**Cairo University**
**Faculty of Economics and Political Science**
**Department of Statistics**

# *On the Selection of Prior Distributions in Bayesian Analysis*

**A Thesis Submitted to the Faculty of Economics and Political Science in Partial Fulfillment of the Requirements for the Degree of Master in Statistics**

Presented by
**Niveen Ibrahim Aly El-Zayat**

Supervisors

**Dr. Laila O. El-Zeini**

Associate Professor
Department of Statistics
Faculty of Economics and Political Science

**Dr. Emad E. A. Soliman**

Assistant Professor
Department of Statistics
Faculty of Economics and Political Science

**Cairo (2007)**

**Name:** Niveen Ibrahim Aly El Zayat

**Nationality:** Egyptian

**Date and Place of Birth:** 3/9/1973 – Agouza – El Giza

**Degree:** M.Sc. in Statistics

**Specialization:** Statistics

**Supervisors:**

Assoc. Prof. Laila O. El-Zeini Assis. Prof. Emad E. A. Soliman

Department of Statistics

Faculty of Economics and Political Science

**Title of Thesis:** On the Selection of Prior Distributions in Bayesian Analysis

**Summary:** The main objective of the current thesis is to review the best known approaches of selecting the prior distributions of unknown parameters in Bayesian analysis. Two main approaches are available in the literature, namely the noninformative prior approach and the informative prior approach. The thesis throws light on the well-known methodologies of each type as introduced in the literature. Moreover, the study considers the definition, motivations, philosophy and derivation of each type. Furthermore, the thesis discusses their merits and drawbacks. The affinities and distinctions of different types are revealed as well. Applications of the prior distributions of both approaches are demonstrated to get the posterior analysis of some well-known models in econometrics and time series analysis; such as general linear model and autoregressive model of order one. A numerical example is introduced, based on simulation studies for AR(1) model, to compare the performance of the studied priors using some criteria. The results of the comparative study suggest that there is no clear-cut prior distribution recommended for usage where elicitation of suitable prior distribution is based on the time series length and properties. Finally, all priors are demonstrated by some real life time series data sets to illustrate the behavior of the candidate priors.

# Abstract

The main objective of the current thesis is to review the best known approaches of selecting the prior distributions of unknown parameters in Bayesian analysis. Two main approaches are available in the literature, namely the noninformative prior approach and the informative prior approach. The thesis throws light on the well-known methodologies of each type as introduced in the literature. Moreover, the study considers the definition, motivations, philosophy and derivation of each type. Furthermore, the thesis discusses their merits and drawbacks. The affinities and distinctions of different types are revealed as well. Applications of the prior distributions of both approaches are demonstrated to get the posterior analysis of some well-known models in econometrics and time series analysis; such as general linear model and autoregressive model of order one. A numerical example is introduced, based on simulation studies for AR(1) model, to compare the performance of the studied priors using some criteria. The results of the comparative study suggest that there is no clear-cut prior distribution recommended for usage where elicitation of suitable prior distribution is based on the time series length and properties. Finally, all priors are demonstrated by some real life time series data sets to illustrate the behavior of the candidate priors.

**Key Words:** Bayesian analysis - Prior distribution - Posterior distribution - Noninformative prior distributions - Jeffreys' prior – Invariance - Locally uniform prior - Data translated likelihood - Maximal data information prior - Informative prior distributions - Natural conjugate prior - G-prior - General linear model - Autoregressive models of order one.

**Supervisors:**

Assoc. Prof. Laila O. El-Zeini  Assis. Prof. Emad E. A. Soliman

Department of Statistics

Faculty of Economics and Political Science

# Abstract

Prior selection is considered as a crucial difficulty that have ever encountered Bayesian framework for many applications, since prior specification is the prime step to perform a Bayesian analysis to the unknown parameters for the decision making. Bayesian machine updates the prior information available about parameters through the prior distribution in the light of information provided by the likelihood function to get finally the so-called posterior distribution. This last distribution contains all possible information about parameters. Thus, it is used for making inference about the parameters.

That essential rule of prior selection in the structure of Bayesian analysis explains the vast literature in prior selection problem. This selection can be done using one of two main approaches, noninformative prior and informative prior according to the existence of prior information about the parameters in the model of interest.

Noninformative prior approaches are used when no or few information are available about parameters. These approaches are widely accepted in literature since they do not require subjective determinations. Besides, they introduce automatic consecutive steps to derive the posterior results. One of the most well-known noninformative prior is the Jeffreys' prior. Such prior has gained widespread acceptance in many fields due to its simplicity. One of its main features is the invariance property. However, Jeffreys' prior can not be applied in some cases when there are different types of parameters or when no regularity conditions are available. That motivates authors to develop some other noninformative prior distributions. These approaches differ in their philosophies. One of those outstanding approaches is the locally uniform prior proposed by Box and Tiao (1973) that is based on the concept of the data translated likelihood. Another one developed by Zellner (1977), is the maximal data information prior that is based on maximizing the data provided by the sample. That last approach requires developing some informational criteria.

On the other hand, the informative prior distributions are used when information are available about the unknown parameters. Many approaches were developed in literature to quantify such information in a form of probability distribution. The progress in computing facilities motivates authors to develop more accepted informative prior distributions. One of the most famous informative prior approaches is the natural conjugate prior developed by Raiffa and Schlaifer (1961). This prior is chosen such that it has the same functional form as the likelihood function when the last is expressed as a function of the parameters. The only difficulty encounters that type is the specification of hyperparameters. However, there are many methods developed in the literature to solve such a problem. Another type of informative priors is the g-prior introduced by Zellner (1986) to formulate the Bayesian analysis of the general linear model. This type of informative prior is a special case of the natural conjugate one but with less effort required to assess the hyperparameters, since it only requires estimating the location hyperparameters of the coefficient parameters while the variance-covariance matrix is estimated using the design matrix.

Since the study aims to investigate the different types of noninformative and informative prior distributions, a complete perspective over both approaches is displayed and applications to these approaches have been introduced to produce the posterior analysis of the general linear model (GLM) and the AR(1) models. A comparative study is also demonstrated through simulation devices to investigate the efficiency of the different prior approaches to produce the posterior analysis of AR(1) models. All priors are demonstrated by some real life time series data sets to compare the performance of the candidate priors.

From all what have been introduced in the thesis, the current thesis emphasizes the great importance of the prior selection according to the characteristics of the model of interest and to the sample size as well. Thus, caution must be given to the different situations that may be encountered and it is recommended to examine the appropriate prior since no clear-cut method tells the investigator which is the best prior to be used.

*Dedicated To my Mother and the Soul of my Grandfather and Grandmother*

# *Acknowledgment*

*Foremost, grateful acknowledgment is made to Dr. Samir Shaarawy for his kindhearted support, valuable and favorable guidance he has given throughout this dissertation. I would like to record that the point of this research is driven back to his marvelous suggestion. I am greatly indebted to him for guiding me to this mesmerizing research area.*

*I would deeply like to thank Dr. Laila El Zeiny for her moral and practical support. Dr. Laila continuous encouragement during the preparation of this dissertation made it possible for me to complete this work.*

*I would like to thank Dr. Ahmed Daif for undertaking the supervision of this work in its early stage.*

*I would also like to express my indebted gratitude to Dr. Emad Soliman for his insightful supervision and considerable effort. His valuable guidance and powerful support were tremendously helpful for the progress of this dissertation.*

*A great appreciation goes to Dr. Mohamed Ismail for his kind assistance in providing me with a lot of reference to complete such research. Special gratefulness should go to the Time Series Group in the Faculty of Economics and Political Science, for their significant advices.*

*I would also like to thank all staff in the Department of Statistics of the Faculty of Economics and Political Science, for their nonstop support.*

*Last but not least, I should add thanks to my mother for her constant support and prayers. My greatest debt is to my nice nephew Ahmed for his positive effect on my progress.*

# Contents

# Contents

# Contents

# Chapter 1

# Introduction

In Bayesian analysis for a parametric statistics problem, it will be inevitable to specify a prior for the unknown parameters. Bayesian analysis makes inference about parameters by assessing a prior distribution and then deriving the posterior distribution via Bayes' theorem. Thus, the Bayesian framework allows one to incorporate prior information into statistical models for decision-making. This prior information is combined with information from the data using the axioms of probability to yield posterior distribution for parameters of interest. This is done using the Bayes' rule which says the posterior is proportional to the likelihood times the prior (Hahn, 2006).

Accordingly, due to the crucial rule of prior selection in Bayesian analysis, various approaches have been developed in literature to assess prior distribution. A prior distribution which is "automatically" specified by the given model, is called a ***noninformative prior*** since no other entries are required to derive such a prior. Alternative names are given to these kinds of priors such as "default", "vague", "reference", "ignorance", "weak", "inner", "invariant", "objective", "flat ", or "diffuse" priors (Ye, 1990). Such approach of determining priors is termed as "objective" and has long been attractive in practice since it involves numerous methodological advantages (Yang, 1994).

However, subjective determination of the prior density has been the foremost philosophical foundation for Bayesian analysis, though it is often criticized. That sort of determination are named as ***informative priors*** since one has a certain "degree of beliefs" and the Bayesian algorithm is followed to study how the data change these beliefs. Probability theories have been developed in the literature to measure these beliefs numerically (Lindley, 1965a and 1965b).

In this chapter, a brief discussion to the perspective and motivation for the prior information are presented, followed by a discussion of difficulties in prior selection.

Then, the most recent approaches to develop prior distributions will be briefly reviewed. Finally, the objectives and the structure of the thesis will be sketched.

## 1.1. Perspective on Prior Information

In many practical applications, the decision maker usually has additional information about the parameters of interest more than those found in current or observed data. For example, a manager may know perfectly that another competitor's factory has burnt down. What should he do? Does he ignore the information or try to make use of it? A trading company knows well that a new legislation will appear in the short coming period (approximately 15% increasing in the tax rate). Will the company make changes in its production activity, such as increasing or decreasing production or ignore the information? Such knowledge is a further form of relevant information that would be desired to combine with the observed data to make a more refined estimation of the parameter of interest (Barnett, 1973). Information of such sort is derived from outside the current situation and termed as a "priori" or "prior information". Prior information is generally of various types, usually one of the following sources or a mixture of them,

1. Information of previous data and studies.
2. Theoretical information.
3. Casual observation.

Methods have been set up to quantify a priori. Generally, the expected effect of such measures is probabilistic. A probability distribution for that expected effect is decidedly required to characterize its uncertainty. A powerful tool to do this task is the so-called "prior distribution". The reader may refer to (Barnett 1973) and (Berger, 1985) for an inclusive discussion to variety of methods of probabilistic determination of prior information. This probability distribution is used to represent the degree of reasonable belief that may be available about the parameter and is always conditional on our state of information. Consequently, this probability distribution is revisable against variation in such state of knowledge. Furthermore, this process of revising probability associated with the priori in the face of new information is the essence of learning form experiment. Incorporating new information made by the use of Bayes'

theorem that is considered as an essential part of Bayesian approach. It is known in literature as "principle of inverse probability", since information from data are used to infer what random process generates them (Zellner, 1971).

A fundamental feature of Bayesian analysis is the use of prior information as well as the observed data in the final analysis. Bayesian mechanism combines information from sample through “likelihood function” with the prior information through “prior distribution” to get the so-called “posterior distribution”, according to Bayes’ rule, that is,

$$\textit{Posterior distribution} \ \alpha \ \textit{Prior distribution} \times \textit{Likelihood function}$$

In this prospect, prior distribution embodies the probability density function based on our initial belief about the parameter. Whereas the posterior density function incorporates our initial information as represented by the prior distribution and our sample information as represented by the likelihood function. Zellner (1971) declared some remarkable characteristics of the posterior distribution are:

1. As the sample information grows, it will more dominate the posterior distribution which will become more concentrated about the true value of the parameter.
2. The posterior distribution will be the same in the case when there are different prior distributions as long as they are combined with common large sample information.

However, Bayesian results could be sensitive to different assumptions on the prior distribution. This is studied in literature under the so called "sensitivity analysis" or "Bayesian robustness".

Nevertheless, in the Bayesian approach, the prior information about parameters of a given model is represented by a chosen probability density function (p.d.f.). That distribution must be adequate in representing prior or initial information about parameters otherwise another prior p.d.f. have to be chosen to represent adequately the prior information. This fact guides us to a very crucial question "How one could be able to assign a prior p.d.f. to represent a state of knowledge about the parameter of the given model? Are the information about parameters always available? These questions will be replied through the following section.

## 1.2. Difficulties of Prior Selection

Bayesian analysis of a statistical problem consists of three stages, namely the prior, posterior and predictive distributions. The prior distributions reflect the expert's beliefs about the parameters, and the posterior distribution is considered as a modification of the prior information in the light of the observed sample. Therefore, the posterior distribution construction is affected by the selection of the prior distributions. Hence, careful specification of the prior distribution is of great importance, since using bad prior will lead to bad posterior results.

Choosing the prior distribution is considered the hardest part in applying the Bayesian framework. Prior selection faces two main problems.

- How to express the case of "knowing nothing" or "knowing little" about the parameters in a probability distribution representation. (For more details about this problem, one may refer to Zellner (1971), Box and Tiao (1973), Berger (1985), and Ye (1990)).
- How to express the information about the parameters, if exist, in a satisfying probability distribution representation. (For more details see Berger (1985)).

Many essays are developed in the Bayesian literature to discuss the above problems and introduce various solutions to overcome them.

As a solution for the first problem, one may use the so-called noninformative prior. That is termed as "weak informative prior" as well since few information is the merely available about parameters. Moreover, noninformative priors are described as "objective" because prior elicitation does not require assigning any personal or subjective consideration. A noticeable remark is that Bayesian statistics are termed as "objective", due to the use of noninformative prior distributions. The outstanding motivation for noninformative priors is that they are considered as a remedy of the often disapproval of "subjectivity" that most Bayesians rely on when quantifying prior distribution through personal judgments. Thus, noninformative priors achieve a conventional agreement as it retains the prevailing preconception that the science must be objective. Moreover, noninformative prior distributions are more practical since they have no population basis and play a minimal role in the posterior distribution. The

idea behind the use of noninformative prior distributions is to make inferences that are not greatly affected by external information or when external information is not available.

The variety of the approaches to develop noninformative prior distributions in Bayesian analysis is vast and complex. Many philosophies are available in the literature to choose a noninformative prior. The most famous types are Jeffreys' prior, locally uniform prior, maximal data information prior and reference prior.

To solve the second problem, one may use the so-called informative prior. Informative prior distributions are used when there is information, usually subjective, about the parameters available before assessing the data. Ignorance of this information, just for the sake of objectivity, is not recommended. Subjective beliefs are usually available in scientific inference. For example, a scientist decides to do a particular experiment in order to confirm some hypothesis about the parameter, see Press (1989). A probability distribution is needed to represent these subjective beliefs.

On the other hand, informative priors have a stronger influence on the posterior distribution. The influence of the prior distribution on the posterior is related to the sample size of the data and the form of the prior. Generally speaking, large sample sizes are required to modify strong priors, where weak priors are overwhelmed by even relatively small sample sizes.

Informative priors are typically obtained from past data and are commonly used in small samples where there is insufficient data to form a convenient conclusion.

Nonetheless, in developing strategies for specifying informative priors, researchers have recognized the importance of carefully eliciting an expert's judgments so that the translation from belief to a probability distribution is as accurate as possible. As a result, a wide variety of procedures for eliciting informative priors have been developed (Hahn, 2006). A brief review of informative prior approaches provides an indication of the extent in this area of research, as represented in a later chapter. Representing the prior information by a proper distribution has been widely covered in statistical literature. Such procedures are available to select informative prior such as conjugate prior, g-prior, predictive density approach and ML-II prior.

Approaches that are widely followed in literature to select both noninformative and informative priors will be discussed in details in subsequent chapters.

## 1.3. Objectives and structure of the thesis

The main objective of the thesis is reviewing the best-known approaches of selecting informative and noninformative prior distributions. The philosophy, derivation, and properties of each type will be studies and demonstrated by some theoretical examples and by real and simulated time series data sets as well.

**In more details, the objectives of the current study can be summarized as follows:**

1. The study reviews the best-known approaches of selecting noninformative prior distributions, such as Jeffreys' prior, locally uniform prior and maximal data information prior.
2. The study reviews the best-known approaches of selecting informative prior distribution, such as natural conjugate prior, and g-prior.
3. The philosophy, procedure for derivation and properties of each type are explained. In addition, the difficulties in the construction of each prior are discussed. Moreover, relations between priors are verified. Finally, some selected examples are devoted to illustrate the derivation techniques of each type.
4. Posterior analysis of the general linear model (GLM) is established using informative priors; natural conjugate prior and g-prior.
5. Posterior analysis of the well known time series model, the autoregressive model of order one (AR(1)), is demonstrated using the noninformative and informative prior distributions.
6. Numerical examples are introduced, based on simulation studies for AR(1) model, to compare the performance of the studied priors using some criteria. Comparative study is implemented concerning the general case of AR(1) process when the stationarity assumption is ignored.
7. All priors are demonstrated by some real time series data sets to illustrate the behavior of the candidate priors.

The thesis is structured as follows:

In chapter 2, a comprehensive discussion of noninformative priors is considered involving definitions, motivations and importance of noninformative priors. Various approaches to develop noninformative priors are introduced. Particular attention is given to certain noninformative priors such as Jeffreys' prior, locally uniform prior and maximal data information prior.

Chapter 3 throws light on the methodologies of informative priors' elicitation as introduced in the literature. Types of prior information that may be available about parameters of a given model are reviewed as well. A specific interest is focused on the natural conjugate prior and the g-prior.

Chapter 4 shows the application of the natural conjugate prior and the g-prior to the well known general linear model (GLM). These informative priors are used to compare the posterior analysis of the GLM resulted from each prior.

Chapter 5 applies some of the preceding noninformative and informative priors to the well known time series model, AR(1). Attention is restricted to the posterior analysis of AR(1) using different priors concerning the general case of the process when stationarity is not assumed. Moreover, a comparative study has been carried out based on simulation to compare the efficiency of the studied priors. Some efficiency criteria are provided to serve the comparative study. Finally, the posterior analysis of some real time series data sets, that follows AR(1), is done. Most of the discussed prior distributions are applied and the posterior analysis is produced using the candidate priors.

Finally, the main results of the current work are summed-up in a concluding chapter (chapter 6). Moreover, some points for future work are presented.

The simulation study and the computations concerning the posterior analysis of the real examples are carried out using Matlab software (version 7.1). The scripts to do such task are exhibited through Appendices II and III.

# *Chapter 2*

# *Noninformative Prior Distributions*

## 2.1. Definitions and motivations

A prior that is constructed by some formal rules or subsequent algorithms and that is specified automatically by the given model is called ‘noninformative prior’. A researcher does not need any other inputs to derive such prior (Ye, 1990).

Noninformative priors are mainly used when the information about the parameters, to be provided by the prior, is little with respect to that from the data. The literature contains many alternative names for this type of priors such as "objective priors", "vague priors", "diffuse priors", "reference priors", and "invariant priors", "default priors", "ignorance priors", "weak informative priors", "inner priors" or finally "flat priors".

Since the use of noninformative priors has been considered as a routine in Bayesian practice, it would be helpful to review the numerous motivations to noninformative priors and the reasons of their importance to Bayesian analysis. These motivations are briefly summarized by Berger and Yang (1996) as follows:

1. Utilizing noninformative priors avoids difficulties and criticisms that face Bayesian analysis when using subjective prior distributions and being away from using of objective inference. Also in the case of large amount of data, there is no need to do more effort by using subjective prior distributions. Therefore, Bayesian analysis with noninformative priors is the most preferred objective inference that is possible.
2. Moreover, elicitation of subjective prior distribution is difficult, because of cost or time constraint. On the other hand, in particular circumstances, a simple and fast approximation is always needed regarding the complexity and high dimensionality of various modern Bayesian models, such as, Bayesian time series models.

Automatic or default prior distributions are then needed since they provide good approximation at much less effort than a full Bayesian analysis.

3. In high dimensional problems, subjective prior elicitations are desirable for "interesting" parameters whereas noninformative priors can be given to the unimportant or "nuisance" parameters. Therefore, in multiparameter statistical problems, using a noninformative prior may be the best method for diminishing nuisance parameters (Ye, 1990).
4. Subjective determination to the prior information may easily result in "poor" prior distributions because of the inherent elicitation bias, where that elicitation typically yields only a few feature of the specified prior (such as its functional form) in addition to some other characteristics that are convenient but inappropriate. Therefore, it is important to compare outputs from a subjective analysis with those from noninformative prior analysis. It is important to check that the expected substantial differences are due to the features of the prior that are trusted.
5. In addition, noninformative priors could be considered as a starting point for investigating the effect of any other suggested subjective priors by comparing the Bayesian analysis using these two approaches.
6. Another motivation to noninformative priors is due to their simplicity in the Bayesian analysis particularly in Bayesian time series analysis. That is because of the difficulties that face posterior computations using other subjective prior distributions. Furthermore, applying such priors in Bayesian time series analysis does not face any problems in dealing with usual presence of constrains on the parameters in time series models, such as stationarity and invertiblity (Ismail, 1994).

On the other hand, there are some difficulties and problems in selecting such priors. The main difficulty is that there is no clear-cut method for saying which noninformative prior should be used. Besides most of noninformative priors are improper, which makes interpretation about posterior results unclear (Berger, 1985).

## 2.2. Literature Review

There has been a tremendous amount in the statistical literature of noninformative priors. Procedures for deriving such priors vary according to the philosophy of each type. Furthermore, several books and articles have been concerned with discussing or comparing different approaches in developing noninformative priors, see (Kass and Wasserman, 1996) and (Berger and Yang, 1996). The last reference is considered as a catalog of most of noninformative priors that have been developed.

The work in developing noninformative priors has begun so early by Bayes (1763), which is known as Bayes' Postulate, and Laplace (1812). They developed a noninformative prior to represent the state of complete ignorance or knowing nothing about the parameters. They depended on the principle of insufficient reason to evolve such prior that is if there is no reason to prefer one value of the parameter to any other then all values should be taken equally likely. Hence, they used the uniform prior as a noinformative prior. Such prior is improper, in the case of infinite parameter space, and is not parameter invariant.

Jeffryes (1961) tried to overcome the lack of invariance to transformations through developing what is the most famous known as Jeffreys' prior. The real contribution due to Jeffreys' work is that his prior is advocated by convention (or international agreement), see Kass and Wesserman (1996). He did not insist on unique representation of ignorance, but he worked to derive the best rule in each of many cases, as it will be shown in next sections. The Jeffreys' prior has gained widespread acceptance on many fields especially Bayesian time series analysis. This wide use is due also to its simplicity to derive so it is considered as a standard noninformative prior in Bayesian time series analysis.

The principal of choosing noninformative priors based on invariance property is widely discussed in the literature beside Jeffreys' prior such as Hartigan (1964) and a recent work for Datta and Ghosh (1996).

Some other modification to Jeffreys' prior is presented according to a different philosophy. Box and Tiao (1973) have derived the locally uniform prior as a

noninformative prior based on the concept of data translated likelihood. They developed a noinformative prior that makes the likelihood independent of the data except for its location. This prior is proposed because of the difficulties that face the use of Jeffreys' prior in dealing with multiparameter case with different types of parameters. Recent work in deriving that type of priors is introduced by Kass (1990).

Maximum entropy is another approach to construct a noninformative prior, where prior with larger entropy is considered as being less informative. This principle seeks the prior that maximizes the Shannon (1948) entropy. Such priors were developed by many authors such as Jaynes (1957, 1968, 1980, 1982, 1983) and Zellner (1991, 1995).

Another criterion for selecting a noninformative prior is that based on the information measurement, the most important studies to derive the prior are due to Zellner (1971, 1977) and Berger and Bernardo (1989). The Zellner's method leads to maximal data information prior, which gives the minimum information compared with the sample information. While Berger and Bernardo (1989, 1992) introduced the most formal rule to derive a noninformative prior that is called reference prior. A very recent work discusses definition and application of the reference prior is due to the work of Berger *et al* (2007). The last prior is often used in the case when there are nuisance parameters, where the Jeffreys' prior does not adopt such case. Many authors have extended the implication of reference priors to multiparameter case through different applications such as Ye (1990) and Yang (1994).

Frequentist coverage matching approach is another method to select the noinformative prior that makes "the data speak for themselves". Such prior is the one that achieves probability agreement between the sample and the posterior distributions. This approach has been widely undertaken to discriminate among alternative candidate prior distributions such as Welch and Peers (1963), Peers (1965 and 1968), Ye (1990) and Yang (1994), Datta and Ghosh (1995), Sun and Ye (1995).

For more review and discussion to a variety of criterion to select noninformative priors reader may be refered to Kass and Wasserman (1996).

## 2.3. Jeffreys' Prior

### 2.3.1. Introduction

Tracing Jeffreys' work throws light on opulent literature with vast writings of other eminent statisticians who undertake exploring Jeffreys' contributions to Bayesian statistics. They have appreciated Jeffreys' work due to its particular influence on their own work in statistics. The influence of Jeffreys' work in the analysis of several statistical problems reflects the power of Jeffreys' contributions and insights. Some of those outstanding writings are for Geisser (1980), Good (1980), Lindley (1980), Kass (1982) and Zellner (1980, 1982a and 1982b).

The current thesis will mainly rely on those writings to present a short summary of Jeffreys' numerous contributions to Bayesian analysis. However the main emphasize will be given to the noninformative prior suggested by Jeffreys, the so-called Jeffreys' prior. Thus, it should be emphasized that the work will highlight these contributions bearing in mind that the prime interest is writing up the Jeffreys' entry to the prior distributions when no or little information is available within the Bayesian work.

Some of Jeffreys' contributions to the philosophy, methodology, and applications of Bayesian analysis could be presented through the following headings:

***Jeffreys as a scientist besides being a statistician***:

Jeffreys was a noted physical scientist who re-established the statistical theory in his time on the Bayesian foundations. Therefore most of his work on Bayesian statistics was oriented towards the natural sciences. In this regard, one can never ignore the very important citation of Lindley (1980, p. 4),

> It is of course, one of Jeffreys' great strength as a statistician that he is a scientist.

