2} \end{pmatrix}$$

We can calculate $r^*(z)$ and $r^{**}(z)$ functions, with there components $r(z)$, $q^*(z)$ and $q^{**}(z)$. The approximation for cumulative distribution function for z given y is

$$F(z|y) \approx \Phi\left(r^*(z)\right) \text{ or } F(z|y) \approx \Phi\left(r^{**}(z)\right) \ .$$

Numerical illustration

A sample of size 10 that is simulated from a $Gamma(4, 2)$. Suppose we know the shape parameter of gamma distribution and set gamma distribution as the prior distribution for the rate parameter, with shape hyperparameters equal 1 and rate hyperparameter equal 1. Table 4.4 contains the main summaries of predictive random variable $Z|Y = y$ calculated with HOPA method and approximated by MCMC and HOPA simulations. The plots in Figure 4.4 show the intermediate and final results of the implementation of Laplace's and HOPA methods.

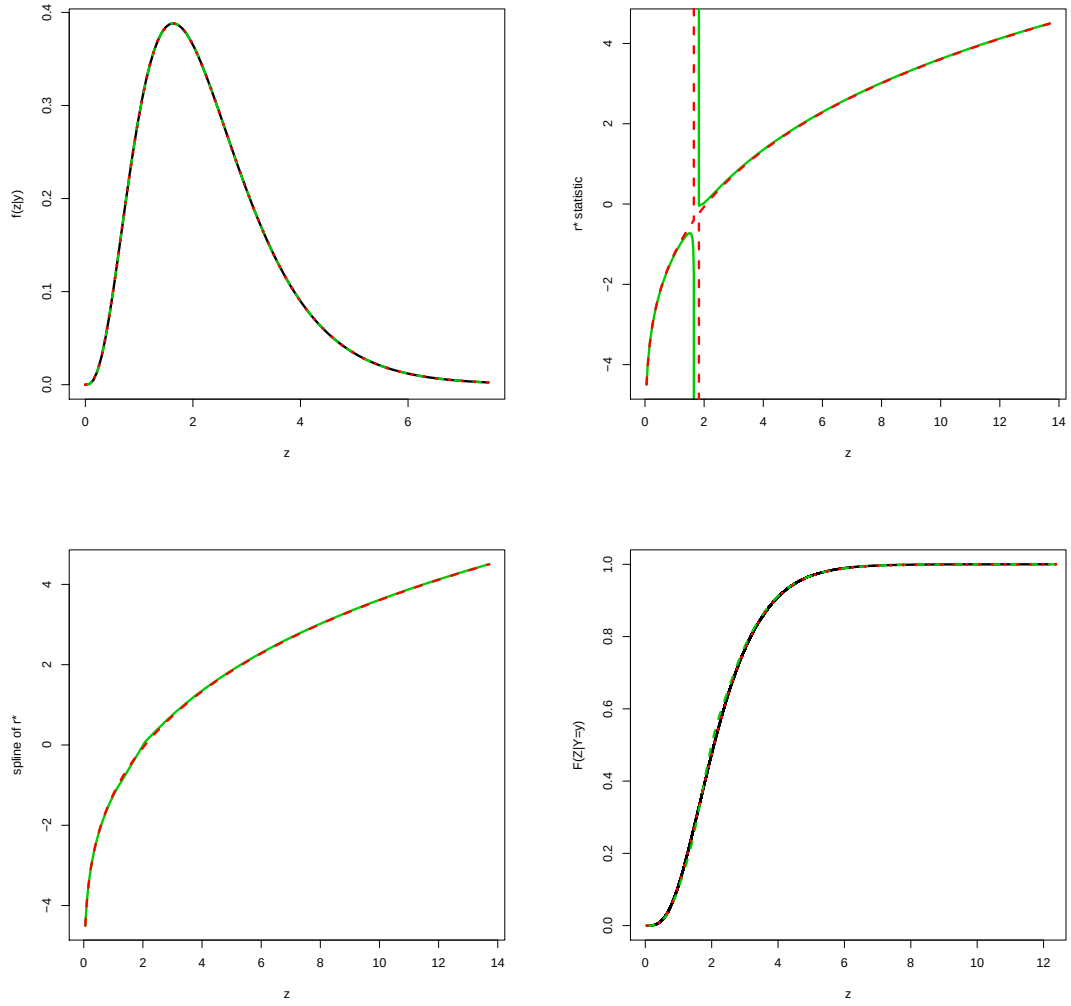

Figure 4.4: Gamma model with unknown rate parameter. *Top-left:* The predictive density functions: exact (black-solid line), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = 0.4$ and $\epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

Table 4.4: Predictive summaries of gamma model with unknown rate parameter using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	0.75	2.07	4.57	2.29	1.45	[0.5 , 4.06]
HOPA1	0.76	1.99	4.56	-	-	-
HOPA1 sim.(10^6)	0.77	1.99	4.55	2.27	1.44	[0.49 , 4]
HOPA2	0.75	2.07	4.59	-	-	-
HOPA2 sim.(10^6)	0.75	2.07	4.6	2.29	1.48	[0.49 , 4.05]

Alternative solution to overcome the problem when shape parameter is lower than 1

As we mentioned in Chapter 3.3 the trick to implement HOPA method in this case is to transform the predictive variable of interest Z into random variable $V = \log Z$. Suppose the case we have the same y realisation of Gamma and V is log-Gamma, with the same parameters. In the new settings we have that $\tilde{l}_v(\lambda)$ becomes

$$\tilde{l}_v(\lambda) = (\alpha(n+1)+a-1)\log(\lambda) - (b+n\bar{y}+e^v)\lambda + (\alpha-1)\sum_{i=1}^{n}\log(y_i) + \alpha v -$$

$$-(n+1)\log\left(\Gamma(\alpha)\right) + a\log(b) - \log\left(\Gamma(a)\right).$$

Now the first order condition of $\tilde{l}_v(\lambda)$ is the system

$$\begin{cases} \partial\tilde{l}_v/\partial\lambda = (\alpha(n+1)+a-1)/\lambda - (b+n\bar{y}+e^v) = 0 \\ \partial\tilde{l}_v/\partial v = -\lambda e^v + \alpha = 0 \end{cases}$$

that has this intermediate solution

$$\begin{cases} \lambda = \frac{\alpha(n+1)+a-1}{b+n\bar{y}+e^v} \\ e^v = \alpha/\lambda \end{cases}$$

The solution of the system with respect to v is

$$\tilde{v} = \log\frac{\alpha(b+n\hat{y})}{n\alpha+a-1}\ .$$

The solution exists if only if $n\alpha+a-1>0$. Note that this is the condition of positivity of logarithm argument and the condition of positivity for solution of λ. The condition holds if $a \geq 1$, i.e. the prior distribution is bounded. Otherwise, for $a<1$, for any α there exist n_0 such that $n\alpha+a-1>0$ for $n>n_0$. This is a mild condition to have $\tilde{z}$ and to implement HOPA method. The matrix J does not change and $J_v = J_z$. The profile function of v changes in this way

$$\tilde{l}_p(v) = (\alpha(n+1)+a-1)\log\left(\frac{\alpha(n+1)+a-1}{b+n\bar{y}+e^v}\right)+\alpha v+ \ ,$$

which has the first derivative equal with

$$\tilde{l}'_p(v) = -e^v\frac{\alpha(n+1)+a-1}{b+n\bar{y}+e^v}+\alpha.$$

The second derivative of $\tilde{l}_v(\lambda)$ with respect to (v,λ) with changed sign, $\tilde{J}_v(v,\lambda)$, is the matrix

$$\begin{pmatrix} \frac{\alpha(n+1)+a-1}{\lambda^2} & e^v \\ e^v & \lambda e^v \end{pmatrix}$$

We can now redifine the quantities $r^*(v)$ and $r^{**}(v)$ using the formulas above.