This feature produces the inherent appearance of mixing those theories and applications found in Jeffreys' work. This also reflects a testimony of the coherent statistics apparent in Jeffreys' work, which was built on the theoretical satisfaction and practical implementation. In this respect, Zellner (1980, p. 4) comments:

> This is a recognition of the pervasive interaction between theory and application that is present in his work.

It is noteworthy that his procedure for estimation, prediction, and inference is applicable in natural and social sciences. So the applications of his techniques in astronomical and geophysical fields are similar to that in econometrics and other social sciences (Zellner, 1980).

***Jeffreys' contributions to probability:***

Most of Jeffreys' contributions to statistics, particularly to Bayesian statistics, are found in probability. Jeffreys was the first who used probability to deal with problems in the philosophy of science in addition to using probability to explain and investigate the reasonability of scientific theories. This work was introduced through his work with Wrinch (1919, 1921 and 1923), and through his famous book *Scientific Inference* (1931). Furthermore, Jeffreys (1939) extended the notion of "degree-of-belief", which was first used by those who adapted the subjective concept of probability. However Jeffreys disagreed with them in their confining on the personal beliefs, so he treated probability in the logical sense. Jeffreys' theory of probability book has been introduced in two more editions, in 1948 and in 1961, but the third in 1961 was of the Bayesian revival. Jeffreys (1961) introduced the logical concept of probability in Bayesian framework based on the principle of inverse probability, to compute probabilities rather than the empirical calculations which followed in the frequentist approach. In other words, Jeffreys' work in probability is developed along Bayesian lines.

Jeffreys defines probability to be the reasonable "degree of belief", or "priori", that an individual has in a proposition "q" given some body of evidence (the observed data) "p". Then, the formal notion p(q | p) expresses the measure of the implication in which "p" support or refutes "q". This concept of probability is considered to be objective through the Jeffreys' view "*there is one and only one opinion "q" justified by any body of evidence "p" which could be the same*". Thus, such probability p(q | p) is considered to be unique impersonal logic one could calculate or estimate only in the context of Bayesian framework (Zellner, 1982b). Then the logical view adapted by Jeffreys straddled the subjective and frequentist view in being objective but expressing degree of belief (Barnett, 1973).

However, Jeffreys does not assume that everyone always have the same prior information (Zellner, 1982b). Moreover, he is the first to adopt the case when the person has no opinion which is the case of formulating “ignorance”, “nothing” or “knowing little” resulting in the so called *Jeffreys' prior*, which will be discussed in the next point. Such type of prior is noninformative. For that last reason, the probability theory book of Jeffreys' is considered as a modern book because it introduces a recent meaning of probability when little information is available (Lindley, 1980). That concept of probability could be updated in the light of new information using the Bayes’ theorem or the principle of inverse probability, where the resulted posterior distribution could be used as a prior distribution taking into account further set of data (Huzurbazar, 1980).

***Jeffreys' ingenuity in quantifying “ignorance”:***

As mentioned above, Jeffreys in developing his theory of probability, has not denied the presence of any type of information (prior information) the investigator may have and need to be tested with data. Such sort of priori is termed as “informative prior”. Jeffreys was aware of many applications in which informative priors rather than noninformative priors should be applied (Zellner, 1980). Furthermore, Jeffreys in his work argued that each scientific law should be assumed to have a priori otherwise no law could ever become probable no matter the evidence includes it (Good, 1980).

On the other hand, it may be the case of lack of information, i.e. the case when the investigator has “no opinion” or “know little” about the proposition. In such case, Jeffreys was considered as the unparalleled statistician that blew up the well-known procedure for formulating “ignorance”, which is translated into the so called *Jeffreys' prior*. However, he had a firm belief in the existence of an “initial” state of knowledge even before data is available, and the importance of being able to make inference merely based on data. Zellner (1982a) describes Jeffreys as pragmatic in his valiant attempts to represent such a state of information, he says as well:

> The situation is similar to the need to formulate the concept of vacuum in physics.

At this point, it is noteworthy to put forward the reasons that motivated Jeffreys to develop procedures to express such state of ignorance; those reasons are as follows:

1. It is a matter of common nature of science to let the data speak for itself. Since in many contexts including; scientific reports, courts of law and many other fields, one wishes to abstract from personal views of an investigator (Zellner, 1980).

2. "A subjective assessor who had some prior information would need to be driven back to the cradle or womb to reveal a time when what he knew was negligible or irrelevant to the matter at hand." (As stated by Geisser, 1980, p. 17).

3. One of the difficulties that faced the Bayesian theory is which prior distribution to be used when the prior knowledge is weak relative to that provided by the data. This difficulty has been considered as a serious block to the universal acceptance of Bayesian approach. Thus, Jeffreys tried to describe that relative lack of information and developed a theory to deal with this difficulty (Lindley, 1980). It is very important here to notice that, this type of prior distribution will set the Bayesian machinery in motion providing indifference or impartial stance (Geisser, 1980).

The most noticeable criticism encountered Jeffreys is the dissatisfaction with his technique to obtain numerical values of such unknown prior probabilities. Jeffreys' respond was "It is not a correct description or an exact quantification, but a type of approximation, to determine some infinite number of initial probabilities, each is consistent, and then choose the best one according to some type of international agreement". Jeffreys (1961) takes on providing satisfying general canonical rules for choosing initial probabilities to express ignorance. On that same matter, Kass and Wasserman (1996, p. 1345) state:

> The real contribution that due to Jeffreys' prior is that he gives an evaluation to the prior distribution base by convention away from unique representation of ignorance. That means that Jeffreys did not insist on unique representation of ignorance, but he did work to find “the best” rule in each of many cases.

Jeffreys also gave a further support of these general rules in case of large amount of data. Where, in such case, the assignment of these initial probabilities by a conventional choice of priors would make little difference in the posterior results. Moreover the results given by the Bayesian inference are indistinguishable from those given by the classical inference (Huzurbazar, 1980).

In view of what have been presented, Jeffreys' prior is considered as an essential part of the furniture of the Bayesian statistics.

### 2.3.2. Derivation

The preceding section discusses the Jeffreys' motivation to express the case when the prior information about the parameter is vague relative to that provided by the observed data and his attempts to seek a general formal rule or a standard prior distribution. Such distribution would be viewed as an approximated representation to a vague prior.

Jeffreys (1961) defined the noninformative prior distribution of the parameter as follows:

> It is a way of saying that the magnitude of the parameter is unknown when none of the possible values need special attention.

Jeffreys stated that if there is no information relevant to the actual value of the parameter then the prior distribution must be chosen to express none or to say nothing about the parameter values. However, it may be restricted within certain constraints. Therefore, Jeffreys (1961) derived some rules for choosing the prior distribution to cover the most common cases of the regular type. He identified rules that should satisfy the following characteristics:

1. Provide a formal way to express ignorance of the parameter value over the permitted range.
2. Make no statement of how frequently that parameter occurs within different ranges.
3. Give the same results in terms of several different sets of parameters, that is, the rules have to be invariant under re-parameterization.

**Jeffreys' first rule:**

If the parameter $\theta$, the mean in location densities for example, may have any value in a finite range or from $-\infty$ to $\infty$, the prior distribution should be taken as uniformly distributed in the form:

$$p(\theta) \propto \text{constant} \tag{2.3.1}$$

The uniform distribution was first used, to express complete ignorance, by Bayes (1763) and later by Laplace (1812). It is based on the "*principle of insufficient reason"* where if there is no reason to prefer one value of the parameter to any other then all values should be taken to be equally likely. The choice of the uniform prior has long been known as *Bayes' postulate* as an indication to Bayes' theorem. Jeffreys indicated that the uniform distribution could not be a final solution for all problems because of its lack of invariance under transformation. In more explicit meaning, the ignorance about the parameter values intuitively implies ignorance about the values of any transformation of the parameter. However, given a certain transformation, the uniform distribution would not be the distribution of such function of the parameter (see Lee, 1989). This concept is usually termed as invariance, which will be explained in a following subsection.

It is obvious that the p.d.f. in (2.3.1) is improper which means that it has infinite mass or the integral on that density over the entire range of the parameter leads to $\infty$ rather than the unity.

Using (2.3.1) involves representing complete ignorance about the parameter values. Jeffreys explained this by the statement $\Pr\{a < \theta < b\} = 0$, where a and b are any finite numbers, however this statement does not mean that $\theta$ is outside the closed interval [a,b], (which resembles the fact that the probability of a continuous random variable taking a specific value equals zero). This property corresponds to the first charachteristic mentioned above.

Based on the previous property, the odds $\Pr\{a < \theta < b\} / \Pr\{c < \theta < d\}$ is indeterminate, where a, b, c and d are any finite numbers. This property corresponds to the second characteristic, since no statement can be made about the odds that $\theta$ lies in any particular pair of intervals. The indeterminacy of this ratio seems to be adequate to justify the use of the rectangular p.d.f. (see Zellner, 1971).

To check whether the distribution in (2.3.1) meets the third characteristic, consider another parameter $\eta = g(\theta)$, say $\eta = \exp(\theta)$ hence, the inverse function given by $\theta = \ln(\eta)$. This is a one-to-one transformation, through which the new parameter $\eta$

will be defined over $(0,\infty)$. Hence if $p(\theta)$ is the density of $\theta$ then the corresponding noninformative prior distribution $p^*(\eta)$ of $\eta$ could be derived as follows:

$$p^*(\eta) \propto p\left(g^{-1}(\eta)\right)\left|\frac{d\theta}{d\eta}\right|,$$

Based on equation (2.3.1.), the above relation could be simplified to:

$$p^*(\eta) \propto \left|\frac{d\theta}{d\eta}\right|,$$

$$p^*(\eta) \propto \left|\frac{d}{d\eta}(\ln(\eta))\right|,$$

then

$$p^*(\eta) \propto \eta^{-1}. \tag{2.3.2}$$

Therefore, the noninformative prior distribution of $\eta$ have to be proportional to $\eta^{-1}$ to maintain consistency and to obtain the same answers in each parameterization. Thus, the consistency could not be satisfied if a constant prior distribution is chosen for both $\theta$ and $\eta$ since (2.3.1) in terms of $\eta$ would not meet the third characteristic (see Berger, 1985).

From what stated above, the argument that the lack of prior information should correspond to the constant density (2.3.1) would be hard to defend in general. Therefore, the lack of invariance of (2.3.1) motivated Jeffreys to search for noninformative priors that are appropriately invariant under transformations.

**Jeffreys' second rule:**

If the parameter $\sigma$, the standard deviation in scale densities for example, may have any value in a semi-infinite range from 0 to $\infty$, the prior distribution of its logarithm should be taken uniformly distributed in the form:

$$p[\ln(\sigma)] \propto \text{constant},$$

which is equivalent to

$$p(\sigma) \propto \frac{1}{\sigma} \tag{2.3.3}$$

This distribution is termed as "Jeffreys-Haldan" distribution. An interesting natural application of this distribution is the "table entry" problem, which represents the study

of positive entries in various natural numerical tables, such as table of population sizes and tables of positive physical constants.

The form in (2.3.3) can be proved using Jeffreys' first rule in (2.3.1). The proof is the same as presented for (2.3.2). The prior distribution in (2.3.3) is again an improper prior distribution.

Jeffreys pointed out that (2.3.3) has the property $\Pr\{0<\sigma<\mathrm{a}\}=\Pr\{a<\sigma<\infty\}=\infty$, which indicates that nothing is known about $\sigma$, the case of complete ignorance provided by the first characteristic.

The previous property implies that the ratio $\Pr\{0<\sigma<\mathrm{a}\}/\Pr\{\mathrm{a}<\sigma<\infty\}$ is indeterminate, where "a" is any finite number, and thus nothing can be said about the odds of these two probabilities, which correspond to the second characteristic. Again this indeterminacy is considered as a formal presentation of ignorance.

With reference to the third characteristic, Jeffreys observed that (2.3.3) is invariant to the one-to-one transformations only in the form $\eta=\sigma^{\mathrm{n}}$, in other words (2.3.3) is invariant under positive or negative powering of $\sigma$. This is an important property, because some investigators parameterize models in terms of the standard deviation $\sigma$ and others in terms of the variance $\sigma^2$, or the precision parameter $\tau=\sigma^{-2}$.
Checking the invariance property of the powering transformation can be done as follows (in case if *n=2* for example):
Let

$$\eta=f(\sigma), \tag{2.3.4}$$

where $f(\sigma)=\sigma^2$ in this case, then applying the change of variable technique using the distribution in (2.3.3) will lead to

$$p^*(\eta)\propto p\left(f^{-1}(\eta)\right)\left|d\sigma/d\eta\right|,$$

where $p^*(\eta)$ is the required noninformative prior, then

$$p^*(\eta)\propto \eta^{-1/2}\,\eta^{-1/2},$$

then

$$p^*(\eta)\propto\frac{1}{\eta}. \tag{2.3.5}$$

Then $p^*(\sigma^2)\propto 1/\sigma^2$ has the same form as (2.3.3). Thus, applying Jeffreys' rule in (2.3.3) to different parameters of the form $\sigma^n$ provides prior p.d.f.'s of the same form

and consistent with each other. These prior distributions are consistent in the sense that posterior probability statements based on the alternative parameters will be also consistent. i.e., if an investigator "A" parameterizes a model in terms of $\sigma$ and uses (2.3.3) as his prior p.d.f., whereas another investigator "B" parameterizes the model in terms of $\eta$ and uses (2.3.5) as his prior p.d.f., they would get their posterior p.d.f.'s in terms of their own parameter. If the invariance property is satisfied, "B" can use (2.3.4) to transform his posterior distribution in terms of $\sigma$ and gets the same posterior distribution that "A" has obtained. Alternatively, "A" can use (2.3.4) to transform his posterior distribution in terms of $\eta$ to get the posterior distribution that "B" has obtained (see Zellner, 1971).

**Jeffreys' general rule:**

Jeffreys (1961) generalized the invariance property base to develop the noninformative prior distribution and hence, solve more general problems such as problems involving multiparameter cases. Jeffreys pointed out that the prior p.d.f. of the parameter vector $\boldsymbol{\theta}$ should be taken as:

$$p(\boldsymbol{\theta}) \propto \left|\mathrm{Inf}_{\boldsymbol{\theta}}\right|^{1/2}, \tag{2.3.6}$$

such that

$$\mathrm{Inf}_{\boldsymbol{\theta}} = -E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right],\ i,j = 1,2,..,k, \tag{2.3.7}$$

where $\boldsymbol{\theta}' = (\theta_1\ \theta_2 \ldots \theta_\kappa)$ is the k-vector of parameters defined on the space $\Omega \in R^K$, $\mathbf{y}$ is the n-vector of observations having the p.d.f. $p(\boldsymbol{y}|\boldsymbol{\theta})$ over the space $S \subset R^n$, which has continuous $\boldsymbol{\theta}$ derivatives for all $\boldsymbol{y} \in S$, $\mathrm{Inf}_{\boldsymbol{\theta}}$ is the (k × k) Fisher's information matrix for the parameter vector $\boldsymbol{\theta}$, and $E$ denotes the expectation with respect to the p.d.f. of $\mathbf{y}$.

The most important property of Jeffreys' prior in (2.3.6) is the invariance property, in the same sense explained in the previous section. Thus if $\boldsymbol{\eta} = \mathbf{G}(\boldsymbol{\theta})$, where $\mathbf{G}$ is a one-to-one differentiable transformation of $\boldsymbol{\theta}$, then the invariance virtue of (2.3.6) involves that, the prior p.d.f. of $\boldsymbol{\eta}$ should be taken as

$$p(\boldsymbol{\eta}) \propto \left|\mathrm{Inf}_{\boldsymbol{\eta}}\right|^{1/2}.$$

Therefore, the posterior probability statements will be consistent for all problems that are parameterized in terms of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ (for the proof of this property, one could refer to Zellner (1971, p. 47)).

It is important here to notice that Jeffreys himself pointed out that such multiparameter rule must be applied with caution, especially in scale and location parameters problems that occur simultaneously. He also emphasized that this rule must be examined to avoid adding some unwanted prior information into analysis. This guides Jeffreys to assume the following rule:

**Jeffreys' independence rule:**

In such cases he suggested treating location parameters separately. Thus, consider the case when the parameters' vector specified by $(\mu_1 \mu_2 ... \mu_\kappa \ \boldsymbol{\theta})'$ such that $\mu_i$'s denote the location parameters whereas $\boldsymbol{\theta}$ is an additional vector of parameters that includes the scalar parameters. Then the modified general rule recommended by Jeffreys is given by

$$p(\mu_1, \mu_2, ..., \mu_\kappa, \boldsymbol{\theta}) \propto \left| \text{Inf}_{\boldsymbol{\theta}} \right|^{1/2}, \tag{2.3.8}$$

Kass and Wasserman (1996) called (2.3.8) Jeffreys' location general rule while called (2.3.6) Jeffreys' non-location general rule.

**Difficulties encounter applying Jeffreys' general rule**

The major difficulty associated with the application of Jeffreys' rule in (2.3.6) arises when parameters of different types are considered simultaneously. For example, in problems containing both location and scale parameters, Jeffreys (1961) avoided applying (2.3.6) and derived alternatively a noninformative prior density assuming independence between parameters of different types. This modification leads to the rule given by (2.3.8). Jeffreys explained the nonuse of his general rule, in (2.3.6), in such cases by his deem that this rule will lead to dissatisfying results and the derived prior density based on it will be inferior. So he, instead, derived the noninformative prior by assuming independence then applying the rule separately to parameters of each type (that is the rule in (2.3.8)). He also proposed that the resulted prior distribution is invariant under transformations of a certain type. Zellner (1971) proved

that Jeffreys' prior assuming independence is "minimal information" prior. This concept will be explained in a next section (see §2.5).

On the contrary, Box and Tiao (1973) stated in this respects, that the independence assumption between parameters of different types could appear inappropriate in certain cases. They showed some applications in which the location and the scale parameters could be dependent according to the nature of data of interest. To overcome this problem, they suggested manipulating data by adapting some appropriate transformations such as taking the logarithm of the original data to remove constrained dependence. For further details and explanations, one could refer to the example which they introduced (see Box and Tiao, 1973).

Another difficulty, that impedes working with Jeffreys' general rule, is that the rule could not be applied with distributions of non-regular type and distributions indifferentiable with respect to parameters (Huzurbazar, 1980). In this respect, Jeffreys himself realized that his assumption only works under regularity conditions in one parameter; in continuous problems (see Irony and Singpurwalla, 1996).

Regarding the invariance requirement attained by Jeffreys' rule, Jeffreys insists on viewing his rule as unique for any given model, which is considered to be wrong by many other statisticians. Huzurbazar (1980) considered seeking a single invariance rule, which is adequately applicable to all distributions, as an impossible hope. He deemed that such hope is as unlikely as discovering a single scientific law to explain satisfactorily all physical phenomena.

In essence, the application of Jeffreys' rule leads to inappropriate results in large dimensional parameter space and in distributions of non-regular type as well. Hence an inevitable technique of noninformative prior will be required. Some other techniques could be available to produce a noninformative prior distribution that fits those cases, see Berger and Bernardo (1989 and 1992) who developed the reference prior.

### 2.3.3. Properties

For long decades and till nowadays, Jeffreys' prior has been considered the most widely used standard noninformative prior in many applications, particularly in

Bayesian time series applications. This is due to its simplicity in being derived automatically. Therefore, it was important to provide a subsection here discussing the two main properties of Jeffreys' prior. First, being an improper prior is considered by many others to be a flaw. Second, the invariance consideration involved by such prior is considered by many as a great contribution.

**Impropriety**

It happens frequently that noninformative priors are improper, which means that it has infinite mass. In such case the function used is not a probability density at all. Many statisticians consider this a serious drawback of the noninformative Bayesian analysis because it is hard to apply it with problems in estimation and inferences (Koop, 1994). A reasonable response to this criticism revealed by Bernardo through his discussion of noninformative priors in Irony and Singpurwalla (1996, answer to question 7) was:

> One should not interpret any noninformative prior as a probability density. Noninformative priors are merely technical devices to produce non-subjective posterior distributions by formal use of Bayes theorem and sensible non-subjective posterior distributions are always proper.

This involves that the improper noninformative priors will be "unacceptable" if they yield improper posterior distributions. So the *propriety* of the posterior distributions should be the first property required when improper noninformative prior is applied even for minimum sample size as it will be illustrated in following parts.

The previous discussion has not confined the noninformative priors to be improper, however, proper noninformative priors are usually found whenever the parameter space is bounded (see example 2.3.1), although this is not a general case.

Jeffreys was the first to propose an axiomatic foundation of improper noninformative priors. The most applicable use of such improper distributions is in elementary quantum mechanics (Good, 1980).

Jeffreys considered using improper noninformative priors as the best way to describe the case of complete ignorance. Commenting on this, O'Hagan (1994) argued that there is no prior information that is completely absent, but improper priors are

used precisely to reflect weak information relative to the data. In such point of view, the posterior distribution will generally be robust to any reasonable choice of noninformative prior even improper one. In the same sense (see Irony and Singpurwalla, 1996) similar representation of noninformative priors are adopted, whether proper or improper, to construct a posterior distribution that reflects data dominance.

Another two interested arguments must be mentioned. First, in the case of large sample sizes choices of noninformative priors, whatever proper or improper, will have minor effect on the posterior results (Gelman, 2002). Second, an improper prior can be approximated by a proper one, for example the Jeffreys-Haldan distribution in (2.3.3) can be approximated very closely by a log Caushy distribution (Good, 1980).

**Invariance**

One of Jeffreys' great contributions to Bayesian inference is that he introduced and developed invariance considerations into the Bayesian system (Geisser, 1980). Furthermore, Jeffreys' prior was the first explicit use of the concept of invariance in statistics and particularly in the selection of noninformative prior distributions (Good, 1980). He was then the first to set up rules for noninformative prior distributions that satisfy various invariance principles, as will be illustrated below. On the other hand, the invariance principle is suitable only when no prior information is available, so the analysis of invariance will correspond to Bayesian analysis with noninformative priors (Berger, 1985).

The invariance requirements are crucial for sensible posterior distributions that are based on noninformative prior. Furthermore, one should not seriously consider an assumption for noninformative Bayesian inference which does not satisfy them. The importance of meeting invariance principles could briefly be due to the following reasons as mentioned in Berger (1985):

1. People who don't like to talk about noninformative priors are welcome to do the same procedure in terms of invariance.

2. Existence of many choices of noninformative prior, which is considered as a crucial criticism to noninformative Bayesian analysis, will be restricted to one particular noninformative prior if invariance is satisfied.

Most of the recent use of invariance could be traced back directly or indirectly to Jeffreys' work (Good, 1980). Later efforts to derive noninformative priors through considerations of transformations of a problem had been extensively used in Hartigan (1964), Jaynes (1983 and 1968), and Villegas (1977, 1981 and 1984) and Berger (1985).

The most apparent noticeable property of Jeffreys' general prior is that it satisfies all requirements of invariance concept as Hartigan (1964) proposed. Zellner (1971) introduced the invariance principles that they all hold through applying Jeffreys' prior in (2.3.6) according to the establishment of Hartigan (1964). These requirements are as follows:

Let $\mathbf{y}$ be the n-vector of observations having the p.d.f. $p(\mathbf{y}|\boldsymbol{\theta})$ over the space $S \subset R^n$, which has continuous $\boldsymbol{\theta}$ derivatives for all $\mathbf{y} \in S$, where $\boldsymbol{\theta} = (\theta_1\ \theta_2 \ldots \theta_\kappa)$ is the k-vector of parameters defined on the space $\Omega \in R^K$. Hartigan (1964) established that if the Jeffreys' prior in (2.3.6) is considered, the Bayesian transformation, by combining the prior information with sample information, will have the following invariance properties:

1. ***S-Labeling Invariance:*** Let $\mathbf{z} = \mathrm{G}(\mathbf{y})$ be a differentiable one-to-one transformation that takes the sample space S for $\mathbf{y}$ into $\mathrm{S}^*$, the sample space for $\mathbf{z}$, then

$$p(\boldsymbol{\theta}|\,\mathbf{z}) \propto p(\boldsymbol{\theta}|\,\mathbf{y}),$$

where $p(\boldsymbol{\theta}|\,\mathbf{z})$ and $p(\boldsymbol{\theta}|\,\mathbf{y})$ denote posterior distributions of $\mathbf{z}$ and $\mathbf{y}$ respectively. This property is important particularly if this transformation in data involves a change in the units of measurement.

2. ***Ω-Labeling Invariance:*** Let $\boldsymbol{\eta} = \mathbf{F}(\boldsymbol{\theta})$ be a differentiable one-to-one transformation of $\boldsymbol{\theta}$, then

$$p(\boldsymbol{\eta}|\,\mathbf{y}) \propto p(\boldsymbol{\theta}|\,\mathbf{y}),$$

where $p(\boldsymbol{\eta}|\,\mathbf{y})$ and $p(\boldsymbol{\theta}|\,\mathbf{y})$ denote posterior distributions of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ respectively. This property, as mentioned above, is important in different parameterization of the problem.

3. ***Ω-Restriction Invariance:*** Assume that $\boldsymbol{\theta} \in \Omega^* \subset \Omega$. Then

$$p^*(\boldsymbol{\theta}|\,\mathbf{y}) \propto p(\boldsymbol{\theta}|\,\mathbf{y}),$$

where $p^*(\boldsymbol{\theta}|\,\mathbf{y})$ is the posterior distribution based on $p^*(\mathbf{y}|\,\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Omega^*$. This property means that Jeffreys' prior is not affected by a restriction on the parameter space. In other words, applying Jeffreys' prior under restriction on the parameter space will lead to the same posterior.

4. ***Sufficiency Invariance:*** Let $\mathbf{t}' = (t_1\, t_2 \ldots t_m)$ be a vector of sufficient statistic of $\boldsymbol{\theta}$. Then:

$$p(\boldsymbol{\theta}|\,\mathbf{t}) \propto p(\boldsymbol{\theta}|\,\mathbf{y}),$$

where $p(\boldsymbol{\theta}|\,\mathbf{t})$ is the posterior distribution obtained from the model $p(\mathbf{t}|\,\boldsymbol{\theta})$. In this regard, Jeffreys' rule will lead to appropriate prior distributions for all well-known distributions that admit a sufficient statistic for a parameter, but merely in the case of single parameter.

5. ***Direct Product Invariance:*** Let $\mathbf{y}_1$ and $\mathbf{y}_2$ be two independent sample vectors each of n×1, then

$$p(\boldsymbol{\theta}|\,\mathbf{y}) \propto p_1(\boldsymbol{\theta}_1|\,\mathbf{y}_1)\; p_2(\boldsymbol{\theta}_2|\,\mathbf{y}_2),$$

where $p(\boldsymbol{\theta}_i|\,\mathbf{y}_i) \propto p_i(\boldsymbol{\theta}_i)\; p_i(\mathbf{y}_i|\,\boldsymbol{\theta}_i)$, for $i = 1,2$, $\boldsymbol{\theta}_1 \in \Omega_1, \boldsymbol{\theta}_2 \in \Omega_2$, $\boldsymbol{\theta} \in \Omega = \Omega_1 \times \Omega_2$, and the prior p.d.f.'s $p_1(\boldsymbol{\theta}_1), p_2(\boldsymbol{\theta}_2)$, and $p(\boldsymbol{\theta})$ are each taken in Jeffreys' form in (2.3.6).

6. ***Repeated Product Invariance:*** Suppose that $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$ are each n×1 independent observations vector and each is from $p(\mathbf{y}|\,\boldsymbol{\theta})$, the same as for $\mathbf{y}$. Then

$$p(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m|\,\boldsymbol{\theta}) = \prod_{i=1}^{m} p(\mathbf{y}_i|\,\boldsymbol{\theta}),$$

and

$$p^*(\boldsymbol{\theta}|\,\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m) = p(\boldsymbol{\theta}) \prod_{i=1}^{m} p(\mathbf{y}_i|\,\boldsymbol{\theta}),$$

and

$$p^*(\boldsymbol{\theta}\,|\,\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_m) = p(\boldsymbol{\theta}\,|\,\mathbf{y}_1)\prod_{i=2}^{m} p(\mathbf{y}_i\,|\,\boldsymbol{\theta}),$$

where $p(\boldsymbol{\theta})$ is the form in (2.3.6).

## 2.3.4. Examples

In this subsection, the derivation of Jeffreys' rule will be illustrated to several distributions from which observations are generated. The same examples will be demonstrated in the following sections for discussing other approaches for selecting noninformative priors.

**Example 2.3.1:** ***Binomial*** **($\theta$)**

According to this distribution, the parameter of interest $\theta$, defined over the range [0,1], is the probability of success in each trial of total fixed number of trials n. Then an observation y (the number of success) will be distributed as follows:

$$p(y\,|\,\theta) \propto \theta^{y}(1-\theta)^{n-y}, y = 0,1,...,n.$$

The importance of providing such a distribution is that there are several candidates in the literature for the noninformative prior form of the Binomial parameter.

Deriving Jeffreys' prior requires computing the square root of the determinant of Fisher's information matrix, which is a scalar in such case, having the form

$$\mathrm{Inf}_{\theta} = -E_{y|\theta}\left[\frac{d^2 \log p(\mathrm{y}\,|\,\theta)}{d\theta^2}\right].$$

Then the computation can be proceeded as follows:

$$\log p(y|\theta) \propto y\log\theta + (n-y)\log(1-\theta),$$

then

$$\frac{d\log p(y\,|\,\theta)}{d\theta} \propto \frac{y}{\theta} - \frac{(n-y)}{1-\theta},$$

and

$$\frac{d^2 \log p(y\,|\,\theta)}{d\theta^2} \propto \frac{-y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2},$$

$$\mathrm{Inf}_{\theta} \propto -E_{y|\theta}\left[\frac{-y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}\right],$$

$$\propto \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2},$$

$$\propto \frac{n}{\theta(1-\theta)}.$$

Jeffreys' prior, which has the form $p(\theta) \propto \sqrt{\text{Inf}_\theta}$, will be

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} \quad . \tag{2.3.9}$$

This is a proper distribution well known as *Beta*(1/2, 1/2) which is also called as the *arc-sine* distribution. Some different plausible alternative suggestions to this distribution will be seen later.

**Example 2.2.3:** ***Normal*** **($\theta$)**

In such distribution the observations are generated from a normal distribution with a known variance. The unknown parameter $\theta$, which is the location parameter, is the parameter of interest defined over the parameter space $(-\infty, \infty)$. This distribution, which belongs to the family of location densities, is in the form

$$p(y|\theta) \propto e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, y \in (-\infty, \infty).$$

As illustrated in the preceding example, the algorithm of driving Jeffreys' prior will be as follow:

$$\log p(y|\theta) \propto (y-\theta)^2,$$

$$\frac{d \log p(y|\theta)}{d\theta} \propto (y-\theta),$$

$$\frac{d^2 \log p(y|\theta)}{d\theta^2} \propto \text{constant}$$

Hence, Jeffreys' prior of such normal mean will be

$$p(\theta) \propto \text{constant},$$

which is in the same form as Jeffreys suggested in his first rule in (2.3.1).

**Example 2.3.3:** ***Normal*** **($\sigma$)**

The observations are also generated from a normal distribution but with a known mean. The unknown parameter $\sigma$, which is the scale parameter, is the parameter of interest, defined over the parameter space $(0, \infty)$. This distribution, which belongs to the family of scale densities, is in the form

$$p(y|\sigma) \propto \sigma^{-1} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, y \in (-\infty, \infty).$$

Deriving Jeffreys' prior can be done through the following steps:

$$\log p(y|\sigma) \propto -\log\sigma - \frac{1}{2\sigma^2}(y-\theta)^2,$$

$$\frac{d \log p(y|\sigma)}{d\sigma} \propto \frac{-1}{\sigma} + \frac{(y-\theta)^2}{\sigma^3},$$

$$\frac{d^2 \log p(y|\sigma)}{d\sigma^2} \propto \frac{1}{\sigma^2} - \frac{3(y-\theta)^2}{\sigma^4},$$

$$\mathbf{Inf}_\sigma \propto -E_{y|\sigma}\left[\frac{1}{\sigma^2} - \frac{3(y-\theta)^2}{\sigma^4}\right],$$

$$\propto \frac{-1}{\sigma^2} + \frac{3}{\sigma^2},$$

$$\propto 2\sigma^{-2},$$

then, Jeffreys' prior in such case which takes the form $p(\sigma) \propto \sqrt{\text{Inf}_\sigma}$ , will be

$$p(\sigma) \propto \sigma^{-1}$$

which is in the same form as Jeffreys suggests in his second rule in (2.3.3).

**Example 2.3.4:** ***Normal*** **$(\theta, \sigma)$**

In such case the observations are generated from a location-scale normal distribution with unknown mean and variance. The parameter space over which the parameters are defined is the same as mentioned in the above two examples, for $\theta$ in example 2.3.2 and for $\sigma$ in example 2.3.3. The form of this distribution is as follow:

$$p(y|\theta,\sigma) \propto \sigma^{-1} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, y \in (-\infty, \infty).$$

To follow the procedure for deriving Jeffreys' prior we need first to calculate,

$$\log p(y|\theta,\sigma) \propto -\log\sigma - \frac{1}{2\sigma^2}(y-\theta)^2,$$

then to find the Fisher's information matrix, which is symmetric having the form

$$\text{Inf}_{\theta,\sigma} \propto -E_{y|\theta,\sigma}\begin{pmatrix} \frac{\partial^2 \log p(y|\theta,\sigma)}{\partial\theta^2} & \frac{\partial^2 \log p(y|\theta,\sigma)}{\partial\theta\,\partial\sigma} \\ \frac{\partial^2 \log p(y|\theta,\sigma)}{\partial\theta\,\partial\sigma} & \frac{\partial^2 \log p(y|\theta,\sigma)}{\partial\sigma^2} \end{pmatrix},$$

$$\propto \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

then, Jeffreys' prior with the form $p(\theta,\sigma) \propto \sqrt{\left|\text{Inf}_{\theta,\sigma}\right|}$, will be

$$p(\theta,\sigma) \propto \sigma^{-2}. \tag{2.3.10}$$

Jeffreys discarded to use this prior, as mentioned in a previous subsection, and recommended alternatively another prior distribution that resulted from assuming independence between both the location and the scale parameters. That leads to the Jeffreys' non-location rule.

The main reason of considering this result inappropriate is that, when the model extended to the k-means and a common unknown variance, the marginal posterior distribution of the location parameters is the student-t with degrees of freedom depend only on the sample size regardless the value of k (see Zellner, 1971). So given this assumption the joint prior density in this case taken as the product of the Jeffreys' priors for the mean parameter $\theta$ and the scale parameter $\sigma$ separately to get the joint prior in the form

$$p(\theta,\sigma) \propto \sigma^{-1}, \tag{2.3.11}$$

which is the result of the product of $p(\theta) \propto$ constant and $p(\sigma) \propto \sigma^{-1}$. This result can be obtained, as well, by applying directly Jeffreys' location general rule in (2.3.8). This final form is the recommended prior distribution by Jeffreys to this problem.

**Example 2.3.5:** ***k-Normal*** **($\boldsymbol{\theta}$, $\sigma$)**

It is essential to provide another example for a multiparameter distribution. Therefore, the distribution of k-normal independent populations with k-vector of unknown means $\boldsymbol{\theta}' = (\theta_1, \theta_2, ..., \theta_k)$, defined on the parameter space $\Omega \in R^k$, and unknown common standard deviation $\sigma$, which is defined over $(0,\infty)$, will be presented here. In such distribution there are k independent random samples $\mathbf{y}_i$'s , each of size $n_i$ defined over the sample space $S \subset R^{n_i}$ and each also generated from *Normal*($\theta_i, \sigma$), where, $i$ =1,2,…,$k$. The joint distribution of the k-vector of samples $\mathbf{y}' = (\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_k)$ will be in the form

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma) \propto \sigma^{-k} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathrm{y}_i - \theta_i)^2}.$$

Derivation of Jeffreys' prior in (2.3.6) requires computing the following term

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma) \propto -k \log \sigma - \frac{\sum_{i=1}^{k}(\mathrm{y}_i - \theta_i)^2}{2\sigma^2}.$$

Then compute the Fisher's information matrix in (2.3.7), with a minor difference that there is another additional parameter $\sigma$, which means that the matrix will be of order *(k+1)×(k+1).* The elements of this matrix will be computed according to the following steps:

The first k elements on the main diagonal will be proportional to

$$\propto -E_{\mathbf{y}|\,\boldsymbol{\theta},\sigma}\left[\frac{\partial^2 \log p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma)}{\partial {\theta_i}^2}\right],\ i=1,2,..,k,$$

$$\propto \sigma^{-2}.$$

But the last $(k+1)^{th}$ element on the main diagonal will equal to

$$\propto -E_{\mathbf{y}|\,\boldsymbol{\theta},\sigma}\left[\frac{\partial^2 \log p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma)}{\partial \sigma^2}\right],$$

$$\propto 2k\sigma^{-2}.$$

Since $\mathrm{Inf}_{\boldsymbol{\theta},\sigma}$ is symmetric, the off-diagonal elements except for the last row and last column will be hence in the form

$$-E_{\mathbf{y}|\,\boldsymbol{\theta},\sigma}\left[\frac{\partial^2 \log p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma)}{\partial\theta_i \partial\theta_j}\right],\ i\neq j=1,2,..,k,$$

$$\propto zero.$$

Similarly, the off-diagonal elements on the last row and the last column will equal :

$$-E_{\mathbf{y}|\,\boldsymbol{\theta},\sigma}\left[\frac{\partial^2 \log p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma)}{\partial\sigma \partial\theta_i}\right],\ i=1,2,..,k,$$

$$\propto Zero.$$

It is noticed, so far, that the Fisher's information matrix for this problem is of diagonal type. Therefore, Jeffreys' prior, which is the square root of the determinate of Fisher's information matrix, will be computed as follows:

$$p(\boldsymbol{\theta},\sigma)\propto\sqrt{(\prod_{i=1}^{k}\sigma^{-2})\times(2k\sigma^{-2})},$$

$$p(\boldsymbol{\theta},\sigma)\propto\sqrt{\sigma^{-2k}\times\sigma^{-2}}.$$

Then, Jeffreys' prior for this problem based on his general rule in (2.3.6) will have the form

$$p(\boldsymbol{\theta},\sigma)\propto\sigma^{-(k+1)}. \qquad (2.3.12)$$

Jeffreys deemed this last derived form as a dissatisfying prior. Instead, he derived another noninformative prior form for this problem assuming independence between both the vector of means and the standard deviation, where both are of different kinds. He then, applied his general rule separately to each type of parameters, to get $p(\boldsymbol{\theta}) \propto \text{constant}$ as a noninformative prior distribution for $\boldsymbol{\theta}$ and $p(\sigma) \propto \sigma^{-1}$ as a noninformative prior distribution for $\sigma$. This can easily be proved in the two cases for this problem. Then the joint noninformative prior distribution will result from the product of these marginal distributions to be

$$p(\boldsymbol{\theta}, \sigma) \propto \sigma^{-1}. \tag{2.3.13}$$

This was the final form accepted by Jeffreys and practically applied in similar problems. Zellner (1971) considers (2.3.12) as more informative for large k than the (2.3.13). In other words, Zellner described (2.3.13) as "minimal information prior". This latter concept introduced by Zellner (1971), as another tool to derive noninformative prior forms, will be discussed later (see §2.5). Zellner generally, and particularly in such problem, explained Jeffreys' departure from his general rule by his concern to add inconvenient prior information to the analysis.

## 2.4. Locally Uniform Prior

### 2.4.1. Introduction

Box and Tiao (1973) objected the reckless application of Bayes' postulate to characterize the situation where nothing is known about the parameter. They also disagreed with the realistic existence of "complete ignorance" state of knowledge about the parameter. The state of "knowing little" is considered to have meaning only relative to the information provided by an experiment. This refutation has been justified by many reasons. The most noticeable one, which has been indicated in the preceding section, is lacking of this postulate leads to consistent posterior distributions if it is applied to different transformations of the parameter using the same data. However, they did not absolutely reject the uniform prior distribution. They permit using it approximately in certain cases such as:

1. In cases where the range of uncertainty of the parameter is not large, since many transformations would be nearly linear such as logarithmic and reciprocal. However, this argument would not necessarily work if a very extreme transformation was considered such as $\phi = \exp(\exp(\theta))$ or $\phi = \theta^{10}$, assuming that $\theta$ is the parameter of interest and $\phi$ is some transformation of this parameter.

2. For large or moderate-sized samples. Since fairly crucial modification of the prior distribution, through transformations in parameter may, only lead to minor modification of the posterior distribution.

Away from these limited cases, they proposed a tool for choosing a particular metric (transformation) in terms of which a uniform, or locally uniform distribution as they call, can be regarded as a noninformative prior distribution about the parameter. Such a noninformative prior distribution is termed by them as a ***reference prior*** which is used as a standard prior to characterize the situation of ignorance about parameter relative to the informative data.

**Relevance Concepts:**

The technique for choosing a noninformative prior distribution proposed by Box and Tiao (1973) is mainly inherent to some basic terms.

***Likelihood Function (LF)***

Suppose that **y** is a vector of n observations whose density $p(\mathbf{y}|\theta)$ depends on the value $\theta$, the parameter of interest. $p(\mathbf{y}|\theta)$ is considered as a function of $\theta$ for fixed **y** not as a function of **y**. In such case, $p(\mathbf{y}|\theta)$ is called the likelihood function (LF) of $\theta$ given **y** and written as $l(\theta|\mathbf{y})$. Further, assuming $p(\theta)$ indicates to the prior distribution for $\theta$, Bayes' theorem is very often written in the form

$$p(\theta|\mathbf{y}) \propto l(\theta|\mathbf{y}) \times p(\theta).$$

Therefore, the LF plays a very important role in Bayes' formula. It is the function through which the data **y** modifies the prior beliefs about $\theta$. It can hence be regarded as representing information about $\theta$ coming from the data. The main properties of the LF can be summarized as follows:

1. Multiplication of LF by a constant, or generally by a function of the data **y** only, leaves the LF unchanged.
2. The LF should not have the same properties as $p(\mathbf{y}|\theta)$, that is $l(\theta|\mathbf{y})$ is not always integrated or summed to unity. In this case, the LF is often scaled so that the area under the curve is one as follow

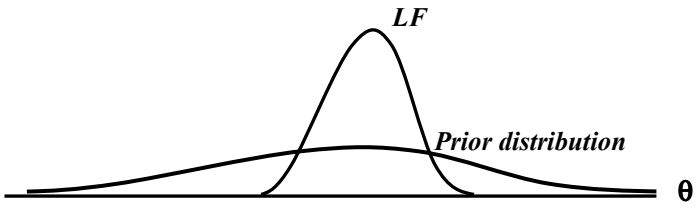
$$\frac{l(\theta|\mathbf{y})}{\int l(\theta|\mathbf{y})\,d\theta}. \tag{2.4.1}$$

This quantity is often termed as ***standardized LF***.

***Dominance LF***

Box and Tiao were concerned with problems of scientific inference occurring in scientific investigation. They deemed that analyzing scientific data would often be sensible on the assumption that the LF ***dominates*** the prior. For more clarification of this concept, consider the example of the normal distribution with known variance, where $\theta$ is the location parameter. The concept of dominant LF may be illustrated by the following figure.

**Figure 2.4.1**
**The LF dominates the prior distribution**




It is obvious from the above figure that the LF reflects less uncertainty about $\theta$ compared to that reflected by the prior distribution. So the LF tends to be more informative about $\theta$ than the prior distribution, whose shape indicates that little is known about $\theta$. Such relationship between the LF and the prior distribution figures that the prior is dominated by the LF.

Box and Tiao motivated the dominance of the LF in scientific investigation for many reasons such as:

1. An experiment, for sake of scientific investigation, is not usually undertaken unless information supplied by it is likely considered more significant than information already available.
2. It is appropriate for a scientist who has strong prior beliefs, which strictly disagree with what others have, to begin with deriving a posterior distribution that represents the view of someone else who has no strong beliefs or knows little about the parameter in the light of data. Such posterior distribution can merely be derived using prior distribution dominated by the LF.

***Locally Uniform Prior***

A basic property of the Bayes' postulate is that it is an ***improper*** distribution. Box and Tiao (1973) were hesitant to employ the improper p.d.f.'s recommended by Jeffreys through (2.3.1) and (2.3.3). They rather used such densities to express the ***local*** behavior of the prior distribution of the parameter over the region where the LF is appreciable but not over its entire admissible range. They have used the term ***local*** as a remedy to impropriety, which have been considered by them as impractical to occur. So by assuming the prior approximately follows (2.3.1) or (2.3.3) only over the range of appreciable LF and tails to zero outside that range, the resulted priors used are actually proper and have hence more practical sense.

Considering the above argument only as for the uniform distribution in (2.3.1), that can be regarded as a normal distribution with infinite variance. It can hence be approximately having this form ***locally*** over some (possibly very large) interval, precisely over the range of appreciable LF, and is never very large outside it. The posterior distribution derived based on such prior distribution is approximately numerically equal to the standardized LF in (2.4.1). It follows that the dominant feature of the posterior is the LF, Lee (1989). Such a prior distribution used in this case, which is a proper one, is termed as ***locally uniform prior***. So Box and Tiao overcome the theoretical difficulty of the impropriety of the uniform distribution in (2.3.1) by introducing instead, the practical sensible, the locally uniform prior distribution.

In general, the locally uniform prior is the prior which is dominated by the LF and does not change very much, or may be further considered as reasonably flat, over the region in which the LF is appreciable and does not assume large values outside that range (see Fig. 2.4.1). Kass and Wesserman (1996) considered such suggestion of the locally uniform prior by Box and Tiao (1973) as a response to the suspicions about the often impropriety property of many noninformative priors. Kass and Wesserman explained the use of locally uniform prior as a truncation of an improper prior to make its domain more compact and it hence becomes a proper distribution.

**Difficulties associated with locally uniform prior:**

It is of interest to bear in mind that, appealing to the locally uniform prior, as a remedy to the impropriety, has not yet so far wiped out the crucial flaw of its being self-inconsistent. It's being so, in the sense when it is applied to different transformations for the original parameter, which has just been indicated at the beginning of this section. Box and Tiao (1973) have devoted an effective technique for choosing a noninformative prior that overcomes such crack, as will be demonstrated.

***Data Translated LF***

Box and Tiao (1973) introduced the notion of data-translated LF to refine the use of locally uniform priors. For more assimilation to such terminology, consider again the example of the random sample **y** of size n from the normal distribution with known variance $\sigma^2$. The LF for $\theta$, the location parameter, can be considered to be normal distribution with mean equal the sample mean $\bar{y}$, and standard deviation $\sigma/\sqrt{n}$. This LF has precisely the following form

$$l(\theta \mid \sigma, \mathbf{y}) \propto e^{\frac{-n}{2\sigma^2}(\theta-\bar{y})^2}.$$

Considering different sets of data represented by different values for $\bar{y}$'s, the standardized LF curves would have the appearance shown in figure 2.4.2(a). It obviously illustrates how different sets of data ***exactly translate*** the LF curves on the $\theta$ axis but leave it otherwise unchanged, with same functional form, except for a shift in location. Now if the locally uniform prior is taken for $\theta$, the posteriors based on these sets of data will be also the same except for their locations. That's why Box and

Tiao considered that it seems sensible to adopt a locally uniform prior when the LF is data translated.

It is now important to provide the case when the LF is not data translated in terms of the parameter of interest. Consider, therefor, the case when the random sample **y** is generated from normal distribution with known mean but unknown variance $\sigma^2$. Then the LF for $\sigma$ will be in the form

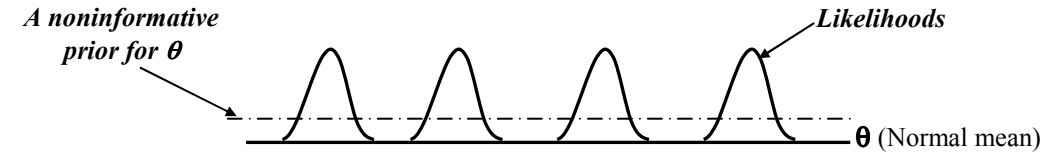
$$l(\sigma \mid \theta, \mathbf{y}) \propto \sigma^{-n} e^{\frac{-1}{2\sigma^2}[(n-1)s^2 + n(\theta - \bar{y})]},$$

where

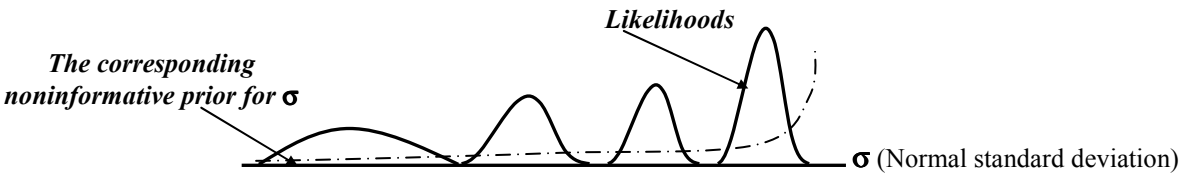
$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}.$$

In such a case, the standardized LF curves in terms of the original metric $\sigma$ with different data sets, expressed by different values of S's, as shown in figure 2.4.2(b).

**Figure 2.4.2**
**The standardized LF shapes relative to different sets of data and the corresponding noninformative prior distributions of the parameter**




**(a)The likelihood is data translated in terms of θ**




**(b)The likelihood is not data translated in terms of σ**

Box and Tiao considered that the noninformative prior for $\sigma$ should not be taken as locally uniform distribution.

In such a case, they suggested to express the unknown parameter $\sigma$ in terms of another metric, say $\phi(\sigma)$, so that the corresponding LF for this transformation is exactly data translated. That is the LF curves for $\phi(\sigma)$ are unchanged via data sets except for their locations. The locally uniform prior could hence be sensible to be

assigned as a noninformative prior for $\phi(\sigma)$. Then the corresponding prior distribution for $\sigma$ could hence be easily derived by the usual change-of-variable rule based on the distribution of $\phi(\sigma)$. Such a resulted prior distribution for $\sigma$ is termed as noninformative prior.

In essence, for the moment, the above argument strengthens the use of locally uniform prior as long as it guarantees the LF to be exactly data translated. That is, in another words, the LF is said to be data translated when the sets of data only serve to relocate the LF with the same functional form. However, if this is not the case another transformation for the original metric is still a natural urge to be sought for and that makes the LF in terms of which exactly data translated. The locally uniform distribution is hence chosen as a noninformative prior for such transformation. Then by formal rules of change-of-variable techniques the corresponding noninformative prior distribution of the original parameter could easily be derived based on the locally uniform distribution of the transformation.

***Multi-parameter Data Translated LF***

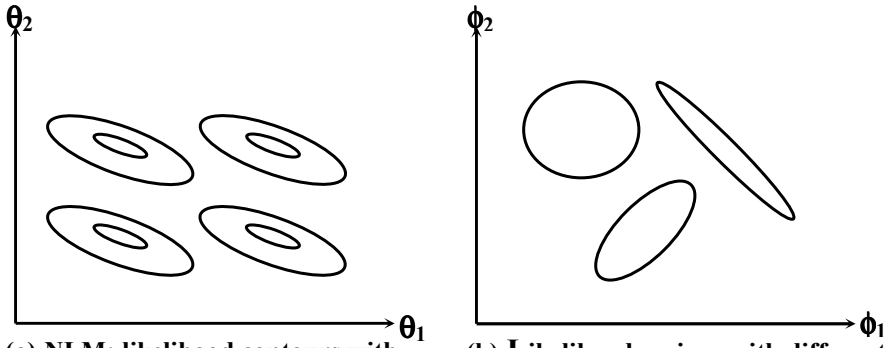
It is of wide interest to point out the manipulation of Box and Tiao to the same concept of "data translated LF" but within multi-parameter models. For illustration, consider the example of Normal linear model (NLM), i.e., $E(\mathbf{y}) = \mathbf{X\theta}$, where $\mathrm{y}' = (y_1\, y_2 \ldots y_n)$ is a set of Normally independent distributed random variables having common known variance $\sigma^2$, $\theta' = (\theta_1\, \theta_2 \ldots \theta_k)$ is a k-vector of unknown parameters, and $\mathbf{X}$ is the design matrix of order (n×k). The LF can be expressed as:

$$l(\boldsymbol{\theta} \mid \mathbf{y}, \sigma) \propto \exp\left\{\frac{-1}{2\sigma^2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\},$$

where $\hat{\boldsymbol{\theta}}$ is the vector of least squares estimate of $\boldsymbol{\theta}$. In case of k=2, figure 2.4.3(a) shows the same shape of LF contours for different sets of data represented by different values of $\hat{\theta}_1$ and $\hat{\theta}_2$. That is, data sets serve only to relocate the LF over $\theta_1$ and $\theta_2$ space and leave it with the same spread. In such case, the LF is data translated and the noninformative prior for $\boldsymbol{\theta}$ would be taken as locally uniform. When this is not the case, a transformation will be needed such that, in terms of which the LF is data translated. Then the procedure of selecting a noninformative prior for the original

parameters will be applied according to the same argument that followed above. Practically it is difficult to find a transformation that fulfills such concept. Generally a transformation would be necessary only to produce LF regions of the same size, see figure 2.4.3(b).