Numerical illustration of HOPA method for transformed predictive random variable V

The dataset is an i.i.d sample of size 10 simulated from a Gamma random variable with shape parameter equal 0.5 and rate parameter equal 1. We set the prior distribution of rate parameter $Gamma(1,1)$. The condition to implement the method holds because $n\alpha+a-1=5>0$, or because for $a=1$ the prior distribution of the parameter is bounded. Table 4.5 contains the main summaries of predictive random variable $V|Y=y$ calculated from MCMC simulation and approximated by HOPA methods and HOPA simulations based on $r^*(z)$ and $r^{**}(z)$. Using formulas (3.7) and (3.8) and the inverse transformation $Z=e^V$, we can compute summaries for predictive

random variable $Z|Y = y$, which are exposed in Table 4.6. The plots in Figures 4.5 and 4.6 show the intermediate and final results of the implementation of the listed methods and the approximations of predictive probability distribution functions $\tilde{f}^*$ and $\tilde{f}^{**}$. Note that the numerical irregularity for the second method has a local character and it has a better accuracy.

Table 4.5: Predictive summaries of log-Gamma distribution with unknown rate parameter and shape parameter lower than 1 using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	-6.48	-1.69	0.58	-2.15	5.15	[-5.46 , 1.05]
HOPA1	-6.01	-1.36	0.24	-	-	-
HOPA1 sim.(10^6)	-6.01	-1.36	0.24	-1.91	4.21	[-4.85 , 0.8]
HOPA2	-6.6	-1.71	0.6	-	-	-
HOPA2 sim.(10^6)	-6.6	-1.71	0.6	-2.18	5.32	[-5.52 , 1.12]

Table 4.6: Predictive summaries of Gamma distribution with unknown rate parameter and shape parameter lower than 1 using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	0.0015	0.1839	1.7869	0.4568	0.5787	[0 , 1.22]
HOPA1	0.0024	0.2577	1.2734	-	-	-
HOPA1 sim.(10^6)	0.0025	0.2573	1.2724	0.4083	0.3612	[0 , 0.88]
HOPA2	0.0014	0.1814	1.8275	-	-	-
HOPA2 sim.(10^6)	0.0014	0.1815	1.8255	0.4597	0.5877	[0 , 1.22]

4.3.2 Gamma model with unknown parameters

Let $y = (y_1, \ldots, y_n)$ is an independent and identicaly distributed sample from $Y \sim Gamma(\alpha, \lambda)$, where α is unknown shape parameter and λ is unknown rate parameter. Suppose that the parameters α and λ are independent and their prior distributions are $Gamma(a, b)$ and $Gamma(c, d)$, respectively. Then the likelihood function is

$$L(\alpha, \lambda) = \frac{1}{\Gamma(\alpha)^n} \lambda^{n\alpha} \left(\prod y_i \right)^{\alpha - 1} e^{-\lambda \sum_{i=1}^{n} y_i} ,$$

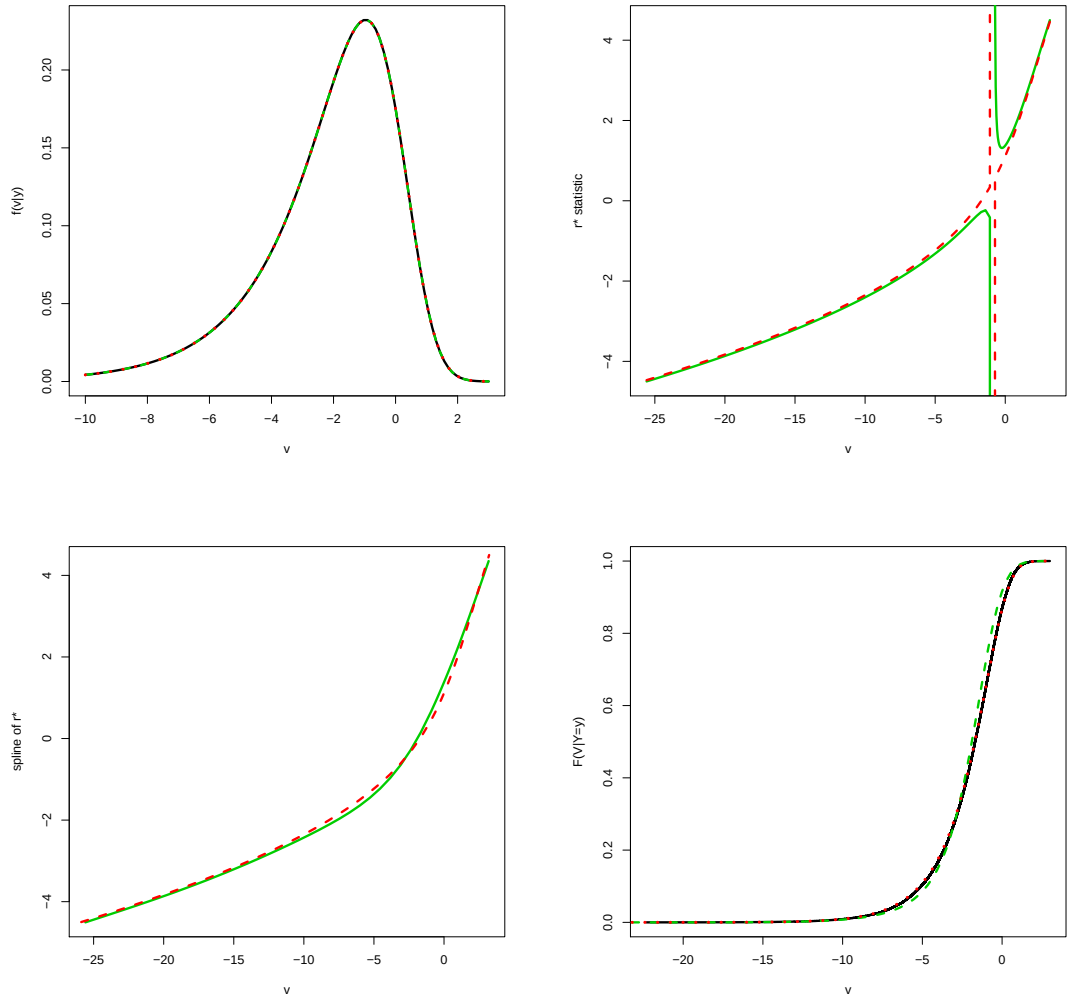

Figure 4.5: The log-gamma distribution with unknown rate parameter and shape parameter lower than 1 . *Top-left:* The predictive density functions: exact (black-solid line), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = \epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