**Figure 2.4.3**
**Likelihood contours relative to different sets of data**




**(a) NLM: likelihood contours with same shape.**

**(b) Likelihood regions with different shapes have the same size.**

## 2.4.2. Derivation

It should be emphasized, so far, that the main issue is how to select a noninformative prior, or a reference prior as Box and Tiao called it, which provides little information about the unknown parameter relative to what is expected to be provided by the projected experiment.

In the light of the above discussion, the best procedure for Box and Tiao, is to recommend the locally uniform distribution as a noninformative prior provided that it satisfies the exact data-translated LF principle. That is because the locally uniform prior under this principle will produce posterior densities with the same form, except for their locations, for different samples. This feature of the locally uniform prior is what makes it noninformative. It is therefore convenient to devote this subsection to hold this concept, on which they based their derivation, in more details.

**(a) <u>Single parameter case</u>:**

Mathematically, according to Box and Tiao (1973), the LF is considered to be exactly data-translated if it may be written in the form

$$l(\theta \mid \mathbf{y}) \propto g[\phi(\theta) - t(\mathbf{y})] \tag{2.4.2}$$

where $\phi(\theta)$ is a one-to-one transformation of $\theta$, g(.) is a known function independent of **y** and t(**y**) is a function of y often expressing a sufficient statistic. If this is so, the noninformative prior of $\phi$ is taken to be locally uniform and the corresponding noninformative prior of $\theta$ is as follows

$$p(\theta) \overset{locally}{\propto} \left|\frac{d\phi}{d\theta}\right| \tag{2.4.3}$$

Box and Tiao stated further that, a transformation that allows the LF to be expressed exactly in the form (2.4.2) is not generally available. Thus, for a moderate sized samples all what would be necessary to require is a transformation $\phi(\theta)$ in terms of which the LF is ***approximately data translated***. That is, the LF is nearly independent of the data **y** except for its locations. So Box and Tiao developed methods for obtaining parameter transformations in terms of which the LF is approximately data translated. These methods are based on approximation of the LF to a quadratic form that is approximately normally distributed. Then, the required transformations will be derived on the principle of variance-stabilizing parameterization, the principle that fulfills to the LF to be nearly data translated. The procedure introduced by Box and Tiao has slight differences according to the type of the p.d.f. as will be shown under the following two titles.

**I.** $p(\mathbf{y} \mid \theta)$ ***belongs to the exponential family:***

Consider $\mathbf{y}' = (y_1\, y_2 \ldots y_n)$ to be a random sample from a distribution $p(\mathbf{y} \mid \theta)$ that follows certain regularity conditions. If this distribution belongs to the exponential family, it could be written in the form

$$p(\mathbf{y} \mid \theta) = h(\mathbf{y})w(\theta)\exp[c(\theta)u(\mathbf{y})] \tag{2.4.4}$$

then the metric $\phi(\theta)$ in terms of which the LF is approximately data translated would be derived such that:

$$\phi \propto \int_{\Theta} K^{1/2}(\theta)\, d\theta, \tag{2.4.5}$$

where $\Theta$ is the parameter space on which $\theta$ is defined

$$K(\hat{\theta}) = \left( \frac{-1}{n} \frac{\partial^2 L}{\partial \theta^2} \right)_{\hat{\theta}}, \tag{2.4.6}$$

and $L$ is the logarithm of the LF, i.e., $L(\theta \mid \mathbf{y}) = \log l(\theta \mid \mathbf{y})$, and $\hat{\theta}$ is the maximum likelihood estimate (MLE) of $\theta$.

Applying Box and Tiao's procedure for selecting noninformative prior of $\theta$ involves taking the locally uniform prior distribution for $\phi$ as an approximately noninformative prior. This in turn implies that the corresponding noninformative prior for $\theta$ approximately follows the form:

$$p(\theta) \overset{locally}{\propto} \left| \frac{d\phi}{d\theta} \right| \overset{locally}{\propto} K^{1/2}(\theta). \tag{2.4.7}$$

**II.** $p(\mathbf{y} \mid \theta)$ ***does not belong to the exponential family (the general rule):***

Since $p(\mathbf{y} \mid \theta)$ is not often expressed in the form (2.4.2), Box and Tiao modified the above argument. Through this refinement, for large n, the quantity in (2.4.6) converges in probability to the expectation form as follow:

$$\zeta(\theta) = -E_{\mathbf{y} \mid \theta} \left[ \frac{\partial^2 \log p(\mathbf{y} \mid \theta)}{\partial \theta^2} \right], \tag{2.4.8}$$

which is the fisher's measure of information about $\theta$ in the sample $\mathbf{y}' = (y_1 \, y_2 \ldots y_n)$, which is generally defined as:

$$\zeta_n = E_{\mathbf{y} \mid \theta} \left( -\frac{\partial^2 L}{\partial \theta^2} \right), \tag{2.4.9}$$

while in case if **y** is a random sample, such a form would be expressed as:

$$\zeta_n(\theta) = n\zeta(\theta) \tag{2.4.10}$$

Consequently, arguing as before, the metric $\phi(\theta)$ for which the locally uniform is approximately noninformative and that makes the LF be approximately data translated will be, in this case, as follow:

$$\phi \propto \int_{\Theta} \zeta^{1/2}(\theta) d\theta. \tag{2.4.11}$$

Hence the corresponding noninformative prior for $\theta$ is approximately distributed as:

$$p(\theta) \overset{locally}{\propto} \left|\frac{d\phi}{d\theta}\right| \overset{locally}{\propto} \zeta^{1/2}(\theta), \tag{2.4.12}$$

which is the same as derived by Jeffreys (1961), on grounds of invariance, and that is discussed in (§2.3). It could easily be shown that, when the distribution $p(\mathbf{y}|\theta)$ is of the form (2.4.4), the forms in (2.4.6) and (2.4.8) are equivalent, i.e., $J(\theta) \equiv \zeta(\theta)$. Whence the prior distribution in (2.4.7) is identical to the prior in (2.4.12) and the latter form can be used generally.

**(b) Multi-parameter case:**

Box and Tiao extended their argument to include the multi-parameter problems, and discriminate between two cases. First, when a transformation that produces data translated LF is available in a sense introduced from the multi-parameter point of view. Second, when such transformation is unavailable, a rule is needed to at least produce LF regions of same size as shown by figure 2.4.3(b).

For further illustration, consider the distribution of data **y,** $p(\mathbf{y}|\boldsymbol{\theta})$, involves k parameters $\boldsymbol{\theta}' = (\theta_1\, \theta_2 \ldots \theta_k)$, the required noninformative prior for $\boldsymbol{\theta}$ could be found through one of the following rules:

**I.** ***A rule fulfills data translation LF:***

Transformation produces LF, as in figure 2.4.3(a), could be available. In this case, the data translated LF in terms of this transformation must be written in the form:

$$l(\boldsymbol{\theta}|\,\mathbf{y}) \propto g[\boldsymbol{\phi}(\boldsymbol{\theta}) - \mathbf{f}(\mathbf{y})], \tag{2.4.13}$$

where g(.) is a known function independent of **y**, $\boldsymbol{\phi}' = (\phi_1\, \phi_2 \ldots \phi_k)$, is a one-to-one transformation of $\boldsymbol{\theta}$, and $[\mathbf{f}(\mathbf{y})]' = [f_1(\mathbf{y})\, f_2(\mathbf{y}) \ldots f_k(\mathbf{y})]$ is a vector of k functions of **y**. The locally uniform distribution is taken as a noninformative prior for $\boldsymbol{\phi}$. The corresponding noninformative prior of $\boldsymbol{\theta}$ is then

$$p(\boldsymbol{\theta}) \propto |J|, \tag{2.4.14}$$

where

$$|J| = \begin{vmatrix} \frac{\partial \phi_1}{\partial \theta_1} & \cdots & \frac{\partial \phi_1}{\partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial \phi_k}{\partial \theta_1} & \cdots & \frac{\partial \phi_k}{\partial \theta_k} \end{vmatrix}.$$

So the matter now is to find a transformation $\boldsymbol{\phi}$ that produces LF form in (2.4.13) that is data translated, but this is not generally available. In the location-scale models, for example, the transformation that leads to LF form in (2.4.13) will not lead to LF contours as in figure 2.4.3(a). That is because, the existence of the scale parameter leads to a transformation that magnifies the volume of the LF, as a proportion of the scale parameter, along the location parameter space, which will be illustrated through examples (2.4.4, figure(2.4.4)).

**II.** ***A rule implies LF regions of same size:***

If the previous case is not available, Box and Tiao provide another less satisfactory method, as described by them, to obtain transformation that produces instead LF regions of same size. Such method depends on approximating the LF to the Normal distribution in a quadratic form and leads eventually to the following noninformative prior for $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}) \propto \left|\zeta_n(\boldsymbol{\theta})\right|^{1/2}, \tag{2.4.15}$$

where $\zeta_n = \underset{\mathbf{y}|\boldsymbol{\theta}}{E}\left(-\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}\right) = n \underset{\mathbf{y}|\boldsymbol{\theta}}{E}\left[-\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]$ for all $i = j = 1,\ldots,k$, is the information matrix about $\boldsymbol{\theta}$ associated with the sample. Therefor the prior form in (2.4.15) is identical to that obtained by Jeffreys' general rule in (2.3.6), but the later is derived on grounds of invariance. Box and Tiao have some interested remarks about applying this rule in some certain problems.

***Comments on Jeffreys' general rule:***

The preceding discussion has exposed to obstacles encounter application of Jeffreys' general rule. However, it is of interest to state here the difficulties of the application of the multi-parameter version of Jeffreys' rule which were introduced by Box and Tiao. They considered that this rule corresponds to less stringent and less

convincing transformation requirement on the LF than data translation. And under approximate Normality assumption, as well, the rule seeks a transformation that produces LF regions of the same size.

They further deem another difficulty associated with the application of this rule to the location-scale models where parameters of different types are considered simultaneously. Where applying this rule to these problems leads to inappropriate priors such as in (2.3.9) and (2.3.11). Therefore, in such problems, seeking transformations that produce LF regions of same size has not been appropriate. Thus they agree with Jeffreys in his assumption of independence between location and scale parameters. In this respects they said "Any prior idea one might have about the location of a distribution would usually not be much influenced by one's idea about the value of its scale parameter". Even though, they considered some problems whereas such assumption is inappropriate. In such cases they recommended applying some manipulation to data to assume independence.

It is of interest, to mention briefly the modification of Box and Tiao's methodology introduced by Kass (1990). He modified their procedure to cover more general location families. Kass extended their work to become group-theoretic. He also modified the concept of "approximate data translated LF" to produce a sharper local approximation.

### 2.4.3. **Examples:**

Box and Tiao's procedure will be illustrated in this subsection for the same models discussed earlier in §2.3.

**Example 2.4.1:** ***Binomial*** **($\theta$)**

According to this distribution an observation y (the number of success within n fixed number of trials) will be distributed as

$$p(y \mid \theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y}, \ \theta \in [0,1], y = 0,1,...,n.$$

It is evident that this density could be written in the form of (2.4.4) as follows:

$$p(y\,|\,\theta)=\binom{n}{y}(1-\theta)^n\exp[y\log(\tfrac{\theta}{1-\theta})].$$

Box and Tiao's procedure for selecting noninformative prior for $\theta$ will be applied for this problem. Start by finding $K(\hat{\theta})=\left(\frac{-1}{n}\frac{\partial^2 L}{\partial\theta^2}\right)_{\hat{\theta}}$, which can easily be proved to equal:

$$K(\hat{\theta})=\hat{\theta}^{-1}(1-\hat{\theta})^{-1}$$

.

Secondly, find the transformation $\phi(\theta)$ that produces approximately data translated LF through equation (2.4.5) as follow

$$\phi(\theta)\propto\int_0^1 K^{\frac{1}{2}}(\theta)\,d\theta\,.$$

It can easily be proved that,

$$\phi(\theta)\propto\sin^{-1}\sqrt{\theta}\,,$$

Then the locally uniform prior is taken as an approximate noninformative prior for $\phi(\theta)$. The corresponding approximate noninformative prior for $\theta$, as shown in (2.4.7), is

$$p(\theta)\overset{locally}{\propto}K^{\frac{1}{2}}(\theta),$$

then

$$p(\theta)\propto\theta^{-1/2}(1-\theta)^{-1/2}\,.$$

This is the same prior derived for the same problem using Jeffreys' rule in §2.3. It could easily be proved to get the same result but using the general rule in (2.4.12), as long as the sample density is expressible in the form of the exponential family.

**Example 2.4.2:** ***Normal*** **($\theta$)**

The sample distribution of such problem is expressed through the form

$$p(y\,|\,\theta)\propto e^{\frac{-1}{2\sigma^2}(y-\theta)^2},\quad y,\theta\in(-\infty,\infty),$$

which can be written in the form in (2.4.4), so both procedures of Box and Tiao in single space parameters will lead to the same result. So the general rule, (2.4.12), will be applied and leads to the result

$$\frac{\partial^2 L}{\partial\theta^2}\propto\text{constant.}$$

Then, by assuming that y is a random sample the quantity $\zeta(\theta)$ is calculated through forms (2.4.9) and (2.4.10) as follow

$$\zeta(\theta) \propto E_{y|\theta}\left(-\frac{\partial^2 L}{\partial \theta^2}\right),$$
$$\zeta(\theta) \propto \text{constant}.$$

Then, through form (2.4.11), the metric $\phi(\theta)$ that leads to LF which is approximately data translated is $\phi(\theta)=\theta$, which has the locally uniform distribution as an approximate noninformative prior. The corresponding noninformative prior of $\theta$, through form (2.4.12) is approximately

$$p(\theta) \overset{locally}{\propto} \zeta^{1/2}(\theta).$$

Then,

$$p(\theta) \overset{locally}{\propto} \text{Consatant}.$$

This is again the same result derived by Jeffreys' for the same model, see §2.3.

**Example 2.4.3:** ***Normal*** **($\sigma$)**

The random sample that is generated from this distribution has a density in the form

$$p(y|\sigma) \propto \sigma^{-1} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, \quad y,\theta \in (-\infty,\infty), \sigma \in (0,\infty).$$

This distribution belongs also to the exponential family, so the procedure of the general rule could be applied and leads to the result

$$\frac{\partial^2 L}{\partial \sigma^2} \propto \frac{n}{\sigma^2} - \frac{3[(n-1)s^2 + n(\theta-\bar{y})^2]}{\sigma^4}.$$

Then, the quantity $\zeta(\sigma)$ is calculated through forms (2.4.9) and (2.4.10) as follow

$$\zeta(\sigma) \propto E_{y|\sigma}\left(-\frac{\partial^2 L}{\partial \sigma^2}\right),$$
$$\propto \sigma^{-2}.$$

Then, again through form (2.4.11), the metric $\phi(\sigma)$ that leads to LF that is approximately data translated is

$$\phi \propto \int_0^\infty \zeta^{1/2}(\sigma)\, d\sigma$$
$$\phi(\sigma) \propto \log \sigma.$$

The corresponding noninformative prior of $\sigma$, through form (2.4.12) is approximately

$$p(\sigma) \overset{locally}{\propto} \zeta^{1/2}(\sigma),$$

then

$$p(\sigma) \overset{locally}{\propto} \sigma^{-1}.$$

This is again the same prior derived by Jeffreys for the same model, see §2.3, but on grounds of invariance.

**Example 2.4.4:** ***Normal*** **$(\theta, \sigma)$**

This is a type of location-scale distribution, where the random sample is generated from distribution of the form

$$p(\mathrm{y} \mid \theta, \sigma) \propto \sigma^{-1} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, \quad y, \theta \in (-\infty, \infty), \sigma \in (0, \infty).$$

The methodology of Box and Tiao for multi-parameter case that leads to the prior in (2.4.14) will be applied to this problem. In such method a transformation is sought such that the LF, in terms of which, could be written in the form in (2.4.13). The matter now is trying to rewrite the LF of this model in the form (2.4.13), and the required transformation could be hence automatically reached.

The LF is expressed by

$$\begin{aligned} l(\theta, \sigma \mid \mathrm{y}) &\propto \sigma^{-n} e^{\frac{-1}{2\sigma^2}\sum (y_i-\theta)^2}, \\ &\propto \sigma^{-n} \exp\left\{-\tfrac{n(\theta-\bar{y})^2}{2\sigma^2} - \tfrac{(n-1)s^2}{2\sigma^2}\right\}, \end{aligned}$$

Multiplying the last form by $s^n$, where multiplication of LF by constant leaves it unchanged, then

$$l(\theta, \sigma \mid \mathrm{y}) \propto \left(\frac{s}{\sigma}\right)^n \exp\left\{-\tfrac{n(\theta-\bar{y})^2}{2s^2}\left(\frac{s^2}{\sigma^2}\right) - \tfrac{(n-1)s^2}{2\sigma^2}\right\},$$

which can be written as

$$\begin{aligned} l(\theta, \sigma \mid \mathrm{y}) &\propto \left(\frac{\sigma}{s}\right)^{-n} \exp\left\{-\tfrac{n}{2}\left(\frac{\theta-\bar{y}}{s}\right)^2\left(\frac{\sigma}{s}\right)^{-2}\right\} \exp\left\{-\tfrac{(n-1)}{2}\left(\frac{\sigma}{s}\right)^{-2}\right\} \\ &\propto \exp\left\{-\tfrac{n}{2}\left(\tfrac{\theta-\bar{y}}{s}\right)^2 \exp[-2\log\left(\frac{\sigma}{s}\right)]\right\} \times \exp\left\{-n\log\left(\frac{\sigma}{s}\right) - \left(\tfrac{n-1}{2}\right)\exp[-2\log\left(\frac{\sigma}{s}\right)]\right\}. \end{aligned}$$

Eventually, the LF can be given by the following form:

$$\begin{aligned} l(\theta, \sigma \mid \mathrm{y}) &\propto \exp\left\{-\tfrac{n}{2}\left(\tfrac{\theta-\bar{y}}{s}\right)^2 \exp[-2(\log\sigma - \log s)]\right\} \\ &\times \exp\left\{-n(\log\sigma - \log s) - \left(\tfrac{n-1}{2}\right)\exp[-2(\log\sigma - \log s)]\right\}. \end{aligned}$$

This last form could be considered as a translation to the form in (2.4.13) such that $\boldsymbol{\phi} \propto \begin{bmatrix} \theta \\ \log\sigma \end{bmatrix}$ and $\mathbf{f}(\mathbf{y}) \propto \begin{bmatrix} \bar{y} \\ \log s \end{bmatrix}$. Then take the locally uniform distribution as a

noninformative prior for $\phi$ and according to the form (2.4.14) the corresponding noninformative prior for both $\theta$ and $\sigma$ will be

$$p(\theta,\sigma) \propto |J|,$$

then

$$p(\theta,\sigma) \propto \sigma^{-1}.$$

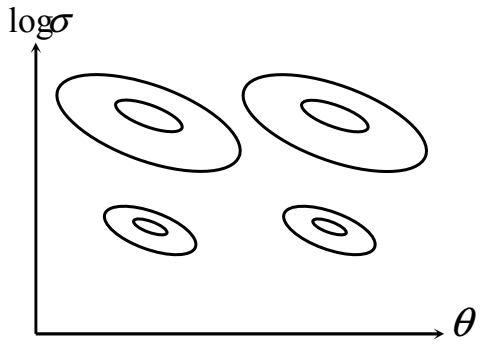
The same result could be reached if one applies Jeffreys' general rule but under assuming independence between location and scale parameters, in the sense as mentioned in §2.3. However, Jeffreys' general rule when independence assumption is not incorporated, will lead to the inappropriate result

$$p(\theta,\sigma) \propto \sigma^{-2}.$$

As mentioned earlier for the location-scale models the transformation leads to LF form in (2.4.13) will not lead to LF contours as in figure 2.4.3(a). Figure (2.4.4) shows the contours of LF of the transformation taken through the preceding methodology.

**Figure 2.4.4**

**Normal $(\theta,\sigma)$: contours of LF of $(\theta, \log\sigma)$ for different data sets**




It is evident that this transformation leads to a bit magnification to the volume of the LF, as a proportion of the scale parameter, along the location parameter space.

**Example 2.4.5:** ***k-Normal*** **$(\theta,\sigma)$**

For another example to the multi-parameter problem, the distribution of k independent normal population has been provided. Assuming for simplicity that the random samples are same sized, say r, the LF of the k-vector of random samples will be in the form:

$$l(\boldsymbol{\theta},\sigma|\,\mathbf{y}) \propto \sigma^{-n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij}-\theta_i)^2}, \quad y_{ij},\theta_i \in (-\infty,\infty),\ j=1,2,...,r,\quad i=1,2,...,k,$$
$$\sigma \in (0,\infty),\ \text{and}\ n=rk.$$

Similarly as shown in the previous example, one could derive the noninformative prior of $\boldsymbol{\theta}$ and $\sigma$ according to Box and Tiao methodology. Again the idea is trying to rewrite the LF of this model in the form (2.4.13), hence the required transformation could be reached automatically.

The preceding LF could be expressed as

$$l(\boldsymbol{\theta},\sigma|\,\mathbf{y}) \propto \sigma^{-n} \exp\left\{-\frac{(n-k)s^2}{2\sigma^2} - \frac{r(\boldsymbol{\theta}-\bar{\mathbf{y}})'(\boldsymbol{\theta}-\bar{\mathbf{y}})}{2\sigma^2}\right\},$$

where

$$\sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij}-\theta_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}[(y_{ij}-\bar{y}_i)-(\theta_i-\bar{y}_i)]^2,$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{k} r(\bar{y}_i-\theta_i)^2.$$

where $\bar{y}_i$ is the arithmetic mean within the sample *i*, where *i=1,2,...,*k.

Multiplying the last form of the LF by $s^n$

$$l(\boldsymbol{\theta},\sigma|\,\mathbf{y}) \propto \left(\frac{s}{\sigma}\right)^n \exp\left\{-\frac{(n-k)}{2}\left(\frac{s}{\sigma}\right)^2 - \left(\frac{s}{\sigma}\right)^2 \frac{r(\boldsymbol{\theta}-\bar{\mathbf{y}})'(\boldsymbol{\theta}-\bar{\mathbf{y}})}{2s^2}\right\},$$

which can easily be rewritten as

$$l(\boldsymbol{\theta},\sigma|\,\mathbf{y}) \propto \exp\left\{-\frac{r}{2}\frac{(\boldsymbol{\theta}-\bar{\mathbf{y}})'(\boldsymbol{\theta}-\bar{\mathbf{y}})}{s^2}\ \exp[-2(\log\sigma-\log s)]\right\}$$
$$\times \exp\left\{-n(\log\sigma-\log s) - \left(\frac{n-k}{2}\right)\exp[-2(\log\sigma-\log s)]\right\}$$

Once again, this last form corresponds to the one in (2.4.13) such that $\boldsymbol{\phi} \propto \begin{bmatrix}\boldsymbol{\theta}\\ \log\sigma\end{bmatrix}$ and $\mathbf{f}(\mathbf{y}) \propto \begin{bmatrix}\bar{\mathbf{y}}\\ \log s\end{bmatrix}$. Then, the locally uniform distribution will be taken as a noninformative prior for $\boldsymbol{\phi}$ and according to the form (2.4.14) the corresponding noninformative prior for $\boldsymbol{\theta}$ will be

$$p(\boldsymbol{\theta},\sigma) \propto |J|,$$

then

$$p(\boldsymbol{\theta},\sigma) \propto \sigma^{-1}.$$

Remember that, Jeffreys reaches the same result but under independence assumption between location and scale parameters, in the sense mentioned in §2.3. However, Jeffreys' general rule when independence assumption is not involved, will lead to an inappropriate result, presented previously by equation (2.3.12).

## 2.5. Maximal Data Information Prior(MDIP)

### 2.5.1. Introduction

Zellner (1971) sustained using "locally uniform" proper priors when an investigator knows something about parameters such as their range, the experimental design and properties of LF. Information of this sort may be available in a perspective that was demonstrated in §2.4. This is usually not the situation in practice. Therefore, when such information is not available, Zellner emphasized that it usually makes very little practical difference whether locally uniform prior or Jeffreys' improper prior is used. In these regards, Zellner (1971) developed a framework that is based upon informational considerations, to derive a noninformative prior that formulates the case of "knowing little" or "ignorance".

Since learning from data and experience is an important activity in science, Zellner's main idea was to reach a prior that leads to a posterior distribution reflecting mainly the information in a given sample or adds little information to the sample information. Thus, his objective was to obtain a prior that maximizes the difference between the average information in the LF, and the information in the prior. A solution to this optimization problem is a "***Maximal Data Information Prior (MDIP)***" or a "***Minimal Information Prior***", as Zellner (1971) called it.

**Definitions**

***Single parameter case***

In order to illustrate this concept in case of one parameter, say $\theta$, notice that, the basic idea underlying MDIPs is that they maximize the gain in the information resulted from the sample. Zellner (1971) introduced the following quantity as a powerful criterion of such gain

$$G = \bar{I}_y - \int_{R_\theta} p(\theta) \ln p(\theta) d\theta \,, \tag{2.5.1}$$

where $p(\theta)$ is the required prior distribution of $\theta$ which is defined on the parameter space $R_\theta$. Whereas, $\bar{I}_y$ is called the prior average information associated with an observation *y* and calculated as

$$\bar{I}_y = \int_{R_\theta} I_y(\theta)\, p(\theta) d\theta, \tag{2.5.2}$$

where $I_y(\theta)$ is defined to be a measure of information in the sample p.d.f. $p(y|\theta)$, and computed as

$$I_y(\theta) = \int_{R_y} p(y|\,\theta)\, \ln p(y|\,\theta) dy, \tag{2.5.3}$$

such that $R_y$ is the sample space on which the sample p.d.f. is defined.

As seen through above relations *G* is just the difference between two information measures. The first relating the data and the second relating the prior.

Zellner (1971) hence defined MDIP or the *minimal information prior* to be the one that maximize *G* for a given $p(y|\theta)$.

***Multi-parameter case**:*

For data p.d.f.'s involving more than one parameter, say a vector of k parameters $\boldsymbol{\theta}' = (\theta_1\, \theta_2 \ldots \theta_k)$, which is defined on the parameter space $R_{\boldsymbol{\theta}}$, *G* will be defined as follows,

$$G = \bar{I}_y - \int_{R_{\boldsymbol{\theta}}} p(\boldsymbol{\theta}) \ln p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.5.4}$$

where,

$$\bar{I}_y = \int_{R_{\boldsymbol{\theta}}} I_y(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

and

$$I_y(\boldsymbol{\theta}) = \int_{R_y} p(y|\,\boldsymbol{\theta})\, \ln p(y|\,\boldsymbol{\theta}) dy. \tag{2.5.5}$$

Moreover, the same definition could be introduced for a random vector of observations $\mathbf{y}' = (y_1\, y_2 \ldots y_n)$ of order n with j.p.d.f. $p(\mathbf{y}|\,\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the k-vector of parameters.

Zellner admits using another definition of a minimal information prior by employing any other measures of information. Much work has been done to derive priors by maximizing the information provided by an experiment, see e.g., Good (1956), Lindley (1956) and Soofi (1994), via calculus of variations techniques and getting no clear-cut analytical result because they provide intractable solutions. Zellner, however, altered the criterion to the form of *G* in (2.5.1), which is considered to be relatively easy to produce, Zellner (1996).

Zellner (1977) provides the same procedure with further application to many problems. Many MDIPs have been developed and further properties to this procedure have been established in many papers in literature such as Sinha and Zellner (1990), Zellner (1991) and Zellner and Min (1993).

### 2.5.2. Derivation

The optimization of G in (2.5.1) with respect to the choice of $p(\theta)$ subject to side conditions is apparently just a standard calculus of variations problem. Zellner (1971) considered the side condition is that the prior $p(\theta)$ is proper. That is

$$\int_{R_\theta} p(\theta)d\theta = 1. \tag{2.5.6}$$

Zellner regarded this side condition provided that $R_\theta$, the region on which $\theta$ is defined, may be, very large but it must be at least a compact region.

The solution to the problem of maximizing (2.5.1) subject to (2.5.6), denoted by $p^*(\theta)$, has been derived by Zellner (1977) to be

$$p^*(\theta) \propto \exp\left\{I_y(\theta)\right\}, \quad \theta \subset R_\theta, \tag{2.5.7}$$

such that $I_y(\theta)$, given in (2.5.3), is the information in the data density $p(y|\theta)$.

Zellner (1996) noticed that if $I_y(\theta)$ is constant, independent of $\theta$, then the MDIP p.d.f. is the uniform distribution. He also pointed out that the rule in (2.5.7) is implemented relatively easily for many problems.

Similarly, expressions as in (2.5.7) could easily be produced for multi-parameter problems. In this respects, the MDIP that maximizes the functional criterion *G* in (2.5.4), as shown by Zellner (1977), has the following form

$$p^*(\boldsymbol{\theta}) \propto \exp\left\{I_y(\boldsymbol{\theta})\right\}, \quad \boldsymbol{\theta} \subset R_{\boldsymbol{\theta}},$$

where the terms $\boldsymbol{\theta}$, $I_y(\boldsymbol{\theta})$, and $R_{\boldsymbol{\theta}}$ are defined above. The same result could be obtained for models that contain random vector of observations, for further details see Zellner (1977).

### 2.5.3. Properties

The use of MDIP approach provides an explicit tool for the problem of selecting noninformative prior distributions. This approach is easy to implement since no asymptotic approximations are involved. Zellner (1996) appends several comparison results provided by alternative procedures for producing noninformative priors, indicates that MDIPs are relatively easy to produce. Besides that, they have reasonable properties which make them helpful to researchers and decision-makers in formulating priors. In this regard, Zellner (1996) epitomized these properties when he said "The MDIP approach allows one to derive diffuse and informative priors that are invariant with respect to relevant transformations is indeed fortunate.". Each of these features mentioned in the above citation shall be considered in details.

***Informational considerations***

Zellner provides an illuminating discussion of information processing rules. These rules derived by optimizing some informational criterion, are 100% efficient (Golan, 2002). This optimization process resulted in MDIPs. Thus it is intuitive to review some general informational features of the MDIP approach.

1. Entropy measure view:

Zellner deemed that there is no way to answer the question about the form of a distribution to express the ignorance without using a measure of information. Therefore, he suggested a measure that is used by many others including Shannon (1948) which is the negative entropy, denoted by -*H*. This measure is used to express the information in a p.d.f. For example the negative entropy of the prior distribution $p(\theta)$, relative to a uniform measure, is given by,

$$-H = \int_{R_\theta} p(\theta) \ln p(\theta) d\theta. \qquad (2.5.8)$$

This last expression is just the second term of the quantity in (2.5.1). Zellner (1971), however, started with this concept to construct his functional criterion *G*. He used the entropy measure to signify the information in the joint density $p(y,\theta)$ which is defined as follow,

$$-H = \int_{R_y}\int_{R_\theta} p(y,\theta)\ln p(y,\theta)d\theta dy. \tag{2.5.9}$$

On using $p(y,\theta) = p(y|\theta)p(\theta)$, the last equation could be passed through the following relations,

$$\begin{aligned} -H &= \int_{R_y}\int_{R_\theta} p(y|\theta)p(\theta)\ln\left[p(y|\theta)p(\theta)\right]d\theta dy, \\ &= \int_{R_y}\int_{R_\theta} p(y|\theta)p(\theta)\left[\ln p(y|\theta) + \ln p(\theta)\right]d\theta dy, \\ &= \int_{R_\theta} p(\theta)\int_{R_y} p(y|\theta)\ln p(y|\theta)\,dy d\theta + \int_{R_\theta} p(\theta)\ln p(\theta)\int_{R_y} p(y|\theta)dy d\theta. \end{aligned}$$

By substituting from both equations (2.5.2) and (2.5.3) in the last quantity, it could be written as

$$-H = \bar{I}_y + \int_{R_\theta} p(\theta)\ln p(\theta)d\theta. \tag{2.5.10}$$

As seen from (2.5.10), which makes up the total information in the joint density $p(y,\theta)$, that it breaks up into two parts. The first is the prior average information in the data density and the second is the information in the prior density.

According to that, Zellner conveniently chooses *G* to be equivalent to the difference of the two terms on the R.H.S.of (2.5.10).

Zellner suggested using another informational measurements to express the criterion *G* rather than the negative entropy based on uniform measure, as mentioned by (2.5.8) or (2.5.9). In this respect he recommended employing the negative entropy defined on other measures rather than the uniform measure to construct *G*, see Zellner (1996) for further details.

Another interpretation of MDIPs according to their entropy view, is mentioned by Jaynes (1982), is that MDIPs are the p.d.f.'s that maximize the entropy associated with

a prior distribution subject to the side condition that the average entropy in the data p.d.f., $p(y|\theta)$, be constant.

2. The average log-ratio of LF to the prior view:

A second interpretation of the criterion *G* from the informational point of view is given in Zellner (1977). This interpretation could be attained through the following steps:

Substituting from equations (2.5.2) and (2.5.3) in (2.5.1), *G* could be written as

$$G = \int_{R_\theta} p(\theta) \int_{R_y} p(y|\theta) \ln p(y|\theta)\, dy d\theta - \int_{R_\theta} p(\theta) \ln p(\theta) d\theta,$$

It can be proved that

$$G = \int_{R_\theta} \int_{R_y} \left[\ln p(y|\theta) - \ln p(\theta)\right] p(y,\theta) dy d\theta.$$

Since $l(\theta|y) \equiv p(y|\theta)$ is the likelihood function (LF) and given the last form, *G* can be expressed as

$$G = \int_{R_\theta} \int_{R_y} \ln\left[\frac{l(\theta|y)}{p(\theta)}\right] p(y,\theta) dy d\theta.$$

According to this last view of *G*, it can be interpreted as the average log-ratio of the LF to the prior p.d.f. Hence, by maximizing *G* by choice of $p(\theta)$, the average log-ratio of the LF to the prior will be made as large as possible.

Given also this view, the forms of the MDIPs will depend on properties of the LF's or on the design of an experiment. This seems natural, since the purpose of the MDIPs is to allow the information provided by an experiment to be featured (Zellner, 1977).

At last, having the LF featured in this fashion is an important aspect of the MDIP approach for selecting noninformative priors (Zellner, 1996).

***Invariance considerations***

Zellner (1977) proposed two important theorems that provides some confining invariance properties of the MDIPs that are relating only to linear transformation of the parameters. According to these theorems:

1. MDIPs are invariant with respect to changes in the unit of measurements. This property of invariance is particularly termed as S-labeling invariance, see subsection 2.3.3. for more discussion to such property.
2. MDIPs are generally invariant with respect to linear transformations of the parameters and observations.

Kass and Wasserman (1996) pointed out that MDIPs are not generally parameterization invariant, specifically they are not Ω-labeling invariante as Hartigan (1964) called it. However, Zellner (1991) argued that invariance under specific classes of re-parameterization can be achieved by adding the appropriate constraints. That is to introduce the invariance conditions as side conditions in the optimization process of the functional criterion *G*.

For more clarification to such a point, consider the case of *m* one-to-one transformations $\eta_i = h_i(\theta),\ i = 1,2,...,m$. Zellner (1991) suggested obtaining MDIP by maximizing instead the following quantity

$$G = \bar{I}_y - \int_{R_\theta} p(\theta)\ln p(\theta)d\theta + \sum_{i=1}^{m}\left( \int_{R_{\eta_i}} p(\theta) I_y(\eta_i)d\eta_i - \int_{R_{\eta_i}} p(\eta_i)\ln p(\eta_i)d\eta_i \right),$$

where $I_y(\eta_i) = \int_{R_y} p(y|\eta_i)\ln p(y|\eta_i)\,dy$.

The optimization process for such quantity will be hold subject to

$$p(\theta)d\theta = p(\eta_i)d\eta_i, \quad \forall i = 1,2,...,m.$$

The solution of this optimization problem has been produced by Zellner (1991) and given by

$$p^*(\theta) = \exp\left\{ I_y(\theta) + \sum_{i=1}^{m} \frac{\ln\left|h_i'(\theta)\right|}{m+1} \right\}.$$

This resulted prior then has the desired invariance properties over the given transformations. Some of interesting examples to such transformations are the reciprocal and power transformations. It is noteworthy that imposition of invariance conditions changes MDIPs that don't incorporate them (Zellner, 1996).

By deeming the above algorithm, Zellner avoids being restricted by broad invariance conditions of the so-called Ω-labeling invariance. As mentioned by Zellner (1977), the objective of having a posterior distribution that reflects mainly the information in the data distribution can be achieved, but, for a particular parameterization. In other words, different investigators will obtain the same posterior distributions given that they use the MDIP procedure to generate priors for any given parameterization.

The problem of achieving invariance to a wide class of re-parameterizations is a problematic issue that has received considerable attention in the literature and must be considered. On this respect Berger (1985) comments *"The major problem with invariance concerns the amount of invariance that can be used."*. Rao (1987), however, discussed degrees of invariance and states that the choice of metric naturally depends on a particular problem under investigation and invariance may or may not be relevant.

***A tool to produce "Informative priors"***

The MDIP approach is designed specifically to provide rules for selecting noninformative priors. One of the greatest contributions of MDIP approach to Bayesian inference is that it can further be employed to produce *informative prior* distributions.

This can be done by incorporating the available *prior information* as side conditions in the process of optimizing *G* in (2.5.1). The prior distribution resulted from this maximization process is informative. For instance, as shown in Zellner (1996), the side condition may include the prior to be proper, that is, in (2.5.6), besides the additional moment conditions with the $m_i's$ given by

$$m_i = \int_{R_\theta} \theta^i p(\theta)\, d\theta, \quad i = 1,2,...,m. \tag{2.5.11}$$

Zellner showed that the prior that maximizes G in (2.5.1) subject to (2.5.6) and (2.5.11) is given by

$$p^*(\theta) = \exp\left\{ I_y(\theta) + \lambda_1\theta + \lambda_2\theta^2 + ... + \lambda_q\theta^q \right\},$$

where $\lambda_i's$ are Lagrange multipliers. A reader may refer to Zellner and Highfield (1988) for a procedure for computing values of these multipliers.

Another type of information could be embodied, as a side condition, may be restricted on the ranges of the parameters. Moreover, Zellner (1996) extended his technique to involve prior information related to prior fractiles. He involved it as a side condition in his algorithm and derived informative priors.

***Affinities with Jeffreys' general rule:***

As most of researches within noninformative prior selection could possibly be traced back directly or indirectly to Jeffreys' prior, it seems natural to inquire about the existence of some affinities between MDIP approach and Jeffreys' prior.

Zellner (1971) considered the asymptotic form of the criterion *G* as follow

$$G_a = \int_{R_\theta} p(\theta) \ln \sqrt{n\left|\mathrm{Inf}_\theta\right|}\, d\theta - \int_{R_\theta} p(\theta) \ln p(\theta) d\theta,$$

where n is the number of independent drawings from $p(y|\theta)$ and $\mathrm{Inf}_\theta$ is the Fisher information matrix defined in (2.3.7). By maximizing $G_a$ subject to (2.5.6) Zellner obtained the following prior

$$p(\theta) = \left|\mathrm{Inf}_\theta\right|^{1/2},$$

which is just Jeffreys' invariant prior given by his general rule in (2.3.6).

Thus from the asymptotic form of *G*, Jeffreys' prior is MDIP. In this respect, it must be recognized, however, that Jeffreys' prior is not always minimal information prior since it does not always maximize *G*. It is convenient to notice some situations in which $G_a$ is not maximized by Jeffreys' prior. This is so when one considers models such as location-scale models where parameters of different types are involved simultaneously, and models of high dimension as well.

At the other extreme, minimal information priors do not have generally the invariance property of Jeffreys' prior as just mentioned above. Zellner (1991) refined his approach to derive MDIP that meets the invariance requirement, but for particular relevant parameterization. Zellner (1971), however, deemed that investigators using

different parameterizations can get compatible (or consistent) results if they adopt the convention of using MDIPs for any given parameterization when they know little about the values of the parameters. That is, the invariance property has not yet been considered as an urgent necessity against getting MDIP.

### 2.5.4. Examples

MDIP approach is applicable to a very wide range of data densities, as shown in Zellner (1977), where many MDIPs for a number of univariate and multivariate data densities are presented. He pointed out, e.g., that for location-scale data densities, the resulted MDIPs are in accordance with usual prescriptions for diffuse or noninformative priors that are in widespread use. Zellner (1996), moreover, discussed deriving MDIPs for parameters of several frequently employed models such as linear models, e.g., General Linear Model (GLM) and Autoregressive Models (AR). He also applied his technique for hierarchical models hyperparameters and for common parameters in different data densities as well.

Throughout this subsection, applications of MDIP approach will be demonstrated for those densities presented in sections 2.3 and 2.4.

**Example 2.5.1:** ***Binomial(θ)***

The well-known Binomial process will be considered but for a single observation, for simplicity, that is the *Bernoulli* process. The probability mass function of such a process, $p(y|\theta)$, is in the form

$$p(y|\theta) = \theta^{y}(1-\theta)^{1-y}, \quad 0 \le \theta \le 1 \text{ and } y = 0,1$$

where $y$ is the number successes and $\theta$ is the probability of success.

Evolving the MDIP in (2.5.7) for the binomial parameter $\theta$ requires computing the quantity $I_y(\theta)$ in (2.5.3), which represent the information in the data mass function $p(y|\theta)$. This term can be computed as follows

$$I_y(\theta) = \sum_{y=0}^{1} p(y|\theta) \ln p(y|\theta),$$

$$= \sum_{y=0}^{1} \theta^{y}(1-\theta)^{1-y}\left[y \ln\theta + (1-y)\ln(1-\theta)\right],$$

$$= \theta \ln \theta + (1-\theta)\ln(1-\theta),$$

$$= \ln\left[\theta^{\theta}(1-\theta)^{(1-\theta)}\right].$$

Thus the MDIP for $\theta$, by using (2.5.7), is

$$p^*(\theta) \propto \theta^{\theta}(1-\theta)^{(1-\theta)}, \quad 0 \le \theta \le 1 \tag{2.5.12}$$

This prior density contrasts sharply with Jeffreys' prior for Binomial parameter that is in (2.3.8), which is the Beta$(½, ½)$. An interesting comparison between these two prior densities will be provided at the end of this chapter.

**Example 2.5.2:** ***Normal*** **($\theta$)**

When the data density belongs to the normal distribution with known variance $\sigma^2$, noted to belong to $(0,\infty)$. Such distribution is a type of location densities and has the form

$$p(y|\theta) = \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, \quad y \in (-\infty,\infty)$$

where $\theta$ is the unknown location parameter that is defined on the parameter space $(-\infty,\infty)$.

The data density information measure $I_y(\theta)$, according to (2.5.3), is given by

$$I_y(\theta) = \int_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2} \ln\left(\tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}\right) dy,$$

$$= \left(\ln\left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)\int_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2} dy\right) + \left(\tfrac{-1}{2\sigma^2}\int_{-\infty}^{\infty} (y-\theta)^2 \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2} dy\right) \tag{2.5.13}$$

Since the integral in the first term on the right side of (2.5.13) is an integral all over the space of the normal p.d.f., it will hence give unity. Whereas, integral in the second term is just $E(y-\theta)^2$, which gives the variance $\sigma^2$. Hence the form in (2.5.13) can be reduced to

$$I_y(\theta) = \ln\left(\tfrac{1}{\sqrt{2\pi}\sigma}\right) + \left(\tfrac{-1}{2\sigma^2}(\sigma^2)\right).$$

$$= -\ln\left(\sqrt{2\pi}\right) - \ln\sigma - \tfrac{1}{2}$$

Finally the measure $I_y(\theta)$ will be given by

$$I_y(\theta) = -\tfrac{1}{2}[\ln(2\pi)+1] - \ln\sigma \tag{2.5.14}$$

It can obviously be seen that the right side of (2.5.14) is constant, independent of $\theta$, then the data density information measure of the normal distribution with unknown mean is as follow

$$I_y(\theta) = \text{constant}.$$

Then the MDIP of the mean parameter $\theta$ can be produced using (2.5.7) to be as follows

$$p^*(\theta) \propto \text{constant}.$$

This prior distribution is in accordance with the Jeffreys' prior of $\theta$ for the same problem, that is, Jeffreys' first rule in (2.3.1).

**Example 2.5.3:** ***Normal*** **($\sigma$)**

Regarding the normal distribution with unknown variance $\sigma^2$, the form of the p.d.f. is given by,

$$p(y|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, \quad y \in (-\infty, \infty)$$

where $\theta$ and $\sigma$ are defined through the previous example.

To illustrate the derivation of the MDIP of $\sigma$ in such problem, the measure of the information in the density $p(y|\sigma)$, which is $I_y(\sigma)$ should be calculated using (2.5.3). This can be done easily by following up the previous steps of computing $I_y(\theta)$ that have been shown in example 2.5.2. Eventually, the measure $I_y(\sigma)$ can easily be proved to have the same form as in (2.5.14). That is it will be given by,

$$I_y(\sigma) = -\tfrac{1}{2}[\ln(2\pi)+1] - \ln\sigma,$$

which is equivalent to,

$$I_y(\sigma) = \text{constant} - \ln\sigma. \tag{2.5.15}$$

Thus, the MDIP of $\sigma$, based on (2.5.7), will be given by

$$p^*(\sigma) = \exp\{I_y(\sigma)\},$$

$$= \exp\{\text{constant} + \ln\sigma^{-1}\},$$

Then, the MDIP of $\sigma$, finally, has the form

$$p^*(\sigma) \propto \sigma^{-1}.$$

Again this prior distribution is equivalent to Jeffreys' prior of the normal standard deviation $\sigma$, which is Jeffreys' second rule that given by (2.3.3).

**Example 2.5.4:** ***Normal*** **$(\theta, \sigma)$**

It is essential to provide a data density that belongs to location-scale densities. Therefore, the normal distribution with mean and variance are both unknown is considered. The form of the p.d.f. is given by,

$$p(y|\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y-\theta)^2}, \quad y \in (-\infty, \infty),$$

where, the parameter space is the same as mentioned in the above two examples.

The information in this density, measured by $I_y(\theta,\sigma)$ using (2.5.5), can be shown to be equivalent to the right side of (2.5.14). It can be accordingly reduced to the same form as in (2.5.15). That is

$$I_y(\theta,\sigma) = \text{constant} - \ln\sigma .$$

As shown throughout the previous example, the prior yielded by MDIP approach has consequently the following form

$$p^*(\theta,\sigma) \propto \sigma^{-1}. \tag{2.5.16}$$

This result corresponds to Box-Tiao *"data translation"* rule for producing noninformative prior presented in §2.4, both in form and being defined over a finite range of parameter space and, thus, being proper density. But this MDIP is in contrast with Jeffreys' general rule, in (2.3.6), which produced the prior form in (2.3.9) that deemed by Jeffreys to be unsatisfactory because it involves adding unwanted information. Jeffreys modified that prior by assuming independence between the location and scale parameters and applying his general rule separately to each parameter and eventually reach the prior in (2.3.10), the prior he employed in practice, which corresponds the form in (2.5.16).

**Example 2.5.5:** ***k-Normal*** $(\boldsymbol{\theta}, \sigma)$

Another example for a multi-parameter case with larger dimension is the distribution of k-normal independent populations with k-vector of unknown means $\boldsymbol{\theta}' = (\theta_1\ \theta_2\ \ldots\ \theta_\kappa)$ defined on the parameter space $\Omega \in R^k$, and unknown common standard deviation $\sigma$ defined over $(0,\infty)$, will be presented here. In such distribution there are k independent random samples $\mathbf{y}_i$'s , each of size $r$ defined over the sample space $S \subset R^r$ and each also generated from *Normal*$(\theta_i, \sigma)$, where $i = 1,2,\ldots,k$. The j.p.d.f. of the k-vector of random samples $\mathbf{y}' = (\mathrm{y}_1\ \mathrm{y}_2\ \ldots\ \mathrm{y}_k)$ will be in the form,

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma) = (\sqrt{2\pi}\sigma)^{-n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{r}(\mathrm{y}_{ij}-\theta_i)^2}, \quad \text{where} \quad n = rk.$$

The information in the data density, measured by $I_{\mathbf{y}}(\boldsymbol{\theta}, \sigma)$, are given by

$$I_{\mathbf{y}}(\boldsymbol{\theta}, \sigma) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma) \ln\left[p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma)\right] d\mathrm{y}_{11} d\mathrm{y}_{12} \ldots d\mathrm{y}_{rk}$$

Then

$$I_{\mathbf{y}}(\boldsymbol{\theta}, \sigma) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{r}(\mathrm{y}_{ij}-\theta_i)^2} \ln\left( (\sqrt{2\pi}\sigma)^{-n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{r}(\mathrm{y}_{ij}-\theta_i)^2} \right) d\mathrm{y}_{11} d\mathrm{y}_{12} \ldots d\mathrm{y}_{rk}$$

$$= \ln\left[(\sqrt{2\pi}\sigma)^{-n}\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{r}(\mathrm{y}_{ij}-\theta_i)^2} d\mathrm{y}_{11} d\mathrm{y}_{12} \ldots d\mathrm{y}_{rk}$$

$$+ \left(\frac{-1}{2\sigma^2}\right) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-n} \sum_{i=1}^{k}\sum_{j=1}^{r}(\mathrm{y}_{ij}-\theta_i)^2 e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{r}(\mathrm{y}_{ij}-\theta_i)^2} d\mathrm{y}_{11} d\mathrm{y}_{12} \ldots d\mathrm{y}_{rk}$$

The right side of the last equation can easily be simplified to the following form,

$$= \ln\left[(\sqrt{2\pi}\sigma)^{-n}\right] \prod_{i=1}^{k}\prod_{j=1}^{r}\int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-1} e^{\frac{-1}{2\sigma^2}(\mathrm{y}_{ij}-\theta_i)^2} d\mathrm{y}_{ij}$$

$$+ \left(\frac{-1}{2\sigma^2}\right) \sum_{l=1}^{n}\prod_{i=1}^{k}\prod_{j=1}^{r}\int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-1} S(m,l)\, e^{\frac{-1}{2\sigma^2}(\mathrm{y}_{ij}-\theta_i)^2} d\mathrm{y}_{ij}, \text{ where } m = (i-1)r + j,$$

and $S(m,l)$ is an indicator function is defined as

$$S(m,l) = \begin{cases} (y_{ij} - \theta_i)^2 & m = l \\ 1 & m \neq l \end{cases}$$

It can be seen, for the first term of the right side of last form of $I_{\mathbf{y}}(\boldsymbol{\theta}, \sigma)$, that each integral within the n multiplied terms is unity. Whereas, the amount of the second term, is a summation of n terms each is a multiplication of n terms of integral. To evaluate this amount, consider for example the first term in the summation where $l = 1$ that is given by,

$$\int_{-\infty}^{\infty}(\sqrt{2\pi}\sigma)^{-1}(y_{11}-\theta_1)^2 e^{\frac{-1}{2\sigma^2}(y_{11}-\theta_1)^2} dy_{11}\Big|_{\ni (m=l=1)} \times \prod_{i=1}^{k}\prod_{j=2}^{r}\int_{-\infty}^{\infty}(\sqrt{2\pi}\sigma)^{-1} e^{\frac{-1}{2\sigma^2}(y_{ij}-\theta_i)^2} dy_{ij}\Big|_{\ni (m\neq l)}$$
$$= \sigma^2 \times \prod_{i=1}^{k}\prod_{j=2}^{r}(1)$$
$$= \sigma^2$$