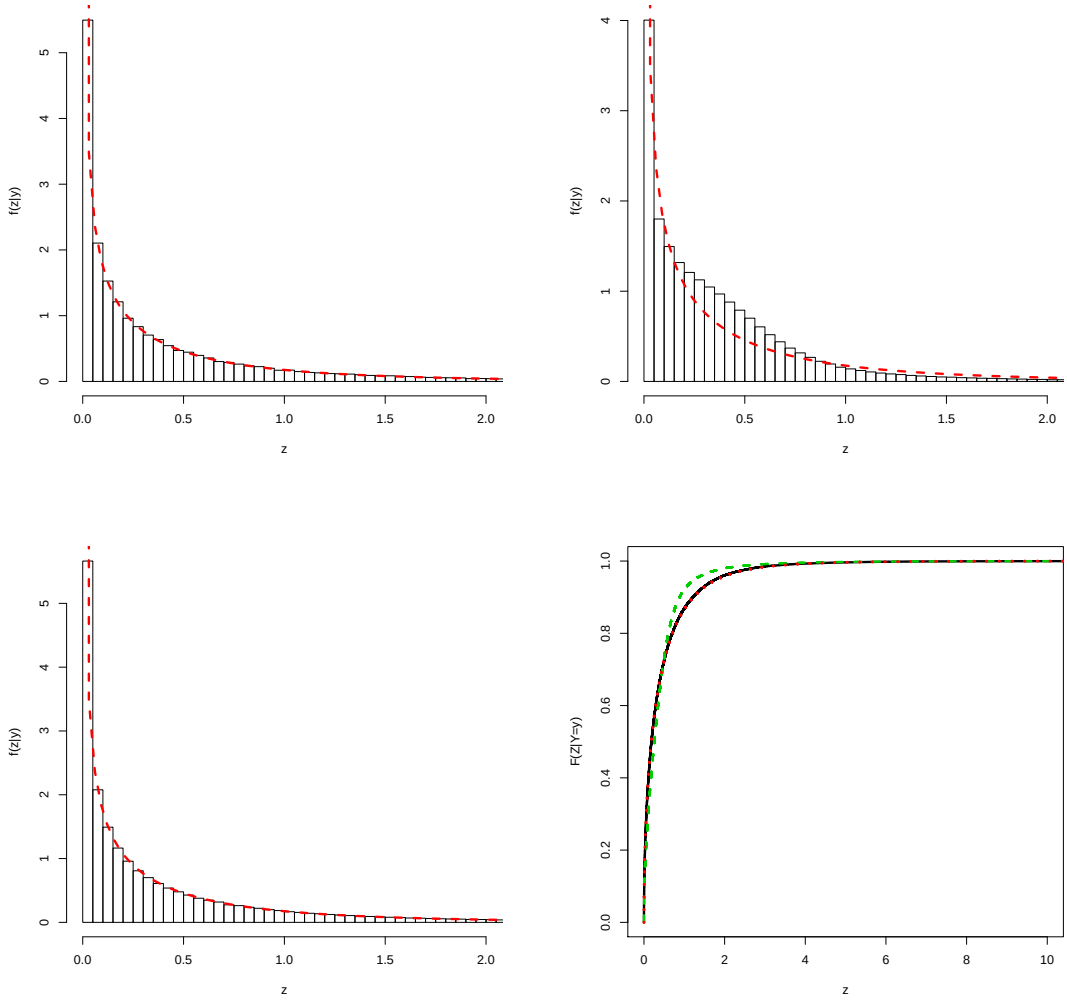

Figure 4.6: Gamma distribution with unknown rate parameter and shape parameter lower than 1. *Top-left:* The predictive density functions: exact (red-dashed line) and MCMC simulation (histogram). *Top-right:* The predictive density functions: exact (red-dashed line) and HOPA1 simulation (histogram). *Bottom-left*: The predictive density functions: exact (red-dashed line) and HOPA2 simulation (histogram). *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

and the prior density of the parameters (α, λ) is

$$\pi(\alpha,\lambda) = \frac{b^a}{\Gamma(a)}\alpha^{a-1}e^{-b\alpha}\frac{d^c}{\Gamma(c)}\lambda^{c-1}e^{-d\lambda} \ .$$

Using the functions above, we can obtain the posterior density of the parameters, which is

$$\pi(\alpha,\lambda|y) \propto \frac{\alpha^{a-1}}{\Gamma(\alpha)^n}\lambda^{n\alpha}\left(\prod y_i\right)^{\alpha-1}\lambda^{n\alpha+c-1}e^{-(d+n\bar{y})\lambda-b\alpha}.$$

The posterior density has no notable distribution. However, a feature of posterior density is that the posterior density function of the parameter λ given α is $Gamma(n\alpha+c, d+n\bar{y})$. This result will be used to implement Metropolis within Gibbs method for simulating a sample from posterior density.

Assume a new observation Z, which is independent from Y. The predictive density function of z given y is

$$f(z|y) = \int_0^{+\infty}\int_0^{+\infty} f(z|\alpha,\lambda)f(\alpha,\lambda|y)d\alpha d\lambda.$$

We can not deduce a closed form for this integral, hence, for the predictive density. For this reason, the two-step simulation method will be used to get a sample from the random variable Z given $Y = y$ and further to approximate both the density function and the cumulative distribution function. The first step is the mentioned Metropolis within Gibbs simulation from posterior density of parameters, $\pi(\alpha, \lambda|y)$. The second step uses the sample from first stage as parameters for $Gamma(\alpha, \lambda)$ to simulate from this distribution the last sample. The elements of this sample are realisations from predictive distribution.

Wishing to implement HOPA method, we will deduce all the need quantities from $\tilde{l}(\alpha, \lambda)$ and $\tilde{l}_z(\alpha, \lambda)$ using numerical methods, where

$$\tilde{l}(\alpha,\lambda) = (\alpha n + c - 1)\log(\lambda) - (d + n\bar{y})\lambda + (\alpha - 1)\sum_{i=1}^{n}\log(y_i) - n\log\left(\Gamma(\alpha)\right) +$$

$$+a\log(b) - \log\left(\Gamma(a)\right) + (a-1)\log(\alpha) - b\alpha + c\log(d) - \log\left(\Gamma(c)\right)$$

and

$$\tilde{l}_z(\alpha,\lambda) = (\alpha(n+1)+c-1)\log(\lambda)-(d+n\bar{y}+z)\lambda+(\alpha-1)\left(\sum_{i=1}^{n}\log(y_i)+\log(z)\right)-$$

$$-(n+1)\log\left(\Gamma(\alpha)\right)+a\log(b)-\log\left(\Gamma(a)\right)+(a-1)\log(\alpha)-b\alpha+c\log(d)-\log\left(\Gamma(c)\right) \ .$$

In order to know if the additional condition to implement HOPA method holds, we will compute numerically $\tilde{z}$ and will check if it is an inner point of $\mathcal{Z} = [0, +\infty)$ or will set the prior distribution for parameters to be bounded.

Numerical illustration

The data consists on a sample of size 20 simulated from a gamma distribution with shape parameter $\alpha = 8$ and rate parameter $\lambda = 2$. Suppose we know only the distribution. Firstly we have setted the prior distribution of both parameters $Gamma(1,1)$. Note that the probability distribution function of $Gamma(1,1)$ is bounded and we can implement HOPA method and to use approximation $\tilde{f}^*$. Indeed, we calculated $\tilde{z} = 3.07$. The summaries of predictive random variable $Z|Y = y$ from MCMC and HOPA methods are displayed in Table 4.7. The plots in Figure 4.7 show all the results.

Table 4.7: Predictive summaries of gamma model with unknown parameters using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	1.45	3.69	7.57	4	3.71	[1.06 , 6.83]
HOPA1	1.45	3.8	7.61	-	-	-
HOPA1 sim.(10^6)	1.45	3.8	7.62	4.03	3.81	[1.06 , 6.89]
HOPA2	1.46	3.71	7.59	-	-	-
HOPA2 sim.(10^6)	1.46	3.71	7.59	4.01	3.75	[1.07 , 6.87]

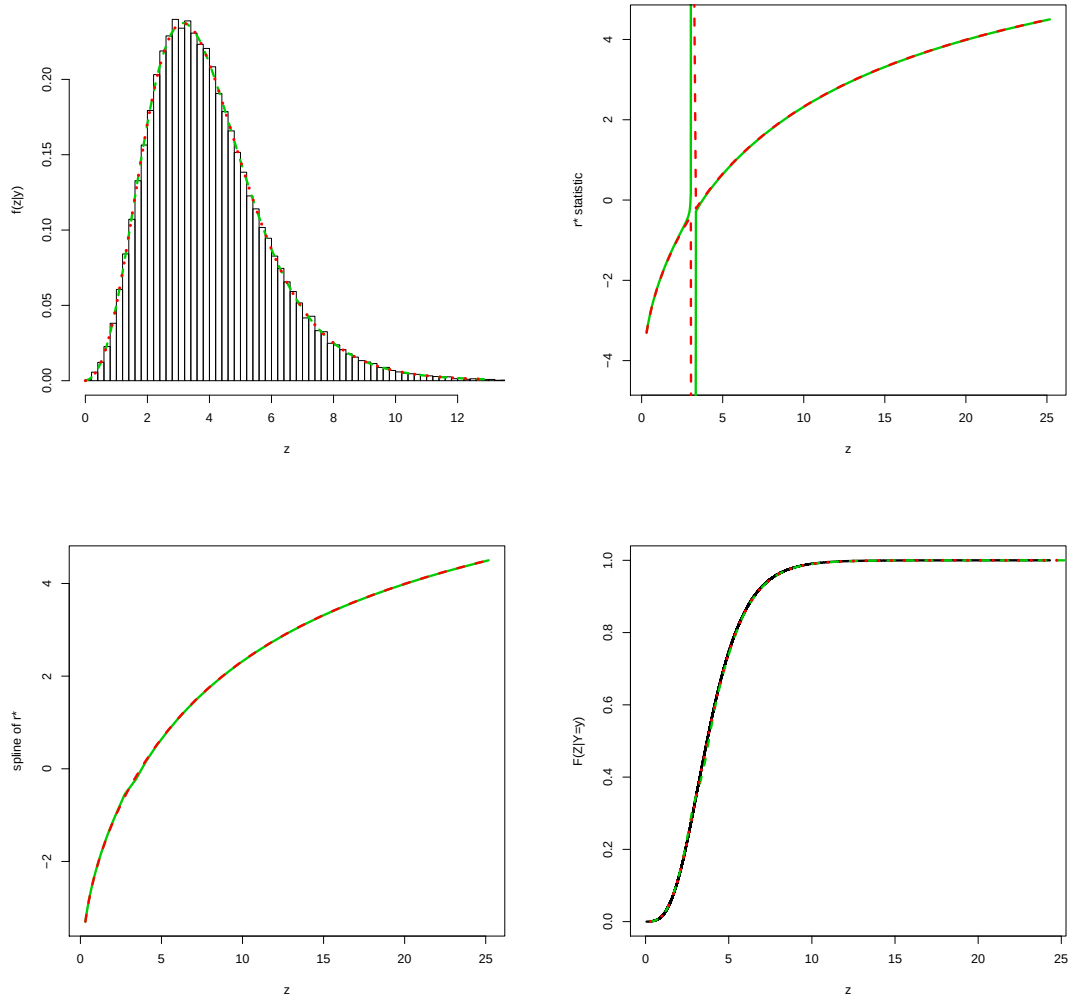

Figure 4.7: **Gamma model with unknown parameters.** *Top-left:* The predictive density functions: MCMC simulation (histogram), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = 0.2$ and $\epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

Practical implementation to overcome the problem when $\tilde{z}$ can not be finded

We proceed with the transformation $V = \log Z$. The functions of interest become

$$\tilde{l}(\alpha,\lambda) = (\alpha n + c - 1)\log(\lambda) - (d + n\bar{y})\lambda + (\alpha - 1)\sum_{i=1}^{n}\log(y_i) - n\log\left(\Gamma(\alpha)\right) +$$

$$+a\log(b) - \log\left(\Gamma(a)\right) + (a-1)\log(\alpha) - b\alpha + c\log(d) - \log\left(\Gamma(c)\right)$$

and

$$\tilde{l}_v(\alpha,\lambda) = (\alpha(n+1) + c - 1)\log(\lambda) - (d + n\bar{y} + e^v)\lambda + (\alpha-1)\sum_{i=1}^{n}\log(y_i) + \alpha v -$$

$$-(n+1)\log\left(\Gamma(\alpha)\right) + a\log(b) - \log\left(\Gamma(a)\right) + (a-1)\log(\alpha) - b\alpha + c\log(d) - \log\left(\Gamma(c)\right) \ .$$

We should solve the system

$$\begin{cases} \partial\tilde{l}_v/\partial\alpha = 0 \\ \partial\tilde{l}_v/\partial\lambda = 0 \\ \partial\tilde{l}_v/\partial v = 0 \end{cases}$$

to conclude if the additional condition holds. An intermediate solution of this system is

$$\begin{cases} v = \log\frac{\alpha(d+n\hat{y})}{n\alpha+c-1} \\ \lambda = \frac{n\alpha+c-1}{d+n\hat{y}} \\ n\log\frac{n\alpha+c-1}{d+n\hat{y}} + \sum_{i=1}^{n}\log(y_i) + \log\alpha - \frac{n+1}{\Gamma(\alpha)}\Gamma'(\alpha) + \frac{a-1}{\alpha} - b = 0 \end{cases}$$

Note that one of the conditions for the existence of the mode $(\tilde{v}, \tilde{\alpha}_v(\tilde{v}), \tilde{\lambda}_v(\tilde{v}))$, as an inner point of $\mathbb{R} \times (0,+\infty)^2$, is the same condition as for log-gamma with unknown rate parameter, i.e. $n\alpha + c - 1 > 0$. The second condition is the existence of $\tilde{\alpha}$ that is the solution for the third equation of the system.