This result could be achieved for all the remaining (n-1) summed terms. Then, the information measure $I_{\mathbf{y}}(\boldsymbol{\theta},\sigma)$ can be eventually reduced to,

$$\begin{aligned} I_{\mathbf{y}}(\boldsymbol{\theta},\sigma) &= \ln(\sqrt{2\pi}\sigma)^{-n} \prod_{i=1}^{k}\prod_{j=1}^{r}(1) + \left(\frac{-1}{2\sigma^2}\right)\sum_{l=1}^{n}\sigma^2 \\ &= \ln(\sqrt{2\pi}\sigma)^{-n} + \frac{-n}{2} \\ &= \frac{-n}{2}\ln(2\pi+1) + \ln\sigma^{-n}. \end{aligned}$$

Using the generalized form of (2.5.7) in multidimensional case, the prior yielded by MDIP approach has consequently the following form

$$p(\boldsymbol{\theta},\sigma) \propto \sigma^{-n}.$$

It is of quite interest to consider such example as it was previously concerned in section 2.3 (see example 2.3.5). The purpose is to compare the results when applying MDIP to the k-Normal distribution with those obtained when applying Jeffreys' prior. In such case, the sample p.d.f. is in the form,

$$p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma) \propto \sigma^{-k} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(y_i-\theta_i)^2}.$$

Again, obtaining the MDIP involves evaluating the quantity $I_{\mathbf{y}}(\boldsymbol{\theta},\sigma)$ that represents the information in the data,

$$\begin{aligned} I_{\mathbf{y}}(\theta,\sigma) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p(\mathbf{y}\,|\,\theta,\sigma)\ \ln\left[p(\mathbf{y}\,|\,\theta,\sigma)\right] dy_1\, dy_2 \ldots dy_k \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(y_i-\theta_i)^2} \ln\left[\left(\sqrt{2\pi}\sigma\right)^{-k} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(y_i-\theta_i)^2}\right] dy_1\, dy_2 \ldots dy_k \\ &= \ln\left[\left(\sqrt{2\pi}\sigma\right)^{-k}\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathbf{y}_i-\theta_i)^2} dy_1\, dy_2 \ldots dy_k \\ &\quad -\left(\frac{1}{2\sigma^2}\right)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k}\sum_{i=1}^{k}(\mathbf{y}_i-\theta_i)^2\ e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathbf{y}_i-\theta_i)^2} dy_1\, dy_2 \ldots dy_k \end{aligned}$$

Since the random sample $\mathbf{y}_i$'$s$ are independent and each has Normal ($\theta_i,\sigma$), then the right hand side of the above equation could be simplified to

$$I_{\mathrm{y}}(\theta,\sigma)=\ln\left[\left(\sqrt{2\pi}\sigma\right)^{-k}\right]\prod_{i=1}^{k}\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-1}e^{\frac{-1}{2\sigma^2}(\mathrm{y}_i-\theta_i)^2}\,dy_i$$
$$-\left(\frac{1}{2\sigma^2}\right)\times\left(\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k}(\mathrm{y}_1-\theta_1)^2\;e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathrm{y}_i-\theta_i)^2}\,dy_1\,dy_2\ldots dy_k+\ldots\right.$$
$$\ldots+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k}(\mathrm{y}_i-\theta_i)^2\;e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathrm{y}_i-\theta_i)^2}\,dy_1\,dy_2\ldots dy_k+\ldots \tag{2.5.17}$$
$$\left.\ldots+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k}(\mathrm{y}_k-\theta_k)^2\;e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathrm{y}_i-\theta_i)^2}\,dy_1\,dy_2\ldots dy_k\right)$$

The integral of the first term in the right side of the above equation is unity. On the other hand, the amount between brackets in the second term is a summation of k terms each is a multiple integral. To evaluate each term, consider for instance the first term as follows:

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-k}(\mathrm{y}_1-\theta_1)^2\;e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{k}(\mathrm{y}_i-\theta_i)^2}\,dy_1\,dy_2\ldots dy_k$$
$$=\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-1}(\mathrm{y}_1-\theta_1)^2\;e^{\frac{-1}{2\sigma^2}(\mathrm{y}_1-\theta_1)^2}\,dy_1\times\prod_{i=2}^{k-1}\int_{-\infty}^{\infty}\left(\sqrt{2\pi}\sigma\right)^{-1}\;e^{\frac{-1}{2\sigma^2}(\mathrm{y}_i-\theta_i)^2}\,dy_i$$
$$=\sigma^2\times\prod_{i=2}^{k-1}(1)$$
$$=\sigma^2$$

Evaluating the integral of each of the other (k-1) terms will lead to the same result.

Eventually, the equation (2.5.17) simplified to

$$I_{\mathrm{y}}(\theta,\sigma)=\ln\left(\sqrt{2\pi}\sigma\right)^{-k}\prod_{i=2}^{k}(1)-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sigma^2$$
$$=\ln\left(\sqrt{2\pi}\sigma\right)^{-k}-\frac{k}{2}$$
$$=\ln(2\pi)^{-k/2}-\frac{k}{2}+\ln\sigma^{-k}$$
$$=-\frac{k}{2}\left(\ln(2\pi)+1\right)+\ln\sigma^{-k}$$

The MDIP can be obtained by substituting in the generalized form of (2.5.7) in the multiparameter case. Then, the MDIP of such problem is given by

$$p(\theta,\sigma)\propto\sigma^{-k}\,. \tag{2.5.18}^{1}$$

[1] It is of great interest to notify that the result in (2.5.18) is different from the one derived by Zellner (1971, p. 53). We contacted professor Zellner to discuss such issue. Professor Zellner confirmed that our proof is correct and there was a typing error in his book.

## 2.6. **Concluding Remarks**

In summary, the noninformative priors derived through the previous examples are presented in the following table.

**Table 2.1: Noninformative priors for some selected sampling distributions**

| **Sampling Distributions** | **Jeffreys ' Prior** | | **Locally Uniform Prior** | **MDIP** |
|---|---|---|---|---|
| | **General Rule** | **Independence Rule** | | |
| Binomial ( $\theta$ ) | $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ | | $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ | $p(\theta) \propto \theta^{\theta}(1-\theta)^{(1-\theta)}$ |
| Normal ( $\theta$ ) | $p(\theta) \propto$ constant | | $p(\theta) \propto$ constant | $p(\theta) \propto$ constant |
| Normal ( $\sigma$ ) | $p(\sigma) \propto \sigma^{-1}$ | | $p(\sigma) \propto \sigma^{-1}$ | $p(\sigma) \propto \sigma^{-1}$ |
| Normal ( $\theta, \sigma$ ) | $p(\theta,\sigma) \propto \sigma^{-2}$ | $p(\theta,\sigma) \propto \sigma^{-1}$ | $p(\theta,\sigma) \propto \sigma^{-1}$ | $p(\theta,\sigma) \propto \sigma^{-1}$ |
| $k$-Normal ( $\theta, \sigma$ ) | $p(\theta,\sigma) \propto \sigma^{-(k+1)}$ | $p(\theta,\sigma) \propto \sigma^{-1}$ | $p(\theta,\sigma) \propto \sigma^{-1}$ | $p(\theta,\sigma) \propto \sigma^{-k}$ |

Regarding the results displayed in the previous table, one may notify some outstanding remarks. First, the different noninformative prior approaches may lead to the same prior p.d.f.. In further details, the noninformative prior p.d.f. of the binomial parameter has the same form that meets both the invariant principle and data translated likelihood concept, however, considering the principle of MDIP leads to a different form, which is the beta form. Zellner (1977) discussed the main properties of those two forms with respect to the uniform prior ( $p(\theta)=1$ ). Hence, he pointed that both Jeffreys' prior and MDIP are proper p.d.f.s, while the MDIP is symmetric around ½ and lies between the Jeffreys' prior and the uniform prior. Moreover, MDIP tends to 1 as $\theta$ tends to 0 or 1. However, Jeffreys' prior tends to $\infty$ as $\theta$ tends to 0 or 1. On the other hand, the three approaches lead to the same prior distribution in case of sampling from Normal($\theta$) and from Normal($\sigma$). Nevertheless, when sampling from Location-Scale-Normal distribution Normal($\theta$, $\sigma$), the Locally uniform prior is the best noninformative approach to be used, since it leads to an accepted noninformative prior form. While Jeffreys' discards his general rule by assuming independence to lead to the same form attained by using the data translated likelihood without assuming independence. Furthermore, adapting the principle of MDIP leads to a different form that contains higher information that is considered by Jeffreys' as dissatisfying information.

# *Chapter 3*

# *Informative prior distributions*

## 3.1. Perspective on informative priors

Bayesian scheme allows one to incorporate prior information into statistical models, before observing data, for decision-making. It then works for producing the posterior distribution by the aid of Bayes' rule. Inference problems concerning the parameters of interest will mainly depend on this distribution since it summarizes all the available information about the parameters, both prior information and sample information. One motivation to incorporate such information is that in certain problems, taking into account cogent information that are not contained in the sampling distribution, can improve the accuracy and the reliability of conclusions (Litterman, 1980). Thus, prior information is a crucial element in Bayesian framework so it attracts numerous statisticians to develop approaches to coin such information. These prior information or beliefs about parameter may be available, usually subjectively, in terms of historical information or expert judgment. It was stated that a convenient way to quantify such prior information is in terms of an appropriate probability density function of the parameter of interest (Berger, 1985). This chosen p.d.f. has to be adequate in representing the prior information otherwise another prior p.d.f. has to be chosen by the investigator to do the same function (Zellner, 1971). Such a prior p.d.f. is the so-called ***informative prior distribution***.

It is worth stressing that, in practice of Bayesian statistics, noninformative prior distributions are used for cases in which expert judgment is unavailable or not of interest. However, it is appealing to incorporate any available information about the parameter as an informative prior to the analysis. Ignoring this information, just for the sake of objectivity, is not recommended. Thus, quantifying prior information is often corresponding to subjective Bayesian system. Prior distributions, in this context, attempt to model the unavoidable ambiguity in life and nature (Pericchi, 1998).

Subjective beliefs are usually available in scientific inference. For example, a scientist decides to do a particular experiment in order to confirm some hypothesis about the parameter (Press, 1989).

### 3.1.2. Interpretations for informative priors

Pericchi (1998) introduced two interesting different broad interpretations for informative prior distributions as follows:

**I. Sensitivity analysis**

This is also often called "Bayesian robustness" or "collection of individual priors". Bayesian robustness aims to establish a neighborhood around a sensible subjective prior. Two intrinsic characteristics of such interpretation are

1. Classes of priors are composed of priors individually judged to be reasonable and compatible with the partial available information.

2. Each prior is consistent with actual prior beliefs but it is recognized that prior beliefs are imprecise.

**II. Collective prior**

In such interpretation the properties and the features of the whole class is the main concern. On the other hand, the practical features of individual priors are unimportant.

Both interpretations are similar in mathematical manipulation but different in assessment strategies. Also problems addressed by each type are different.

Informative prior distribution is commonly used in small samples where there is insufficient data to form a convenient conclusion. A probability distribution is needed to represent these subjective beliefs.

Reviewing the development elicitation methodologies for informative priors recommends the following broad ***principles*** that guide beneficially these efforts (see Hahn, 2006):

[1]. Elicitation methodologies have to be flexible enough in the form in the sense that they would generate a wide range of distributions expressing a wide range of propositions deemed by the experts.

[2]. Probabilistic judgments coined by distributions generated by these methods have to be simple and consistent. However, in case of complex ones, it is beneficial to break them down into a series of more straightforward ones.

[3]. Methods have to minimize the computational efforts of statistician. That is, these methods have to be easy to implement.

[4]. Methodologies for prior elicitation have to be applicable to a wide range of models or scenarios. That is, it is more desirable to have more broad methodologies that could be used in various settings. For example, a unified approach is desired to be applied to real-valued parameters, strictly positive parameters and parameters that exist on unit interval could save development work.

Before overviewing methodologies for prior elicitation, it is important to recognize possible types of prior information that might be quantified by prior distribution.

### 3.1.2. Types of prior information

Zellner (1971) discussed a considerable broad classification of prior information. He distinguished between what is called *data-based prior (DB)* and *non data-based prior (NDB)*. Moreover, Berger (1985) introduced an extensive summary for many other types of NDB priors.

**I. <u>Data-based prior (DB)</u>**

In this type, the prior p.d.f. represents information contained in a sample of past data that have been generated in a scientific manner. There is an inherent disapproval of such a prior because of its practical dependence on data since the idealized Bayesian view is that the prior does not depend anyway on the data. Berger (1985) described this view as not very realistic for some of the following reasons:

1. The model that describes data is often chosen after examining the data and one goes on then to define the parameter.
2. Even when the parameter is well defined outside of the experiment, yet specifying subjective prior information becomes a very sophisticated serious task in multivariate situations, particularly when the parameters have dependent coordinates. On the contrary, one should peek at the data in order to find out where prior elicitation

efforts should be concentrated. That is, to ignore the components of parameter vector that are well assessed by the data.

3. It is noticeable that even methods for developing noninformative priors mostly yield priors dependent on the model. Therefore, noninformative priors are not pure.

It must be recognized that when other reasonable choices of the prior that are not DB yield the same conclusion as the DB prior does, then the details of the prior development will not be of much concern. On the other hand, Bayesian robustness or sensitivity analysis plays a concrete role in alleviating criticism of DB priors.

Another considerable remark is that it is possible that two investigators working with the same model and DB prior information can arrive to different posterior beliefs if they base their prior information on different bodies of past data. The results could be brought into agreement by pooling their past samples to produce the same DB prior information (see Zellner, 1971)

## II. Non data-based prior (NDB)

In this type, the prior p.d.f. quantifies personal or subjective information about the parameters of the model. These subjective beliefs about parameters may be arising from introspection or theoretical considerations. Thus, such a type is relevant to the subjective view of probability. The main idea of subjective view is to let the probability of an event reflect the personal beliefs about the chance of the occurrence of the event. For more details about subjective probability in comparison with other probability views and for knowledge about methods to assess such type of probability, see Barnett (1973) and Berger (1985).

Berger (1985) proposed an interesting discussion to various sorts of subjective information that could be available about the parameters, particularly for parameters of continuous type, and that be beneficial in elicitation of an appropriate prior p.d.f.

### *The Histogram approach*

In such sort of information, the space of the parameter is divided into intervals. Subjective information about parameter could be available in a form of subjective probabilities assigned to each interval. In the sake of constructing an appropriate prior

p.d.f., plot the probability histogram. Then, smoothing this histogram will lead to the prior density. This technique for developing informative prior is known as the histogram approach. There are some difficulties in applying this approach. Since, there is no clear-cut rule to control number of intervals. Moreover, it is hard to be applied in infinite intervals (with tails).

***The Relative likelihood approach***

Here again, the parameter space, say $\Theta$ is a subset of real line, is divided into intervals. Subjective probabilities could be assigned to the relative "likelihoods" or "odds" ratios of various pairs of points in the space. A direct sketching to these points could bring a prior density. It is evident that such a method involves comparing a vast pairs of points to produce an accurate sketch.

There are several advantages to the relative odds ratio prior methodology. First, the task is straightforward to the expert. Second, it is quite general and applicable for many parameter cases. Third, it is easy to be used to produce graphical output that can be used to provide additional feedback to the expert. However, a difficulty is encountered when using such methodology with unbounded $\Theta$ where tails may not be included in such algorithm, since it is applicable in finite region. A possible reply to such problem is that the expert is free to continue adding intervals to the p.d.f. until it has been sufficiently well specified. A comprehensive discussion to such problem is covered in Berger (1985).

A recent work by Hahn (2006) has refined this approach to be implemented with Markov Chain Mont Carlo (MCMC) methods. In that work, $\Theta$ is divided into $k$ intervals. Denote the $i^{th}$ interval as $\theta_i$. Subjective information are assigned in a form of a series of expert's judgments indicating the relative likelihoods or odds of $\theta_i$ compared to $\theta_j$ where $i, j=1,2,...,k$ and $i < j$. This process is repeated by eliciting relative odds ratios for all $\theta_i$ and $\theta_j$, which requires $\frac{1}{2}k(k-1)$ judgments from the expert that can be incorporated in a matrix. The resulted matrix is termed as the matrix of judgments. Hahn used the principle of Kullback-Leibler divergence to derive the prior p.d.f. The resulted informative prior p.d.f. has the interpretation of being the best estimate of the expert's underlying distribution that generates his judgments, for more details see Hahn (2006).

***The cumulative distribution function approach (CDF)***

Subjective information may also be available for several α-fractiles. Then, the CDF of the parameter of interest, say $Z(\alpha)$, could be constructed for each α. The prior p.d.f. could hence be assessed by plotting and smoothing the curve joining the points $(\alpha, Z(\alpha))$ for all α.

***Information match a given functional form***

The preceding types of subjective information has so far been discussed are of nonparametric nature. However, another parametric type of information is useful by assuming that the prior density is of a given functional form, which may belong to a standard density function. It is evident that this given distribution will be a function of another, frequently unknown, parameters. Those parameters are called the hyperparameters, therefore, this technique is described as parametric (Berger, 1985).

## 3.2. **Literature Review**

Overview of the literature on developing elicitation methodologies for informative priors shows a vast history with several controversies that are still not entirely resolved (Jaynes, 1985). On the late 1940's, the prior information idea was strongly instructed, however written work on such issue does not appear at all, possibly, since prior knowledge was hard to document.

Representing the prior information by a proper distribution has been widely covered in statistical literature. A statistician may represent his subjective prior beliefs using "some functional form" without any restrictions. This approach usually requires an application of numerical integration methods to get the posterior distribution.

Another well known and widely used approach is the so-called "conjugate priors", discussed by Raiffa and Schlaifer (1961), DeGroot (1970) and Berger (1985). These priors are chosen such that they have the same functional form as the likelihood when the last is expressed as a function of the parameters. These priors have many useful properties that will be discussed later. A recent work for Packiorek (2006) discussed a certain type of conjugate prior for the normal linear model that is called the unit information prior.

Zellner (1986) presented “g-prior” as an informative prior in the Bayesian regression analysis.

Another type of informative priors, introduced by Kadane (1980), Kadane *et al.* (1980), Geisser (1990), Winkler (1967 and 1980) and West *et al*. (1994), are called predictive distribution priors. This type of prior involves assessment of the expert’s beliefs based on the sample from the process under interest. Thus, this approach suggests using the marginal distribution of the observed sample to determine the prior distribution.

The ML-II (the type II maximum likelihood prior) is another powerful technique to select an informative prior distribution. This approach is developed and applied by many authors such as Good (1983a) and Berger and Berliner (1983). Such technique involves assuming that the prior p.d.f. belongs to a given functional form then determine the prior parameters, the hyperparameters, using the maximum likelihood approach. A similar approach discussed by Berger (1985) is the moment approach that is to determine the hyperparameters using the sample moments.

Lindley and Smith (1972) and Good (1983b) and a recent work for Berger and Strawderman (1993) developed another important approach of informative priors that is called "hierarchical priors". Such approach is used when one has more than one type of prior information at the same time. Hierarchical approach involves modeling these information in stages.

## 3.3. **Natural Conjugate Priors**

A class $\Pi$ of prior distributions is called conjugate class for the class of density functions $\mathrm{F}$, if the resulted posterior density $\pi(\theta \mid x)$ belongs to the same class $\Pi$ for any prior distribution $\pi(\theta) \in \Pi$, and any density function $f(x \mid \theta) \in \mathrm{F}$, see Berger (1985). Some illustrative applications for the use of the natural conjugate prior will be introduced through the following two sections. The most important type in conjugate class is the so-called ***natural conjugate (NC) prior***. It is constructed by choosing the

conjugate family class having the same functional form (kernel) of the likelihood function. Natural conjugate prior is also called “convenience prior”.

### 3.3.1. Properties

Raiffa and Schlaifer (1961) proposed to generate the family $\Pi$ so that it satisfies the following properties:

1. Closure property**:** Conjugate priors allow one to begin with a certain family of distributions and end up with a posterior distribution of the same family, but with parameters updated by the sample information. Therefore, conjugate priors are called "closed under sampling" or "closed under multiplication".
2. Property of tractability: The conjugate priors are analytically tractable so that they ease the computations of the posterior distribution given a certain sample. This property is the main reason of their popularity in time series analysis. They are frequently used in time series analysis such as Broemeling (1985).
3. Richness property: The conjugate family of priors is very rich. It contains many members from well-known standard forms that are able to express the prior information in various situations.
4. Interpretable property: The conjugate family, $\Pi$, should be parameterized in a manner which can be interpreted so that it will be easy to verify that the chosen member of the family is really in close agreement with the decision-maker's prior judgments about $\theta$.

### 3.3.2 Derivation

Raiffa and Schlaifer (1961) have developed a class of distributions that attains the above properties. However, they confined the development to the case where the sample observations are independent and admit sufficient statistics of fixed dimensionality. Their main idea to develop the natural conjugate class is to use the sample kernel as a prior kernel.

Definition:

Consider the i.i.d. random sample $X_1, X_2, ..., X_n$ such that for any n and any sample $(x_1, x_2, ..., x_n)$, there exist a sufficient statistic $\tilde{y}_n(x_1, x_2, ..., x_n) = y = (y_1\ y_2 \dots y_s)$

where $y_i's$ are real numbers in the range Y and the dimensionality *s* does not depend on *n*. Then, the LF is give by $l(x_1, x_2, ..., x_n|\theta) \propto k(y|\theta)$. In such representation of the LF, *k* is called the sample kernel. The natural conjugate prior with parameter $y'$ in the range Y can then be given by

$$P_{NC}(\theta) \propto k(y'|\theta), \quad y' \in Y \tag{4.1}$$

Then, the posterior distribution of $\theta$ will be given by

$$P(\theta|y) \propto k(y'|\theta)\ k(y|\theta) \tag{4.2}$$

Raiffa and Schlaifer (1961, p.53) illustrated some considerable examples for some data-generating processes. It is of quite interest to show some of them through the following table:

**Table3.1: Some Natural Conjugate prior distributions**

| **Data Generate from:** | **Bernoulli Process** | **Poisson Process** (Exponential distribution[2]) | **Normal Process** (Both Mean and precision unknown) |
|---|---|---|---|
| Sample Mass / Density Fun. | $f(x\|\theta) = \theta^x(1-\theta)^{1-x}$, where $x = 0,1,\ 0 < \theta < 1$ | $f(x\|\lambda) = \lambda e^{-\lambda x}$, where $x \geq 0, \lambda > 0$ | $f(x\|\mu,\tau) = (2\pi)^{-\frac{1}{2}}\tau^{\frac{1}{2}} e^{-\frac{\tau}{2}(x-\mu)^2}$, where $-\infty < x, \mu < \infty,\ \tau > 0$ |
| Likelihood Fun. | $l(x_1, x_2, .., x_n\|\theta) \propto \theta^r(1-\theta)^{n-r}$ | $l(x_1, x_2, ..., x_n\|\lambda) \propto \lambda^n e^{-\lambda r}$ | $l(x_1, x_2, ..., x_n\|\mu,\tau) \propto \tau^{\frac{\nu}{2}} e^{-\frac{1}{2}\nu s^2\tau} \times \tau^{\frac{1}{2}} e^{-\frac{1}{2}\tau(m-\mu)^2}$ |
| Sufficient Statistics | $y = (n, r)$ where $r = \sum x$ | $y = (n, r)$ where $r = \sum x$ | $y = (\nu, m, s)$ where $\nu = n-1$ $m = \frac{1}{n}\sum x_i$ and $\nu s^2 = \sum (x_i - m)^2$ |
| Natural Conjugate (NC) Prior Dist. is: NC prior p.d.f. | **Beta** $p(\theta\|r', n') \propto \theta^{r'-1}(1-\theta)^{n'-r'-1}$ | **Gamma-I** $p(\lambda\|n', t') \propto \lambda^{n'-1}e^{-\lambda r'}$ | **Normal Gamma-I** $p(\mu,\tau\|m', s'^2, \nu') \propto \tau^{\frac{\nu'}{2}-1} e^{-\frac{1}{2}\nu' s'^2\tau} \times \tau^{\frac{1}{2}} e^{-\frac{1}{2}\tau(m'-\mu)^2}$ |

[2] *x* is a random variable denoting the time between two successive occurrences of a random event. Such an event is generated from Poisson distribution, hence *x* is exponentially distributed.

### 3.3.3. Difficulties in assessment of Natural Conjugate Priors

The main problem in using a conjugate class is that one must estimate the parameters of the prior distribution, namely, the hyperparameters. There is no ideal method to give a general rule by which the hyperparameters can be estimated. However, there are some ways to estimate the hyperparameters, as follows:

*1. Historical relative frequency distribution method*

In some applications, there are previously available relative frequency distributions for the values of the prior parameters. It is reasonable to match the parameters of the current prior distribution with the historical frequency distribution of their values. Then, choose the prior distribution that gives the closest form to the historical distribution (see Raiffa and Schlaifer, 1961).

*2. Moment method*

If there are available information about the prior moments, then the hyperparameters can be estimated by expressing them as functions of these moments. This method is not recommended in the case of skewed prior distributions because of their drastic effect on their moments, see Berger (1985).

*3. Fractiles method*

Another method for assessing prior parameters starts by a subjective determination of the prior median and some other odd fractiles such as odd quartiles and odd octiles. Then, choose the parameters of the given prior distribution to obtain a density that matches these fractiles as closely as possible, see (Berger, 1985). This method depends on little trail-and-error calculations, and is also called “subjective betting odds”. For more details, one may refer to (Raiffa and Schlaifer, 1961) and (Lempers, 1971).

*4. Predictive method*

The prior parameters can be estimated also in terms of the predictive density of the observations. This approach is sometimes called “the device of imaginary results” (Berger, 1985) and (Broemeling, 1985).

Consider a prior distribution depends on the hyperparameter $\alpha \in \mathrm{A}$, then the predictive density depends on $\alpha$ through the prior can be as follows:

$$p(x|\alpha)=\int_{\Omega} f(x|\theta)\,\pi(\theta|\alpha)d\theta\ ,\ X\in S\ ,\ \alpha\in \mathrm{A} \tag{4.3}$$

where $f(x|\theta)$ is the data density function defined on the sample space $S$ for given values of $\theta\in\Omega$, whereas $\pi(\theta|\alpha)$ is the prior distribution of $\theta$ given the unknown hyperparameter $\alpha$ that requires to be estimated. One can observe values $x_1, x_2, ..., x_n$, imaginary future or past values, from the density in (4.3) and choose $\alpha$ that is compatible with this predictive density. That means to use this predictive density incorporated with the future or past observed values to estimate $\alpha$ with the standard known methods of estimation such as moment or maximum likelihood methods. Broemeling (1985) preferred this method for estimating the prior parameters because of its property to think about $\theta$ as an observed random variable rather than as an unobserved parameter.

5. Training Sample method

Another helpful approach to estimate the hyperparameters is the training sample approach. Such a method has been widely applied in the area of the objective Bayesian analysis. Since the training sample admits utilization of improper objective priors (noninformative priors) to a subset of the observed data to obtain a proper posterior distribution. This last distribution is used to estimate the hyperparameters. Then the Bayesian structure is applied to the rest of the sample as if it was the actual sample to obtain the posterior analysis. Various Bayesian applications used such approach in literature such as Broemling (1985) and Ismail (1994). A recent work to develop a variety of methods of choosing training samples is due to the work of Berger and Pericchi (1996), Pérez and Berger (2002) and Berger and Pericchi (2004). Berger and Pericchi (2004) developed some new definition of training samples that can overcome a wide range of problems in Bayesian analysis. However, they deemed that it is unable to define any type of "optimal" training sample. According to the revision of many training sample techniques discussed in literature, a training sample could be chosen to be as small as possible and that convert improper objective prior into a proper distribution. This type of training samples is called "minimal training sample". It is of limited use particularly when the data set is small, see Berger and Pericchi (2004). Another solution reviewed in Ismail (1994) is the "overlap training sample". This technique suggests

using the whole data set including the training sample that is used to estimate the hyperparameters. That is the actual data overlaps the training sample. A more generalized view to select a training sample is discussed in Berger and Pericchi (2004), where they introduced what is called "randomized and weighted training samples" that chooses a training sample according to sampling mechanism, they also discussed the "imaginary training samples". In such type, training samples are not obtained from the real data, but from some specified distribution.

### 3.3.4. Examples

DeGroot (1970) presented several types of the natural conjugate priors for samples from various distributions. Table 3.2, in addition to table 3.1, summarize some of the natural conjugate priors that correspond to different populations.

**Table 3.2 Some Natural Conjugate prior distributions**

| Sampling distributions | Natural conjugate prior distributions |
|---|---|
| 1. Bernolli | Success probability is Beta |
| 2. Binomial | Success probability is Beta |
| 3. Negative binomial | Success probability is Beta |
| 4. Poisson | Mean is Gamma |
| 5. Uniform $(\kappa_1, \kappa_2)$ | $(\kappa_1, \kappa_2)$ has joint bilateral bivariate Pareto |
| 6. Exponential with mean $\lambda^{-1}$ | $\lambda$ is Gamma |
| 7. Normal with $\sigma^2$ is known | Mean is Normal |
| 8. Normal with $\mu$ is known | Variance is Inverted Gamma |
| 9. Normal with $\mu$ and $\sigma^2$ are unknown | $(\mu, \sigma^2)$ has joint Normal-Gamma |

The Normal-Gamma conjugate prior is widely used in Bayesian literature, especially in time series field, since sampling from normal distribution is the most common case.

## 3.4. G-Prior

### 3.4.1. Introduction

As shown in the previous section, the natural conjugate prior technique is an appealing one for assessing informative prior distributions that lead to relatively simple posterior results. Zellner (1985) considered the class of natural conjugate prior distributions as a very useful class of "*reference informative priors (RIPs)*". However, that technique encounters a serious pitfall in evaluating the prior covariates of the parameters. That motivates Zellner(in 1983 and 1986) to seek another prior, belonging to the same class, that figures out this problem and has the same attractive properties as the natural conjugate priors. Zellner's main concern was with simplifying the Bayesian results for one of the most well known models in econometrics, the **general linear model (GLM)**, therefore he confined his work to the derivation of RIPs for the regression parameters. Such work leads to what is called ***g-priors***, the class of priors that provides a middle ground of sorts between an informative natural conjugate prior and a diffuse prior, see Karlsson (2001). The main feature of g-prior is that it allows the investigator to introduce information about the location of the regression parameters without having to think about the most difficult aspects of prior specification, which is the prior covariates structure of the regression parameters.

Zellner's g-prior has later been extensively utilized for many problems in econometrics. For instance, Zellner (1985) applied the g-prior for a simple-structural econometric model. Moreover, g-prior has become a standard choice for the regression coefficients in the field of Bayesian model averaging (BMA) for several practical reasons, See, e.g., Fernández, *et. al.* (1998), Jörnsten and Yu (2002), Clyd (2003), and Koop and Potter (2003).

In this section, a standard GLM will be considered for an $n \times 1$ vector of observations on the dependent variable **y**, wherein **y** is generated through the following model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \tag{3.4.1}$$

where $\boldsymbol{X}$ is an $n\times k$ non-stochastic design matrix of rank $k$, $\boldsymbol{\beta}$ is a $k\times 1$ regression parameter vector and the error vector $\boldsymbol{u}$ is assumed to be a $N(\mathbf{0},\sigma^2\boldsymbol{I}_n)$, where $\sigma^2$ is finite unknown value. The likelihood function (LF) for the GLM is given by

$$l(\boldsymbol{\beta},\sigma\,|\,\boldsymbol{y},\boldsymbol{X})\propto\sigma^{-n}\exp\left\{-(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/2\sigma^2\right\} \tag{3.4.2a}$$

It is desirable to rewrite the quadratic quantity in the exponent of (3.4.2a) in terms of the least squares estimates $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, as follows:

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

Now completing the square in the right side of the last quantity with respect to $\boldsymbol{\beta}$ implies to

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) &= \boldsymbol{y}'\boldsymbol{y}+\left((\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})\right)-\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= \boldsymbol{y}'\boldsymbol{y}+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})-\boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

then completing the square with respect to $\boldsymbol{y}$ leads to

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}-\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

Since $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}=\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$, the form of the LF, in (3.4.2a), can be eventually written as

$$l(\boldsymbol{\beta},\sigma\,|\,\boldsymbol{y},\boldsymbol{X})\propto\sigma^{-n}\exp\left\{-[\nu s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]/2\sigma^2\right\} \tag{3.4.2b}$$

where $\nu s^2=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ and $\nu=n-k$.

Procedures for assessing informative prior distributions for the GLM's parameters have been used by many authors such as Winkler (1967, and 1977), Kadane *et. al.* (1980) and Zellner (1985). In what follows, the derivation of the g-prior distribution for the GLM parameters will be illustrated. In a next chapter, The posterior results based on the g-prior will be displayed and compared with those based on the natural conjugate approach.

### 3.4.2. **Derivation**

Zellner (1983 and 1986) innovates an approach to derive a reference informative prior (RIP) distribution, as he called, as the joint g-prior distribution of $\boldsymbol{\beta}$ and $\sigma$ through the following steps:

1. Before observing ***y,*** consider another imaginary or conceptual sample $\boldsymbol{y_0}$ assumed to be generated by

$$\boldsymbol{y_0} = \boldsymbol{X\beta} + \boldsymbol{u_0} \tag{3.4.3}$$

where $\boldsymbol{X}$ is the same design matrix defined by (3.4.1), but $\boldsymbol{u_0}$ is assumed to be $N(\mathbf{0}, \sigma_0^2 \mathbf{I_n})$, where $\sigma^2 = g\sigma_0^2$, and $\boldsymbol{u} = \sqrt{g}\,\boldsymbol{u_0}$, where g is assumed to be initially given.

2. Assume the Jeffreys' independent rule for the joint prior p.d.f. of $\boldsymbol{\beta}$ and $\sigma$. Then $p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1}$. Let $S_0$ denotes the conceptual sample information in (3.4.3), then the posterior p.d.f. $p(\boldsymbol{\beta}, \sigma \mid S_0)$ can be evaluated by combining the LF of the model in (3.4.3) and the Jeffreys' prior p.d.f. $p(\boldsymbol{\beta}, \sigma)$. Then the posterior p.d.f. of $\boldsymbol{\beta}$ and $\sigma$ will have the following form:

$$p(\boldsymbol{\beta}, \sigma \mid S_0) \propto \sigma^{-(n+1)} \exp\left\{-g[\nu s_0^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})' \boldsymbol{X}'\boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})] / 2\sigma^2\right\} \tag{3.4.4a}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{0}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y_0}$, $\nu s_0^2 = (\boldsymbol{y_0} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}})'(\boldsymbol{y_0} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}})$ and $\nu = n - k$.

It can be seen that the posterior p.d.f. in (3.4.4a) is a normal inverted-gamma distribution given by

$$p(\boldsymbol{\beta}, \sigma | S_0) \propto \sigma^{-(\nu+1)} \exp\left\{-g\nu s_0^2 / 2\sigma^2\right\} \times \sigma^{-k} \exp\left\{-g[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})' \boldsymbol{X}'\boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})] / 2\sigma^2\right\} \tag{3.4.4b}$$

where the first part of (3.4.4b) is the inverted-gamma(I Γ) of type II. Raiffa and Schlaifer (1961) have represented diifferent forms of gamma and inverted gamma distributions. It is worthwhile to shed further light on some of these distributions in a separate appendix (see appendix-I).

Hence, the marginal posterior p.d.f. for $\boldsymbol{\beta}$ is obtained by integrating the form in (3.4.4a) with respect to $\sigma$, which gives the kernel of I Γ–II distribution with parameters $\left(r = n, r\lambda^2 = g[\nu s_0^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})' \boldsymbol{X}'\boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})]\right)$, which turns out, as shown in distribution IV in appendix-I that

$$p(\boldsymbol{\beta} \mid S_0) \propto \left[1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})' \boldsymbol{X}'\boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})}{\nu s_0^2}\right]^{\frac{-(\nu+k)}{2}} \tag{3.4.5a}$$

which is obviously the k-variate t distribution with ν degrees of freedom, mean vector $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$, and a dispersion matrix proportional to $(\boldsymbol{X}'\boldsymbol{X})^{-1} \nu s_0^2 / (\nu - 2)$.

Similarly, the marginal p.d.f. for $\sigma$ can be obtained from (3.4.4b) by integration with respect to $\boldsymbol{\beta}$, that is integrating the k-variate normal part in (3.4.4b) which gives constant. Hence the marginal posterior p.d.f. for $\sigma$ is

$$p(\sigma \mid S_0) \propto \sigma^{-(\nu+1)} \exp\left\{- g\nu s_0^{\,2} / 2\sigma^2\right\} \tag{3.4.5b}$$

which, through the form IV in appendix-I, is I Γ–II distribution with parameters $\left(r = \nu, \lambda^2 = g s_0^2\right)$.

3. Zellner suggests using anticipated values for $\boldsymbol{\beta}$ and $\sigma^2$ denoted by $\boldsymbol{\beta}_a$ and $\sigma_a^2$ respectively. Zellner applied the Muth's (1961) rational expectations hypothesis that is taking them respectively equal to $E(\boldsymbol{\beta} \mid S_0)$ and $E(\sigma^2 \mid S_0)$, the posterior means derived based on (3.4.5a) and (3.4.5b). Thus $\boldsymbol{\beta}_a = E(\boldsymbol{\beta} \mid S_0) = \hat{\boldsymbol{\beta}}_{\mathbf{0}}$ and $\sigma_a^2 = E(\sigma^2 \mid S_0) = \nu g s_0^2 / \nu - 2$.

4. Zellner recommends using the joint g-prior distribution as given by

$$p_g(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\theta}_0)\, \alpha\, \sigma^{-(n+1)} \exp\left\{-[\nu \bar{\sigma}_a^2 + g(\boldsymbol{\beta} - \boldsymbol{\beta}_a)' \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_a)] / 2\sigma^2\right\} \tag{3.4.6}$$

which is still the form of the normal inverted-gamma, where $\boldsymbol{\theta}_0' = (\boldsymbol{\beta}_a', \bar{\sigma}_a^{\,2}, g, \nu)$ is the vector of hyperparameters, $\bar{\sigma}_a^2 = g s_0^2 = (\nu - 2)\sigma_a^2 / \nu$ and $\nu = n - k$. It is evident that, this prior form is the same as in (3.4.4a) and (3.4.4b) but the first is in terms of the anticipated values $\boldsymbol{\beta}_a$ and $\sigma_a^2$. It can also be seen, as shown above in (3.4.5a) and (3.4.5b), that the marginal g-prior p.d.f. of $\sigma$ is the I Γ-II$\left(r = \nu, \lambda^2 = \bar{\sigma}_a^2\right)$, which takes the following form,

$$p_g(\sigma \mid \nu, \sigma_a^2) \propto \sigma^{-(\nu+1)} \exp\left\{- \nu \bar{\sigma}_a^2 / 2\sigma^2\right\} \tag{3.4.7a}$$

So it is seen that $E(\sigma^2) = \nu \bar{\sigma}_a^2 / (\nu - 2) = \sigma_a^2$. Where the marginal g-prior for $\boldsymbol{\beta}$ is of the following form,

$$p_g(\boldsymbol{\beta} \mid \boldsymbol{\beta}_a, g, \nu, \bar{\sigma}_a^2) \propto \left[1 + g \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})' \boldsymbol{X}'\boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})}{\nu \bar{\sigma}_a^2}\right]^{\frac{-(\nu+k)}{2}} \tag{3.4.7b}$$

which is the multivariate t distribution with ν degrees of freedom, mean vector $\boldsymbol{\beta}_a$, and a precision matrix proportional to $g(\boldsymbol{X}'\boldsymbol{X}) / \bar{\sigma}_a^2$, hence the dispersion matrix is proportional to $(\boldsymbol{X}'\boldsymbol{X})^{-1} \nu \bar{\sigma}_a^2 / (\nu - 2) g = (\boldsymbol{X}'\boldsymbol{X})^{-1} \sigma_a^2 / g$.

### 3.4.3. **Properties**

Zellner discussed some properties of the g-prior in (3.4.6) which can be summarized as follow:

1. When g value is unknown, a prior p.d.f. for g, noninformative or informative, can be introduced and g-prior can be integrated out.
2. When g is unknown, it could be taken as a function of the sample size, $g = g(n)$, say $g = \frac{1}{n}$ or $g = \frac{\log n}{n}$. In such assumption, g controls the dependence of the prior precision on the sample size, so an appropriate choice of this function can allow prior precision to grow with *n* and if desired at a rate less than the rate at which the sample precision grows. Other potential values for g are investigated in literature and will be exhibited in next subsection.
3. In case when another design matrix, say $\boldsymbol{X_0}$, is given, the form of the regression model is the same for both design matrices $\boldsymbol{X_0}$ and $\boldsymbol{X}$. That is (3.4.3) can be rewritten as $\boldsymbol{y_0} = \boldsymbol{X_0}\boldsymbol{\beta} + \boldsymbol{u_0}$, with $\boldsymbol{y_0}$ and $\boldsymbol{u_0}$ each of dimension $n_0 \times 1$ and $\boldsymbol{X_0}$ of dimension $n_0 \times k$ and same approach, discussed above to derive the g-prior for the regression parameters, can be proceeded.

### 3.4.4. **Potential values for g**

The choice of the unknown hyperparameter g is crucial for obtaining sensible results. Therefore, several methods are followed to assign the value of g.

***Information criteria methods***

Fernández, *et. al.* (1998) investigated the properties for many choices for the unknown scalar g. That work shows that some of these choices yield posterior results that have properties similar to commonly used information criteria. They concluded that g-prior could possibly be assigned as a function of the sample size or the number of regressors in the data set. So based on simulation studies, they analyze the consequences of using many different choices of g-prior. It is of interest to introduce herein some of them.

[1]. $\boxed{g = \frac{1}{n}}$

This prior corresponds to assigning the same amount of information, the same weights, to the conditional prior of $\boldsymbol{\beta}$ as contained in one observation. This comes up with the spirit of the "unit information prior" of Kass and Wesserman (1995).

[2]. $\boxed{g = \frac{k}{n}}$

Here more information are assigned as many regressors have been entered in the model. That involves more shrinkage induced in $\boldsymbol{\beta}$ to the prior mean $\bar{\boldsymbol{\beta}}$ as the number of regressors grows, see equation (4.1.3b) in the next section.

[3]. $\boxed{g = \sqrt{\frac{1}{n}}}$

The value of g using this prior behaves asymptotically like Schwarz criterion.

[4]. $\boxed{g = \sqrt{\frac{k}{n}}}$

As in prior [2], more shrinkage induced as number of regressors increases.

[5]. $\boxed{g = \frac{1}{(\ln n)^3}}$

[6]. $\boxed{g = \frac{\ln(k+1)}{\ln n}}$

The priors given in [5] and [6] behave asymptotically like Hnnan-Quinn criterion.

[7]. $\boxed{g = \frac{1}{k^2}}$

Using this prior implies the Risk Inflation Criterion (RIC) of Foster and George (1994).

Fernández, *et. al.* (1998) concluded, on the ground of consistency, that it is better to suggest making g-prior as a decreasing function of the sample size *n*. Moreover, they deduced using simulation that the most reasonable choices of g-prior are

- $g = \frac{1}{k^2}$, for $n \le k^2$.
- $g = \frac{1}{n}$, for $n > k^2$.

In addition to such choices for g listed above, there is another method to assign values for g and in a data-based manner. This is the so called empirical Bayesian methodology, which was first coined by Robbins (1956).

***Empirical Bayesian methods***

In the context of this approach, the Bayesian estimation structural is used with a pre-assigned prior distribution to obtain the Bayes estimator. However, the parameters of the prior distribution, or the hyperparameters, are not assessed subjectively, rather they are estimated through the current data. Often the hyperparameters are estimated by maximizing the marginal likelihood, to get the maximum likelihood (ML), or by sample moments. Nevertheless, empirical Bayesian methodology can be criticized because allowing a prior to depend on data violates the rules of conditional probability, the Bayes' rule that requires the prior distribution depend only on its parameter not on the data set. However, the empirical Bayesian methods are popular for many practical econometricians. For more details about empirical Bayesian approach, see Press (1989) and O'Hagan (1994). Koop and Potter (2003) adopt such approach to estimate the value of a single prior hyperparameter that is g-prior here, using the maximum likelihood estimate (MLE). They adopt this methodology in the application of forecasting dynamic factor model using Bayesian model averaging.

# *Chapter 4*

# *Posterior analysis to GLM*

The g-prior is often described as a less informative prior, and has numerous terms in literature, where it is called as "objective", "benchmark", "shrinkage", and at last as "reference informative" prior. Furthermore, Fernández, *et. al.* (1998) considered the g-prior as a slightly "noninformative prior" that is related to a natural conjugate structure with g-prior specification to the hyperparameters. In addition, they considered such prior specification as a one that lead to sensible results in the sense that data information dominates prior assumptions. That is because such prior does not require substantive amounts of subjective prior election by the researcher except for the scalar parameter g, however the choice of g may be determined subjectively. Nevertheless, some objective methods to specify g are discussed in literature. Therefore, the affinity of the g-prior of the regression parameters with the natural conjugate prior must be emphasized. This difference will be clarified through the posterior analysis of the GLM.

## 4.1. Based on the g-prior distribution

Zellner(1986) introduced the following particular g-prior to derive the posterior distribution for $\boldsymbol{\beta}$ and $\sigma$,

$$p_g(\boldsymbol{\beta},\sigma) \propto p(\sigma) \times p(\boldsymbol{\beta}|\sigma, g) \tag{4.1.1}$$

where $p(\sigma) \propto 1/\sigma$, and $p(\boldsymbol{\beta}|\sigma, g) \propto \sigma^{-k} \exp[-g(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})/2\sigma^2]$, then combining this prior distribution with the LF in (3.4.2a) will lead to the following joint posterior distribution:

$$p(\boldsymbol{\beta},\sigma|S) \propto \sigma^{-(n+k+1)} \exp\left\{-[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + g(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})]/2\sigma^2\right\} \tag{4.1.2a}$$

where $S$ denotes the sample and prior information. Some simplifications will be considered now to the term in the square brackets in the exponent, say $Q$, in the right side of (4.1.2a), where $Q$ can be proved to equal,

$$Q = [\boldsymbol{y}'\boldsymbol{y} + g\bar{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\bar{\boldsymbol{\beta}}] - [(\boldsymbol{y}'\boldsymbol{X} + g\bar{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}] - [\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{y} + g\boldsymbol{X}'\boldsymbol{X}\bar{\boldsymbol{\beta}})] + [\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + g\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}]$$

If the two matrices $\boldsymbol{w}' = \left(\boldsymbol{y}' : g^{1/2}\bar{\boldsymbol{\beta}}'\boldsymbol{X}'\right)$ and $\boldsymbol{W}' = \left(\boldsymbol{X}' : g^{1/2}\boldsymbol{X}'\right)$ are considered then $Q$ can be expressed as,

$$Q = \boldsymbol{w}'\boldsymbol{w} - \boldsymbol{w}'\boldsymbol{W}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{W}'\boldsymbol{w} + \boldsymbol{\beta}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{\beta}$$

Similar simplification made previously in (3.4.2b) will be applied herein. Thus, through completing the square with respect to $\boldsymbol{\beta}$, $Q$ can be rewritten as,

$$\begin{aligned} Q &= \boldsymbol{w}'\boldsymbol{w} + [\boldsymbol{\beta} - (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{w}]'\boldsymbol{W}'\boldsymbol{W}[\boldsymbol{\beta} - (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{w}] - \boldsymbol{w}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{w} \\ &= \boldsymbol{w}'\boldsymbol{w} + (\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}}) - \boldsymbol{w}'\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}} \end{aligned}$$

where $\overline{\overline{\boldsymbol{\beta}}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{w}$, then completing the square, in the right side of the last form of $Q$, with respect to $\boldsymbol{w}$ implies:

$$\begin{aligned} Q &= (\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}}) + \overline{\overline{\boldsymbol{\beta}}}'\boldsymbol{W}'\boldsymbol{w} - \overline{\overline{\boldsymbol{\beta}}}'\boldsymbol{W}'\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}} + (\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}}) \\ &= (\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}}) + (\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}}) \end{aligned}$$

where $\overline{\overline{\boldsymbol{\beta}}}'\boldsymbol{W}'\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{w} = \overline{\overline{\boldsymbol{\beta}}}'\boldsymbol{W}'\boldsymbol{w}$, thus (4.1.2a) can be finally expressed as

$$p(\boldsymbol{\beta}, \sigma \mid S) \propto \sigma^{-(n+k+1)} \exp\left\{-[(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}}) + (\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})] / 2\sigma^2\right\} \qquad (4.1.2b)$$

It is evident that the joint distribution in (4.1.2b) is in the normal inverted gamma form, where the marginal posterior distribution of $\boldsymbol{\beta}$ is obtained from it by integrating with respect to $\sigma$ to get the kernel of I Γ–II distribution with parameters $\left(r = n + k, r\lambda^2 = (\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}}) + (\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})\right)$, that eventually implies the following form:

$$p(\boldsymbol{\beta} \mid S) \propto \left[1 + \frac{(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{W}'\boldsymbol{W}(\boldsymbol{\beta} - \overline{\overline{\boldsymbol{\beta}}})}{(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w} - \boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})}\right]^{\frac{-(n+k)}{2}} \qquad (4.1.3a)$$

which is simply the k-variate t distribution with $n$ degrees of freedom, posterior mean $\overline{\overline{\boldsymbol{\beta}}}$, and variance covariance matrix $V(\boldsymbol{\beta}|S)=\frac{(\boldsymbol{W}'\boldsymbol{W})^{-1}(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})}{(n-2)}$. It is important to notice that,

$$\overline{\overline{\boldsymbol{\beta}}}=(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{w}$$

$$=(1+g)^{-1}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}+g\overline{\boldsymbol{\beta}}\right)$$

Where, the matrices $\boldsymbol{w}$ and $\boldsymbol{W}$ are defined above. Then, the posterior mean $\overline{\overline{\boldsymbol{\beta}}}$ is finally given by

$$\overline{\overline{\boldsymbol{\beta}}}=\frac{\left(\hat{\boldsymbol{\beta}}+g\,\overline{\boldsymbol{\beta}}\right)}{1+g} \tag{4.1.3b}$$

However, the posterior dispersion matrix of $\boldsymbol{\beta}$, as shown above, is $V(\boldsymbol{\beta}|S)=(\boldsymbol{W}'\boldsymbol{W})^{-1}a^2=(\boldsymbol{X}'\boldsymbol{X})^{-1}a^2/(1+g)$, where,

$$(n-2)\,a^2=(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})$$
$$=(\boldsymbol{y}-\boldsymbol{X}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{y}-\boldsymbol{X}\overline{\overline{\boldsymbol{\beta}}})+g(\overline{\boldsymbol{\beta}}-\overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{X}'\boldsymbol{X}(\overline{\boldsymbol{\beta}}-\overline{\overline{\boldsymbol{\beta}}})$$

then, at last

$$V(\boldsymbol{\beta}|S)=\left[(n-2)^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left((\boldsymbol{y}-\boldsymbol{X}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{y}-\boldsymbol{X}\overline{\overline{\boldsymbol{\beta}}})+g(\overline{\boldsymbol{\beta}}-\overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{X}'\boldsymbol{X}(\overline{\boldsymbol{\beta}}-\overline{\overline{\boldsymbol{\beta}}})\right)\right]/(1+g) \tag{4.1.3c}$$

Similarly, the marginal posterior p.d.f. of $\sigma$ can be evaluated by integrating the k-multivariate normal part in (4.1.2b) which gives the distribution of the form

$$p(\sigma\,|\,S)\propto\sigma^{-(n+1)}\exp\left\{-(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})/2\sigma^2\right\} \tag{4.1.4a}$$

which is the I Γ–II distribution with parameters $\left(r=n,r\lambda^2=(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})\right)$. Thus, as shown in distribution IV in appendix-I, the posterior mean of $\sigma^2$ is $E(\sigma^2\,|\,S)=a^2=(n-2)^{-1}(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{w}-\boldsymbol{W}\overline{\overline{\boldsymbol{\beta}}})$. This quantity can be simplified to the following form, as shown above to give (4.1.3c)

$$E(\sigma^2\,|\,S)=(n-2)^{-1}\left[(\boldsymbol{y}-\boldsymbol{X}\overline{\overline{\boldsymbol{\beta}}})'(\boldsymbol{y}-\boldsymbol{X}\overline{\overline{\boldsymbol{\beta}}})+g(\overline{\boldsymbol{\beta}}-\overline{\overline{\boldsymbol{\beta}}})'\boldsymbol{X}'\boldsymbol{X}(\overline{\boldsymbol{\beta}}-\overline{\overline{\boldsymbol{\beta}}})\right] \tag{4.1.4b}$$

## 4.2. Based on the Natural Conjugate Prior

Zellner (1986) has discussed assessing another prior distribution to the GLM parameters which is the natural conjugate prior developed by Raiffa and Schlaifer (1961). This approach, using a certain prior p.d.f. of such class, will lead to a posterior p.d.f. belongs to the same class. This family of distribution involves herein representing the prior information to the GLM using the normal gamma joint prior distribution for the parameters $\boldsymbol{\beta}$ and $\tau$, where $\tau$ is the precision parameter and $\tau^{-1} = \sigma^2$. That is the joint prior distribution, $p_{NG}(\boldsymbol{\beta}, \tau)$ is given by

$$p_{NG}(\boldsymbol{\beta}, \tau) \propto p_G(\tau) \times p_N(\boldsymbol{\beta}|\tau) \tag{4.2.1a}$$

with

$$p_N(\boldsymbol{\beta}|\tau) \propto \tau^{k/2} \exp\left\{-\tfrac{\tau}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})'\boldsymbol{A}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right\} \tag{4.2.1b}$$

and

$$p_G(\tau) \propto \tau^{a-1} \exp\{-b\tau\} \tag{4.2.1c}$$

where $p_N(\boldsymbol{\beta}|\tau)$ is a k-variate normal distribution for $\boldsymbol{\beta}$ given $\tau$, with prior mean vector $\bar{\boldsymbol{\beta}}$ and prior precision matrix $\tau\boldsymbol{A}$, whereas $p_G(\tau)$ is the marginal prior distribution of $\tau$ which is the gamma distribution with parameters *a* and *b*.