We consider a simulated sample of size 30 from $Gamma(0.5, 1)$. The prior distribution is $Gamma(1, 1)$ for both α and λ parameters, which is a

bounded function. So we conclude that the additional condition holds and we can apply HOPA method. After the implementation of MCMC method, HOPA methods based on $r^*(z)$ and $r^{**}(z)$ for the predictive random variable $V|Y = y$ we have $k = 0.216$ and the next summaries that are showed in Table 4.8 and in Figure 4.8. Implementing the inverse transformation from V to Z, we obtain the results showed in Table 4.9 and in Figure 4.9. We can note that the numerical irregularity for the second method has a local character and it has a better accuracy.

Table 4.8: Predictive summaries of log-gamma distribution with unknown parameters using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	-7.64	-1.82	0.58	-2.46	7.36	[-6.23 , 1.13]
HOPA1	-1.46	6.03	5.32	-	-	-
HOPA1 sim.(10^6)	-1.46	5.68	6.4	4.55	8.14	[1.21 , 6.4]
HOPA2	-7.52	-1.79	0.59	-	-	-
HOPA2 sim.(10^6)	-7.55	-1.79	0.59	-2.42	7.12	[-6.11 , 1.2]

Table 4.9: Predictive summaries of gamma model with unknown parameters using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	5e-04	0.162	1.7849	0.438	0.5189	[0 , 1.21]
HOPA1	0.23	415.1	205	-	-	-
HOPA1 sim.(10^6)	0.23	293.4	599.8	293.3	48581.9	[0 , 588.9]
HOPA2	5e-04	0.166	1.812	-	-	-
HOPA2 sim.(10^6)	5e-04	0.166	1.8048	0.444	0.5182	[0 , 1.22]

4.4 Stationary first order autoregressive process AR(1)

Suppose that $y = (y_1, \ldots, y_T)$ is a sample from a stationary first order autoregressive process, where $Y_t = \rho Y_{t-1} + \epsilon_t$, with $\epsilon_i \sim N(0, \sigma^2)$ for $i =$

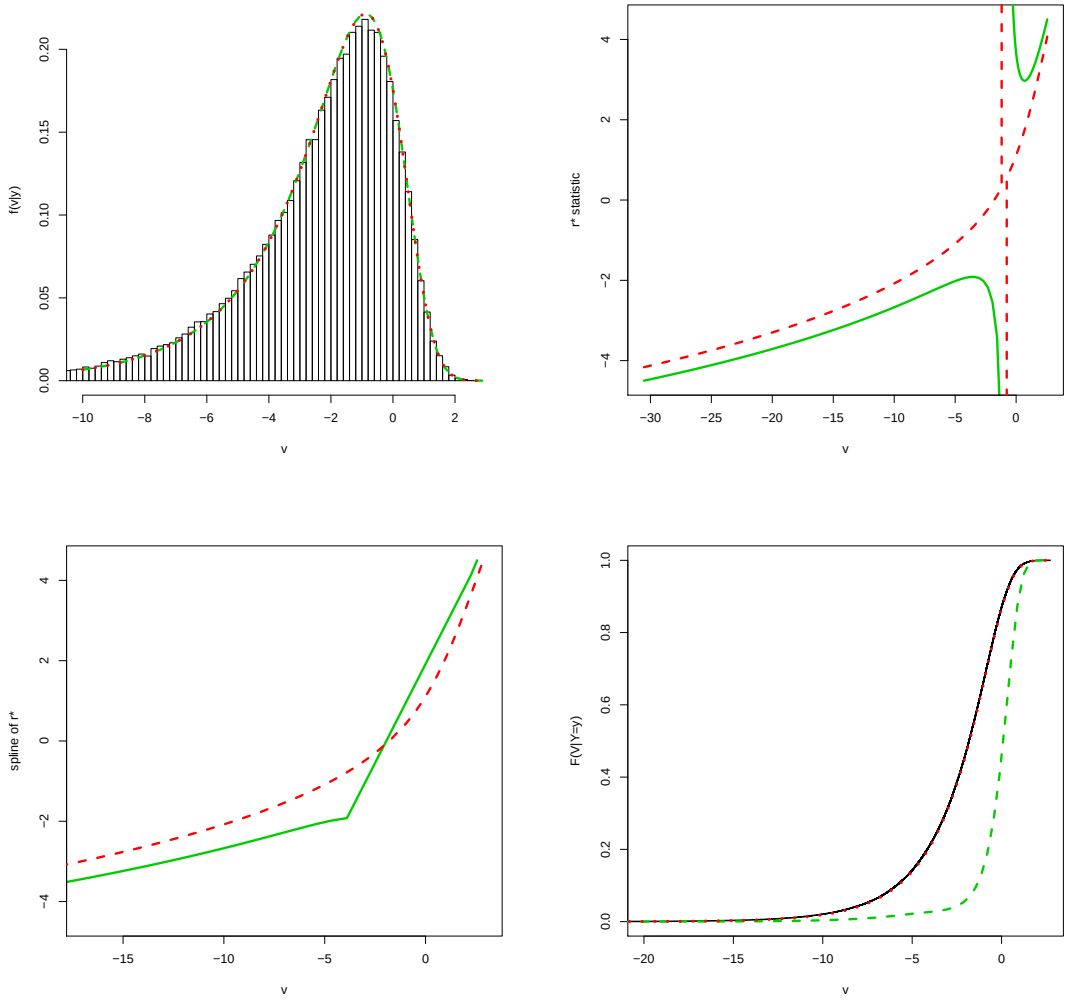

Figure 4.8: Log-gamma distribution with unknown parameters. *Top-left:* The predictive density functions: MCMC simulation (histogram), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = 1.5$ and $\epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

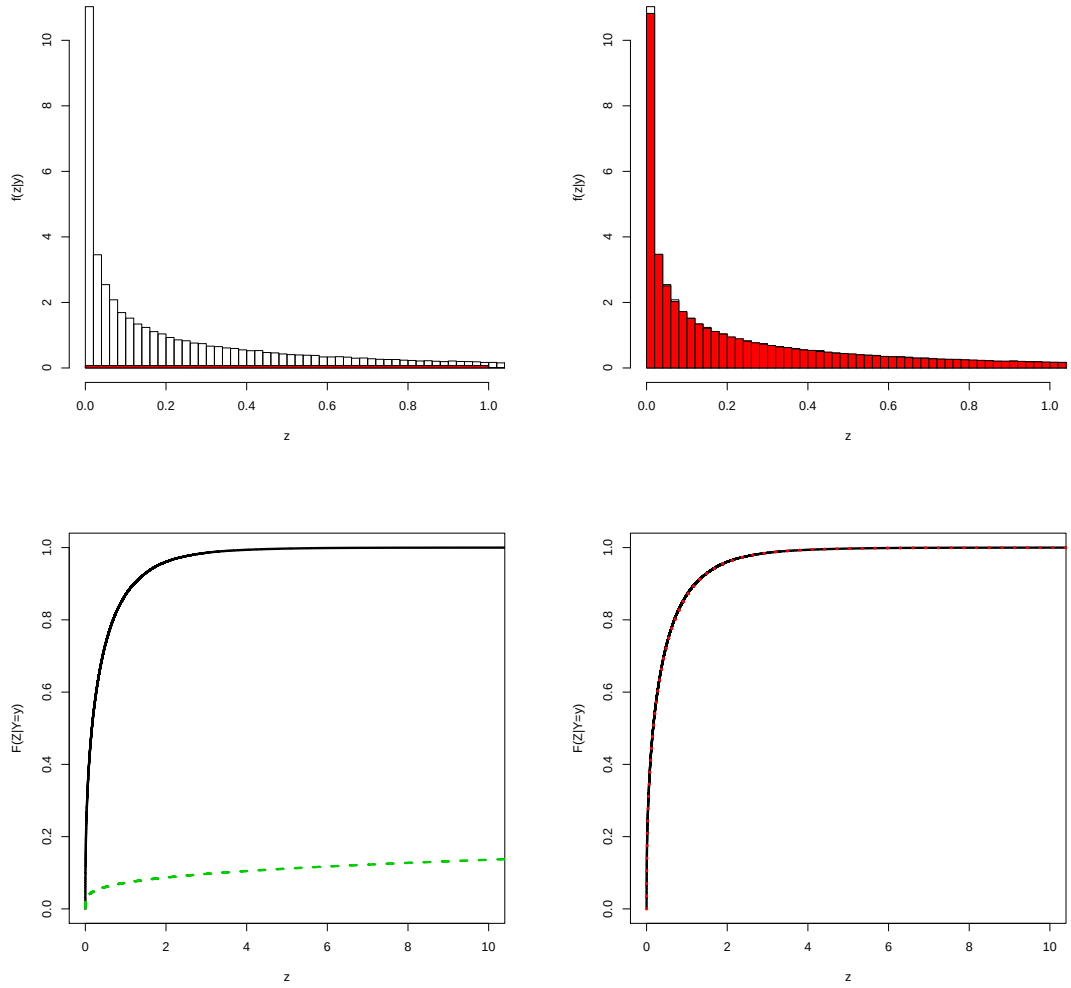

Figure 4.9: Gamma model with unknown parameters. *Top-left:* The predictive density functions: MCMC simulation (histogram) and HOPA1 simulation (red histogram). *Top-right:* MCMC simulation (histogram) and HOPA2 simulation (red histogram). *Bottom-left*: MCMC simulation (black-solid line) and HOPA based on $r^*(z)$ (green-dashed line). *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line) and HOPA based on $r^{**}(z)$ (red-dotted line).

$2, \ldots, T$. Let $Y_1 \sim N(0, \sigma^2/(1-\rho^2))$, then the likelihood function is

$$L(\sigma^2, \rho) = \phi\left(\frac{y_1\sqrt{1-\rho^2}}{\sigma}\right)\prod_{t=2}^{T}\phi\left(\frac{y_t - \rho y_{t-1}}{\sigma}\right) .$$

Let the parameters σ^2 and ρ be a priori are independent and with prior densities $Gamma(a, b)$ and $Unif(-1, 1)$, respectively. So the prior density of the vector of the parameters is

$$\pi(\sigma^2, \rho) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{a-1}e^{-b\sigma^2}I_{(0,+\infty)}(\sigma^2)\frac{1}{2}I_{(-1,1)}(\rho) .$$

The posterior probability density of the parameters (ρ, σ^2) has not a closed form and to approximate this density we will use a random walk Metropolis-Hastings simulation algorithm.

Assume a new observation Z at time $T+1$. We can not find a closed form for the probability distribution function of the random variable $Z|Y = y$. For this reason, the two-step simulation method will be used to generate a sample from the random variable Z given $Y = y$, which we will use to approximate both the density function and the cumulative distribution function.

For the implementetion of the HOPA method, we will use numerical methods to compute the needed quantities, starting from

$$\tilde{l}(\sigma^2, \rho) = \log\left\{\phi\left(\frac{y_1\sqrt{1-\rho^2}}{\sigma}\right)\right\} + \sum_{t=2}^{T}\log\left\{\phi\left(\frac{y_t - \rho y_{t-1}}{\sigma}\right)\right\} + a\log(b) -$$

$$- \log\left(\Gamma(a)\right) + (a-1)\log(\sigma^2) - b\sigma^2 + \log\left(\frac{1}{2}I_{(-1,1)}(\rho)\right)$$

and

$$\tilde{l}_z(\sigma^2, \rho) = \log\left\{\phi\left(\frac{y_1\sqrt{1-\rho^2}}{\sigma}\right)\right\} + \sum_{t=2}^{T}\log\left\{\phi\left(\frac{y_t - \rho y_{t-1}}{\sigma}\right)\right\} +$$

$$+ \log\left\{\phi\left(\frac{z - \rho y_t}{\sigma}\right)\right\} + a\log(b) - \log\left(\Gamma(a)\right) + (a-1)\log(\sigma^2) - b\sigma^2 + \log\left(\frac{1}{2}I_{(-1,1)}(\rho)\right) .$$

Numerical illustration

The dataset is an sample of size 30 generated from AR(1) with $\sigma^2 = 1$ and $\rho = 0.5$. The prior distribution for σ^2 is $Gamma(1, 1)$, which is bounded. So we can implement HOPA method. Table 4.10 contains the main summaries of predictive random variable $Z|Y = y$ calculated from MCMC sample and approximated by HOPA methods and HOPA simulations. The plots in Figure 4.10 show the intermediate and final results of the implementation of HOPA methods and approximations of predictive probability density with $\tilde{f}^*$ and $\tilde{f}^{**}$. Figure 4.11 illustrates the one-step prediction of the stochastic process.

Table 4.10: Predictive summaries of AR(1) stochastic process using different methods.

	$\alpha_{0.05}$	median	$\alpha_{0.95}$	mean	variance	HPD (90%)
MCMC simulation	-2.11	-0.47	1.23	-0.46	1.03	[-2.08 , 1.25]
HOPA1	-2.07	-0.48	1.17	-	-	-
HOPA1 sim.(10^6)	-2.07	-0.48	1.17	-0.47	0.88	[-2.07 , 1.17]
HOPA2	-2.12	-0.47	1.21	-	-	-
HOPA2 sim.(10^6)	-2.12	-0.47	1.21	-0.47	1.03	[-2.14 , 1.19]

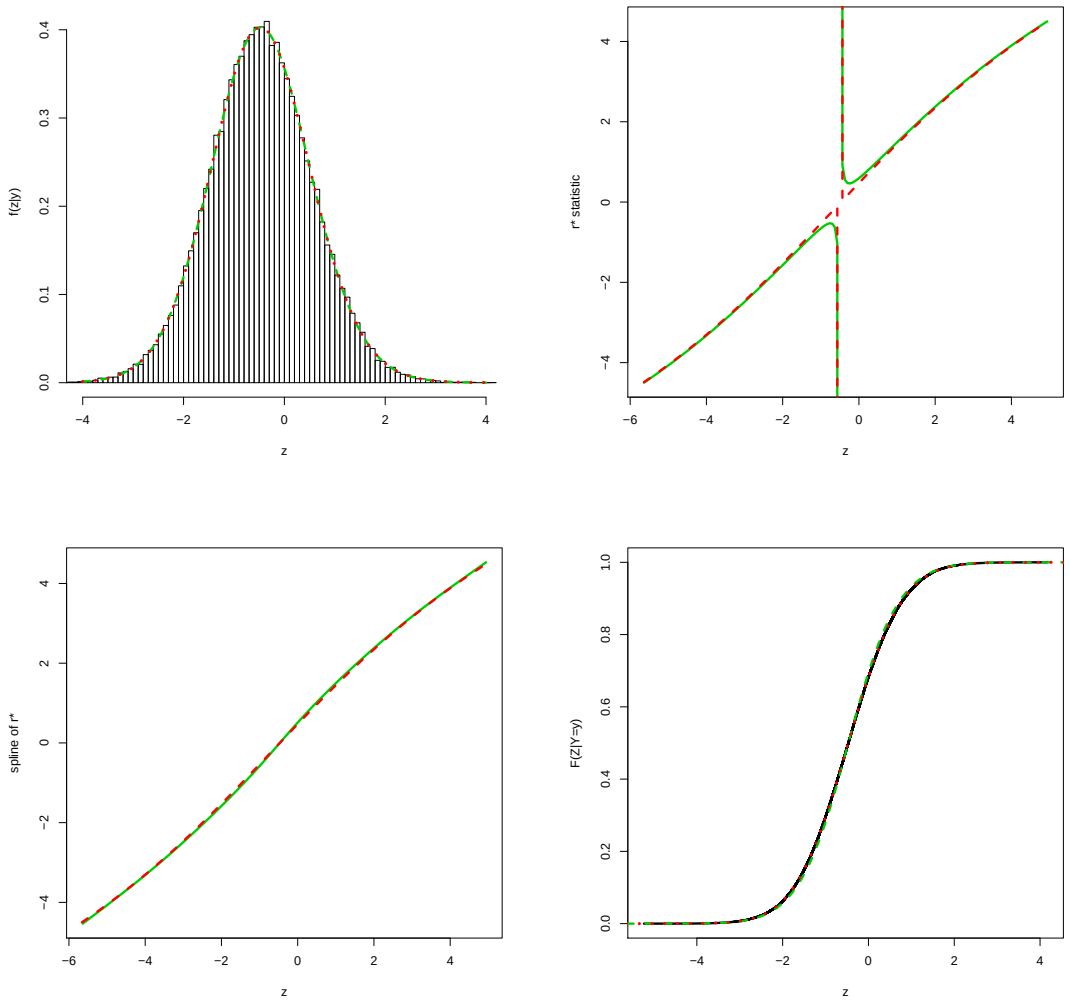

Figure 4.10: Stationary AR(1) stochastic process. *Top-left:* MCMC simulation (histogram), approximation $\tilde{f}^*$ (green-dashed line) and approximation $\tilde{f}^{**}$ (red-dotted line). *Top-right:* $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) . *Bottom-left*: Spline approximations of $r^*(z)$ (green-solid line) and $r^{**}(z)$ (red-dashed line) with $\epsilon^* = 0.2$ and $\epsilon^{**} = 0.1$. *Bottom-right:* The predictive cumulative distribution functions: MCMC simulation (black-solid line), HOPA based on $r^*(z)$ (green-dashed line) and HOPA based on $r^{**}(z)$ (red-dotted line).

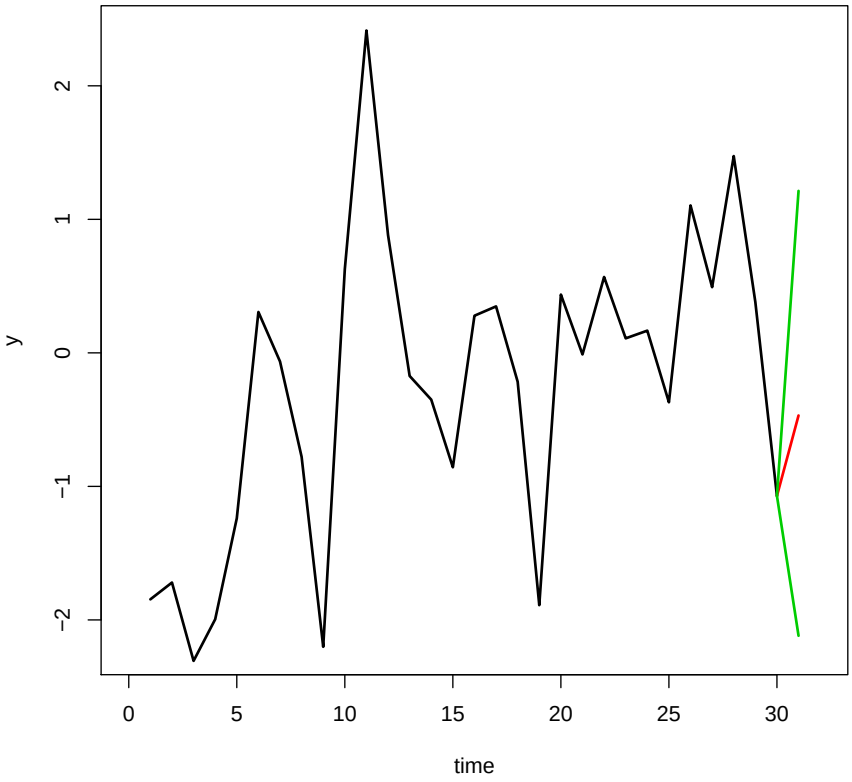

Figure 4.11: Stationary AR(1) stochastic process and its one step prediction: median (red-solid line) , 90% HPD interval (green-solid line).

CONCLUSIONS

As we live in a world of huge information, in constant evolution, quick accurate prediction at a minimal computational cost is a natural desire. In this respect this thesis developed a new method for Bayesian approximate prediction that has the required characteristics. The new method is based on higher-order asymptotics results, i.e. the accuracy of the method depends on the size n of observable data. For this reason we named it Higher-Order Predictive tail Area (HOPA) method. The method can be applied to a wide variety of regular statistical models, with the essential requirement of the posterior mode being unique.

The development of the method is based on the posterior predictive density function, for which we found two formulae that can be seen as ratios of two integrals. The difference between these formulae is the expression of the denominator, i.e. the marginal distribution of the observables. We applied Laplace's method for integrals and obtained formulae that approximate the posterior predictive density with relative error of order $O(n^{-1})$. In practice, the procedures have two ingredients: the log-posterior function based only on the observables, and based on the joint vector of the observables and unobservables; these log-posteriors are supposed to be unimodal smooth functions. The only requirement for the implementation of the approximations of the posterior predictive distribution is the computation of the mode of the log-posterior, which need to be an inner point of the sample space, and its second derivative.

We have found two different approximations, where one of them relies entirely on quantities derived from the log-posterior of the observables and unobservables. Using these expressions we have implemented a third-order

approximation to the tail area, with the assumption that the unobservable vector is univariate. The results are two distinct HOPA methods that approximate the univariate predictive cumulative distribution function and its inverse, i.e. the quantiles. The methods use the standard normal cumulative distribution function and quantities derived from input functions. From these approximations we also implemented the HOPA simulation scheme, which relies on the simple inverse transformation of a standard normal sample. The generated sample from the predictive random variable allow us to compute other quantities of interest, such as predictive moments.

A necessary condition for the implemention of the method is the existence of a unique mode for the input function, which has to be an inner point of the domain space set for the function under consideration. This condition holds when we have a regular model with a bounded prior distribution for the parameters, and the conditional probability distribution function of the unobservable is a unimodal function, with the mode an inner point of the support. If the prior distribution of the parameters is unbounded, then we can overcome the problem by means of reparametrizations. Even in simple models, it may happen that the distribution for the unobservable given observed data and parameters has the mode on the boundary of the support. For such cases a simple solution was founded, which consists of a one-to-one trasformation of the initial predictive random variable.

From a practical point of view the construction of the approximation of the posterior predictive cumulative distribution, requires the computation of all needed functions for every point of the support. However, typically, these functions have a numerical discontinuity in mode, and for this reason we used numerical spline interpolation. The technique may exclude values in a neighborhood of the mode, but this do not affect the approximation when the function has a regular discontinuity.

In the examples of Chapter 4 we have implemented both HOPA methods. The comparisons were done graphically and by means of summary statistics, such as 5% and 95% quantiles, the median, the mean, the variance, and 90% HPD intervals. We note that even if the differences between the approximations fo ther density are insignificant, the results of two HOPA methods can

be quite different. The second method provided better approximations for the considered quantities in all examples, because the numerical irregularity around the mode for this method is more local. This version of the HOPA method is highly accurate in many situations even for limited sample sizes.

Compared to standard MCMC methods, the HOPA simulation scheme has the advantage of giving independent samples at a negligible computational cost. Moreover, it doesn't need a proposal distribution, nor any check of convergence. Finally, we also note that MCMC simulation may have poor tail behavior, especially when the number of parameters is large.

In conclusion we can say that the best performing version of the HOPA method could be useful in practice, possibly even in conjunction with other Bayesian procedures. The method could be easily applicable to prior sensitivity analyses comparing different posterior distribution under the same Monte Carlo variation. A limitation of the method is the fact that it only works for univariate predictive distributions. An extension to the multivariate case is certainly worth further research efforts.

Bibliography

Aitchison, J. and Dunsmore, I. R. (1975) *Statistical prediction analysis*. Cambridge: Cambridge University Press.

Azzalini, A. (2001) *Inferenza statistica: una presentazione basata sul concetto di verosimiglianza*. Milano: Springer.

Barnard, G. A. and Sprott, D. A. (1983) Likelihood. *Encyclopedia of Statistical Sciences* **4**, 639–644.

Barndorff-Nielsen, O. E. and Cox, O. E. (1996) Prediction and asymptotics. *Bernoulli* **2**, 319–340.

Barnett, V. (1999) *Comparative statistical inference*. Chichester: Wiley.

Berger, J. O. and Wolpert, R. L. (1988) *The Likelihood Principle*. Second edition, volume 6 of *Lecture Notes — Monograph Series*. Hayward, California: Institute of Mathematical Statistics.

Bjornstad, J. F. (1990) Predictive likelihood: a review. *Statistical Science* **5**, 242–265.

Box, G. E. P. and Tiao, G. C. (1992) *Bayesian inference in statistical analysis*. New York: Wiley.

Brazzale, A. R., Davison, A. C. and Reid, N. (2007) *Applied Asymptotics*. Cambridge: Cambridge University Press.

Bruijn, N. G. D. (1961) *Asymptotic methods in analysis*. Second edition. Amsterdam: North-Holland.

Butler, R. W. (1986) Predictive likelihood inference with applications (with discussion). *Journal of the Royal Statistical Society* **48**, 1–38.

Congdon, P. (2001) *Bayesian statistical modelling*. Chichester: Wiley.

Corcuera, J. M. and Giummolè, F. (1999) On the relationship between α-connections and the asymptotic properties of predictive distributions. *Bernoulli* **5**, 163–176.

Davison, A. C. (1986) Approximate predictive likelihood. *Biometrika* **73**, 323–332.

Davison, A. C. (2003) *Statistical models*. Cambridge: Cambridge University press.

Geisser, S. (1993) *Predictive Inference: An Introduction*. London: Chapman and Hall.

Gilbert, P. and Varadhan, R. (2012) *numDeriv: Accurate Numerical Derivatives*. URL http://CRAN.R-project.org/package=numDeriv: R package version 2012.9-1.

Guttman, I. (1970) *Statistical Tolerance Regions: Classical and Bayesian*. London: Griffin.

Harris, I. R. (1995) Predictive fit for natural exponential families. *Biometrika* **76**, 675–684.

Hills, M. (1998) Likelihood. *Encyclopedia of Biostatistics* **3**, 2239–2244.

Iversen, G. R. (1984) *Bayesian Statistsical Inference*. Beverly Hills: Sage Publications.

Jacod, J. and Protter, P. (2003) *Probability Essentials*. Berlin: Springer.

Kass, R. E., Tierney, L. and Kadane, J. B. (1990) The validity of posterior expansions based on laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, pp. 473–487.

Komaki, F. (1996) On asymptotic properties of predictive distributions. *Biometrika* **83**, 299–313.

Norris, J. R. (1997) *Markov chains.* Cambridge: Cambridge University.

O'Hagan, A. (1994) *Kendall's Advanced Theory of Statistic 2B: Bayesian Inference.* London: Edward Arnold.

Pace, L. and Salvan, A. (1997) *Principles of statistical inference from a Neo-Fisherian perspective.* Singapore: World scientific.

Ripley, B. D. (1987) *Stochastic Simulation.* New York: Wiley.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods.* Second edition. New York: Springer.

Robert, C. P. and Casella, G. (2010) *Introducing Monte Carlo methods with R.* New York: Springer.

Roberts, G., Gelman, A. and Gilks, W. (1997) Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability* **7**, 110–120.

Ruli, E., Sartori, N. and Ventura, L. (2013) Marginal posterior simulation via higher-order tail area approximations. *Bayesian Analysis* p. to appear.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *Journalof the American Statistical Association* **81**, 82–86.

Venn, J. (1886) *The logic of Chance.* London: MacMillan.

Vidoni, P. (1995) A simple predictive density based on the p^*-formula. *Biometrika* **82**, 855–863.

Vidoni, P. (1998) A note on modified estimative prediction limits and distributions. *Biometrika* **85**, 949–953.