Expressing the LF of (3.4.1) in terms of $\tau$ will give

$$l(\boldsymbol{\beta}, \tau | \boldsymbol{y}, \boldsymbol{X}) \propto \tau^{n/2} \exp\left\{-\tfrac{\tau}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\} \tag{4.2.2}$$

Now, joining the joint prior distribution in (4.2.1) with the LF in (4.2.2) will give the following joint posterior distribution

$$l(\boldsymbol{\beta}, \tau | S) \propto \tau^{\frac{n+2a+k}{2}-1} \exp\left\{-\tfrac{\tau}{2}\left[2b + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})'\boldsymbol{A}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right]\right\} \tag{4.2.3a}$$

this is again in the normal gamma form. To focus more on the posterior results, consider the quantity in the exponent, except for the *2b* term, say *Q* that is

$$\begin{aligned} Q &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})'\boldsymbol{A}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{A}\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}'\boldsymbol{A}\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) - (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'\boldsymbol{\beta} + \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}} \end{aligned}$$

completing the square with respect to $\boldsymbol{\beta}$ in the right side of *Q* will give

$$\begin{aligned} Q = \boldsymbol{y}'\boldsymbol{y} &+ \left[\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)\right] \\ &- (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}} \end{aligned}$$

then the posterior p.d.f. in (4.2.3a) can be rewritten as

$$p(\boldsymbol{\beta},\tau \mid S) \propto \tau^{\frac{n+2a}{2}-1} \exp\left\{-\frac{\tau}{2}\left[2b + \boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}}\right]\right\} \tag{4.2.3b}$$

$$\times \tau^{k/2} \exp\left\{-\frac{\tau}{2}\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)\right\}$$

Obtaining the marginal posterior distribution of $\boldsymbol{\beta}$ is obtained by integrating the form in (4.2.3b) with respect to τ, that gives the kernel of the gamma distribution with parameters $r = \frac{n+2a+k}{2}$ and

$$\lambda = b + \frac{\boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}} + \left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)}{2}.$$

Thus, the marginal posterior distribution of $\boldsymbol{\beta}$ will be of the form

$$p(\boldsymbol{\beta}|S) \propto \left[1 + \frac{\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})\left(\boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})\right)}{2b + \boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}}}\right]^{-\left(\frac{(n+2a)+k}{2}\right)} \tag{4.2.4a}$$

which is the k-variate t distribution with $n+2a$ degrees of freedom and a posterior mean vector $\bar{\bar{\boldsymbol{\beta}}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})$ , It can be shown that

$$\begin{aligned}\bar{\bar{\boldsymbol{\beta}}} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) \\ &= \bar{\boldsymbol{\beta}} + \left(\mathrm{I} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{A}\right)\hat{\boldsymbol{\beta}} - \left(\mathrm{I} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{A}\right)\bar{\boldsymbol{\beta}}\end{aligned}$$

then at last, the posterior mean of $\boldsymbol{\beta}$ is expressed as

$$\bar{\bar{\boldsymbol{\beta}}} = \bar{\boldsymbol{\beta}} + \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{A}\right)\left(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\right) \tag{4.2.4b}$$

Whereas, the posterior dispersion matrix of $\boldsymbol{\beta}$ is given by

$$V(\boldsymbol{\beta}|S) = (n + 2a - 2)^{-1}\left(2b + \boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}}\right)(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1} \tag{4.2.4c}$$

which is proportional to $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}$.

Similarly, the marginal posterior p.d.f. of τ will be derived by integration on the multivariate k-normal distribution part in (4.2.3b) with respect to $\boldsymbol{\beta}$ that gives a constant. It eventually leads to

$$p(\tau|S) \propto \tau^{\frac{n+2a}{2}-1} \exp\left\{-\frac{\tau}{2}\left[2b + \boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{A}\bar{\boldsymbol{\beta}}) + \bar{\boldsymbol{\beta}}'\boldsymbol{A}\bar{\boldsymbol{\beta}}\right]\right\} \tag{4.2.5a}$$

that is the gamma distribution with $r = \frac{n+2a}{2}$ and $\lambda = b + \frac{y'y-(X'y+A\bar{\beta})'(X'X+A)^{-1}(X'y+A\bar{\beta})+\bar{\beta}'A\bar{\beta}}{2}$. As shown in distribution III in appendix-I, the posterior mean of $\tau^{-1} = \sigma^2$ is given by

$$E(\tau^{-1}|S) = E(\sigma^2|S) = (n+2a-2)^{-1}\left(2b + y'y - (X'y + A\bar{\beta})'(X'X + A)^{-1}(X'y + A\bar{\beta}) + \bar{\beta}'A\bar{\beta}\right) \quad (4.2.5b)$$

whereas, the posterior variance is

$$V(\sigma^2|S) = (n+2a-2)^{-2}(n+2a-4)^{-1}\left(2b + y'y - (X'y + A\bar{\beta})'(X'X + A)^{-1}(X'y + A\bar{\beta}) + \bar{\beta}'A\bar{\beta}\right)^2 \quad (4.2.5c)$$

## 4.3. Concluding Remarks

From the above discussion, it is important to summarize some remarkable notes for the consequence of using the g-prior technique versus the natural conjugate one to assess the RIPs for the GLM parameters.

***First***, it is evident that, using the g-prior leads to a posterior covariance matrix of $\boldsymbol{\beta}$ proportional to $(X'X)^{-1}$. This is the main property that motivated Zellner to investigate an approach that lead to a natural conjugate prior distribution with prior precision proportional to $X'X$, so it is simple to be assessed. Since all what is required, to assess such a prior in (3.4.6); a prior mean vector $\boldsymbol{\beta}_a$, a prior mean for the error term variance $\bar{\sigma}_a^2$, and a choice of the value of g. The value g in this case, measures the amount of information in the prior relative to the sample, that is, setting g=0.1 gives the prior the same weight as 10% of the sample. Whereas, using the natural conjugate prior approach leads to posterior covariance matrix of $\boldsymbol{\beta}$ proportional to $(X'X + A)^{-1}$, see (4.2.4c). The posterior covariate structure is thus completely determined by the prior, by evaluating the elements of the matrix ***A,*** and the design matrix.

***Second***, one can notice that the g-prior distribution is a special case of the natural conjugate prior one. Where the joint g-prior in (3.4.6) is in the natural conjugate form (4.2.1) with $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}_a$, $A = gX'X$, $a = \nu$, and $2b = \nu\bar{\sigma}_a^2$. Thus, using g-prior reduces the choice of the $k \times k$ prior covariance matrix $A$ to a single scalar hyperparameter g.

***Third***, the posterior mean $\bar{\bar{\boldsymbol{\beta}}}$, produced by using the g-prior, form (4.1.3b) is just a simple average of $\hat{\boldsymbol{\beta}}$, the least squares estimate, and $\bar{\boldsymbol{\beta}}$, the prior mean vector, with the

parameter *g* involved in the weights. Another noticeable remark on (4.1.3b) that as g getting small $\bar{\bar{\boldsymbol{\beta}}} \approx \hat{\boldsymbol{\beta}}$, the LS quantity, while as *g* is large $\bar{\bar{\boldsymbol{\beta}}} \approx \bar{\boldsymbol{\beta}}$, the prior mean vector. On the other hand, following the natural conjugate technique leads to a posterior mean $\bar{\bar{\boldsymbol{\beta}}}$, in (4.2.4b), that can be viewed as a "shrinkage" estimate. Shrinkage phenomenon arises naturally in many Bayesian analysis, in the sense that, the influence of the prior distribution is to "pull" the likelihood towards the prior, and hence the posterior estimate can often be seen in terms of a classical estimate being pulled towards the prior estimate. The shrinkage phenomenon is not only a property of estimates such a posterior mean but also it affects many other posterior aspects. Shrinkage is also common in hierarchical models, for more details see O'Hagan (1994)

Shrinkage also obtained through the "ridge regression", where the ridge estimator is given by $\boldsymbol{b}(k) = (k\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ will be identical to the posterior mean $\bar{\bar{\boldsymbol{\beta}}}$ given by (4.2.4b) by setting $\mathbf{A} = k\mathbf{I}$ and $\bar{\boldsymbol{\beta}} = \mathbf{0}$. So on the algebraic level there is a close similarity between Bayesian analysis to the GLM using the natural conjugate prior and the ridge regression. However, in contrast to ridge regression where shrinkage is toward zero, the Bayes estimate shrinks toward the prior mean, Karlsson (2001). See Birkes and Dodge (1993) for more details about ridge regression.

# *Chapter 5*

# *Bayesian Time Series: AR(1) Models*

## 5.1. Introduction

A time series is a sequence of numerical data in which each item is associated with a particular time. Univariate time series is a single sequence of data such as monthly unemployment and daily closing prices of stock indices. Whereas, multivariate time series consists of several sets of data for the same sequence of period, such as, monthly unemployment, price levels, and monthly income that are considered over a certain period. One broad technique of analyzing time series is the "time-domain" methods, where they are based on direct modeling of the lagged relationships between a series and its past. Such a modeling technique, theoretically, views a time series as a stochastic process and regards an observed series as a particular or single "realization" of that process.

On a further clarification, suppose the stochastic process $\{Y_t\}$ of T-dimension is a set of autocorrelated random variables. A sample of size 1 of each random variable is hence drawn to form an observed time series. Thus, the observed time series is regarded as a realization of a stochastic process and there is no way to have another observation of each variable that is why it is called "single". These two features, dependence and lack of replication, enforce statisticians to specify some restrictive models for the statistical structure of that type of stochastic process (Maddala, 1988). Stochastic process can be described generally by a T-dimensional probability distribution $p(y_1, y_2, ..., y_T)$, so that the relationship between a realization and a stochastic process is parallel to that between the sample and population in classical approach (Mills, 1990). Instead of capturing a complete form of probability distribution to identify the stochastic process that generate the time series, one can concentrate on the two moments beside the covariance statistics of that distribution.

Classical approach undertakes the same view to analyze time series. According to classical prospect, statistical inference about parameters is explained using repeated sample concept under the same conditions. Practitioners usually do not accept this concept especially in fields such as economic, engineering and environment, whereas it is impossible to obtain another realization at the same time points as just mentioned previously. Hence, a non-classical approach that overcomes the need to repetition and avoids learning the large number of sample theory techniques as well, is required in time series analysis. Such approach is exactly the Bayesian technique, which gives an acceptable interpretation for point estimation, confidence intervals construction, tests of hypothesis, and predictions that are requested by many researchers in various fields.

In general, reasons of involving Bayesian approach in time series analysis are as follows:

1. This approach can assimilate new information different from that one used in the original analysis, so results can always be updated.
2. This approach can successfully give logical interpretation for statistical inferences in time series analysis, especially for constructing confidence intervals.
3. Experience plays an important role as a source of information in economic time series and other fields.

Nevertheless, adopting such an approach encounters some obstacles due to the adherent complicated nature of most of time series models, since the likelihood function is analytically intractable for the majority of ARMA models. That is, due to its nonlinearity which leads to problems with complicated posterior computations.

For simplicity, the current work will mainly focus on the linear autoregressive time series models, where the likelihood function will produce analytically tractable posterior distributions. Hence, a complete Bayesian analysis is possible.

Section 5.2 will focus on representing some basic concepts of the first-order autoregressive time series model, AR(1). Whereas section 5.3 is devoted to develop the posterior analysis of AR(1) model using some noninformative and informative priors that have been represented in the preceding two chapters.

A remarkable comparative study is introduced through section 5.4 to investigate the performance of the studied prior distributions based on simulation tools for the AR(1) process. Section 5.5 presents the posterior analysis of some real time series data sets for AR(1) model.

## 5.2. AR(1) models: Basic concepts

Suppose the discrete stochastic process $\{Y_t\}$ that is given by

$$y_t = \phi y_{t-1} + \varepsilon_t \tag{5.2.1}$$

where $\{\varepsilon_t\}$ is the white noise process, which is purely random process that is a sequence of mutually independent identically (i.i.d.) normally distributed with zero mean and common variance $\sigma^2$, i.e.,

$\mathrm{E}(\varepsilon_t) = 0,$
$\mathrm{Var}(\varepsilon_t) = \sigma^2$, and
$\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-k}) = 0, \text{for all} k \neq 0.$

The model (5.2.1) is called autoregressive model of order one. A main characteristic of that model, by using the Wold's decomposition, that it can be expressed as a "linear filter" of a sequence of white noise process. That can be shown as follows:

$$\begin{aligned} y_t &= \varepsilon_t + \phi y_{t-1} \\ &= \varepsilon_t + \phi(\varepsilon_{t-1} + \phi y_{t-2}) \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2(\varepsilon_{t-2} + \phi y_{t-3}) \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3(\varepsilon_{t-3} + \phi y_{t-4}) \\ &\quad\vdots \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \cdots + \phi^k\varepsilon_{t-k} + \ldots \end{aligned}$$

Finally, the process $\{Y_t\}$ can be written as

$$y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}, \qquad \text{where } j = 1,2,... \tag{5.2.2}$$

The last infinite series in (5.2.2) is also called the "*infinite moving average process"* and is denoted by MA(∞). Using this transformation, taking into consideration of the information about the white noise process, leads to the following derivation of the first two moments of AR(1)

$$\mathrm{E}(y_t) = \mathrm{E}(\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j})$$

$$= \sum_{j=0}^{\infty} \phi^j \mathrm{E}(\varepsilon_{t-j}) = \text{zero}$$

While the variance of the process $\{Y_t\}$, $\gamma_0$, is give by:

$$\begin{aligned}\gamma_0 &= \mathrm{Var}(y_t) \\ &= \mathrm{Var}(\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}) \\ &= \sum_{j=0}^{\infty} \phi^{2j} \mathrm{Var}(\varepsilon_{t-j}) + 2\Sigma_{i\neq j} \Sigma \phi^i \phi^j \mathrm{Cov}(\varepsilon_{t-i}\varepsilon_{t-j}) \\ &= \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} \qquad (5.2.3)\end{aligned}$$

Similarly, the covariance function at lag *k*, $\gamma_k$ is given by

$$\begin{aligned}\gamma_k &= \mathrm{Cov}(y_t, y_{t-k}) \\ &= \mathrm{E}(y_t, y_{t-k}) \\ &= \mathrm{E}\left(\begin{matrix}\left(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + .... + \phi^k\varepsilon_{t-k} + \phi^{k+1}\varepsilon_{t-k-1} + \phi^{k+2}\varepsilon_{t-k-2} + ...\right)\times \\ \left(\varepsilon_{t-k} + \phi\varepsilon_{t-k-1} + \phi^2\varepsilon_{t-k-2} + ...\right)\end{matrix}\right) \\ &= \phi^k \mathrm{E}(\varepsilon_{t-k}) + \phi^{k+2} \mathrm{E}(\varepsilon_{t-k-1}) + \phi^{k+4} \mathrm{E}(\varepsilon_{t-k-2}) + ... \\ &= \sigma^2 \sum_{j=0}^{\infty} \phi^{2j+k} \qquad (5.2.4)\end{aligned}$$

Consequently, the autocorrelation function at lag *k* can be given by

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\sigma^2 \sum_{j=0}^{\infty} \phi^{2j+k}}{\sigma^2 \sum_{j=0}^{\infty} \phi^{2j}},$$

Since $\sigma^2$ is a non-zero positive quantity, then,

$$\rho_k = \frac{\sum_{j=0}^{\infty} \phi^{2j+k}}{\sum_{j=0}^{\infty} \phi^{2j}} \qquad (5.2.5)$$

It is quite important to notice that the sequences of $\varepsilon$'s for $y_t$ as shown in (5.2.2) will accumulate rather than die out if $\Sigma|\phi^j| \to \infty$, which is equivalent to $|\phi| \geq 1$. Consequently, all the moments of $y_t$ given through (5.2.3) to (5.2.5) will not exist.

However, in the case when $\sum|\phi^j|<\infty$ and hence $|\phi|<1$ that is the weights converge absolutely. That condition is equivalent to assuming that the stochastic process $\{Y_t\}$ is *covariance stationary*, which guarantees that all moments exist and are independent of time, particularly, the variance $\gamma_0$ is finite. Moreover, under that condition, the covariance between $y_t$ and $y_{t-k}$ depends only on the lag *k*, the length of the time separating observations and not on the time itself (Mills, 1990). Thus, for any stationary covariance stationary process $\gamma_j=\gamma_{-j}$ for all integer *j*. That is called *weak stationarity*. Furthermore, the covariance stationary Gaussian AR(1) process is *strictly stationary*, since the latter definition requires that all joint distributions of any subset of the time series are unaffected by a change of time origin , however, they just depend on the lags. According to the AR(1) process, only the first two moments are needed to identify the distribution completely. That is why for such a process weak stationarity is equivalent to strict stationarity.

Given the assumption of stationarity the equations through (5.2.3) to (5.2.5) can be simplified as follows

$$\begin{aligned}\gamma_0 &= \sigma^2 \sum_{j=0}^{\infty}\phi^{2j} \\ &= \sigma^2(1+\phi^2+\phi^4+\phi^6+\cdots)\end{aligned}$$

The right hand side of the above equation is an infinite geometric series with base $|\phi|<1$, so the variance of stationary AR(1) process is given by

$$\gamma_0 = \frac{\sigma^2}{1-\phi^2} \tag{5.2.6}$$

Similarly, the covariance function can be derived as follows

$$\begin{aligned}\gamma_k &= \sigma^2 \sum_{j=0}^{\infty}\phi^{2j+k} \\ &= \phi^k \sum_{j=0}^{\infty}\sigma^2\phi^{2j},\end{aligned}$$

then,

$$\begin{aligned}\gamma_k &= \phi^k \gamma_0 \\ &= \sigma^2 \frac{\phi^k}{1-\phi^2}.\end{aligned} \tag{5.2.7}$$

Similarly, the autocorrelation function is given by

$$\begin{aligned}\rho_k &= \frac{\gamma_k}{\gamma_0} \\ &= \frac{\phi^k \gamma_0}{\gamma_0} \\ &= \phi^k\end{aligned} \tag{5.2.8}$$

Stationary time series is also called non-explosive time series, whereas, non-stationary time series is described as explosive.

Techniques of time series analysis are not confined to the analysis of stationary or non-explosive time series. Pragmatically, most of the time series encountered are nonstationary. However, some transformations could be applied to achieve stationarity such as taking difference of successive orders until achieving stationarity. Stationarity is beneficial in reducing the number of parameters of the investigated model. However, the current study will focus on the posterior analysis of AR(1) when stationarity is not assumed as will be discussed below.

## 5.3. Posterior Analysis of AR(1) Models

Autoregressive (AR) models are regularly used for the analysis of time series data. Bayesian analysis of AR models began with the early work of Zellner and Tiao (1964) who considered the AR(1) process. Bayesian analysis of higher order of AR model are given in Zellner (1971). Lahiff (1980) developed a numerical algorithm to produce posterior and predictive analysis for AR(1) process. Diaz and Farah (1981) devoted a Bayesian technique for computing posterior analysis of AR process with an arbitrary order. Broemling (1985) adapt many types of AR models discussed in literature in a very general framework. Philips (1991) discussed the use of different prior distribution to develop the posterior analysis of AR models with fitted trend with no stationarity

assumption assumed. Koop *et al.* (1995) investigated the effect of the prior distribution choices on the prediction particularly when stationarity condition is imposed. Ghosh and Heo (2000) introduced a comparative study to some selected noninformative priors for the AR(1) models.

In this section, the posterior analysis to the AR(1) model, in (5.2.1), is developed using some of selected noninformative and informative priors that have been introduced in the current thesis. Such development will be carried out only for the general case when no attention to the stationarity condition, this case would be applicable for stationary or nonstationary time series.

### 5.3.1. Based on Noninformative Priors

The AR(1) process is generated using the formula in (5.2.1) that is

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad t = 1,2,...,T$$

where $\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$ for all *t=1,2,3,...,T.* the parameters $\phi$ and $\sigma$ are unknown parameters such that $-\infty < \phi < \infty$ and $0 < \sigma^2 < \infty$. In addition, $y_0$ is an initial observation assumed to be known constant. Note that there is no restriction for the autoregressive coefficient $\phi$ to be within the stationary interval -1 and +1. The probability density function of $y_t$ is given by

$$f(y_t \mid \phi, \sigma, y_0) \propto \sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}(y_t - \phi y_{t-1})^2\right\}, \quad t = 1,2,...,T \text{ and } y, \phi \in (-\infty,\infty), \sigma \in (0,\infty). \qquad (5.3.1)$$

The likelihood function of the parameters $\phi$ and $\sigma$ given the observations is given by

$$l(\phi,\sigma \mid y_0, y_1, y_2, ....y_T) \propto \sigma^{-T} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2\right\} \qquad (5.3.2)$$

Consider $\boldsymbol{\theta}' = (\phi \ \ \sigma)$ is the vector of parameters and $\boldsymbol{y}' = (y_1 y_2 ....y_T)$ is the vector of observed data of length *T*. The issue now is to derive the noninformative priors according to the techniques introduced in chapter 2 that would be as follows:

***Jeffreys' Prior***

Applying Jeffreys' general rule given by (2.3.5) and (2.3.7) but in terms of the likelihood function will lead to the following results

$$\mathrm{Inf}_{\boldsymbol{\theta}} \propto -E_{\mathbf{y}|\,\boldsymbol{\theta}}\left[\frac{\partial^2 \log l(\phi,\sigma|\,\mathrm{y})}{\partial\theta_i \partial\theta_j}\right],\ i,j=1,2$$

The logarithm of the likelihood function is given by

$$\log l(\phi,\sigma|\mathbf{y}) \propto -T\log\sigma - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-\phi\, y_{t-1})^2,$$

then

$$\mathrm{Inf}_{\phi,\sigma} \propto -E_{\boldsymbol{y}|\,\phi,\sigma}\begin{pmatrix} \dfrac{\partial^2 \log l(\phi,\sigma|\boldsymbol{y})}{\partial\phi^2} & \dfrac{\partial^2 \log l(\phi,\sigma|\boldsymbol{y})}{\partial\sigma\partial\phi} \\ \dfrac{\partial^2 \log l(\phi,\sigma|\boldsymbol{y})}{\partial\sigma\partial\phi} & \dfrac{\partial^2 \log l(\phi,\sigma|\boldsymbol{y})}{\partial\sigma^2} \end{pmatrix},$$

$$\propto -E_{\boldsymbol{y}|\,\phi,\sigma}\begin{pmatrix} \dfrac{-\sum_{t-1}^{T} y_{t-1}^2}{\sigma^2} & \dfrac{-2\sum_{t-1}^{T}(y_t-\phi y_{t-1})(y_{t-1})}{\sigma^3} \\ \dfrac{-2\sum_{t-1}^{T}(y_t-\phi y_{t-1})(y_{t-1})}{\sigma^3} & \dfrac{T}{\sigma^2}-\dfrac{3\sum_{t=1}^{T}(y_t-\phi\, y_{t-1})^2}{\sigma^4} \end{pmatrix},$$

Since $E_{\mathbf{y}|\,\phi,\sigma}(y_t-\phi y_{t-1})=0$ and $E_{\mathbf{y}|\,\phi,\sigma}(y_t-\phi y_{t-1})^2=\sigma^2$, then the Fisher's information matrix could be simplified to

$$\mathrm{Inf}_{\phi,\sigma} \propto \begin{pmatrix} \dfrac{-\sum_{t-1}^{T} y_{t-1}^2}{\sigma^2} & 0 \\ 0 & \dfrac{-2T}{\sigma^2} \end{pmatrix},$$

Jeffreys' prior, hence, using the formula $p(\phi,\sigma) \propto \sqrt{|\,\mathrm{Inf}_{\phi,\sigma}\,|}$, will be in the form

$$p(\phi,\sigma) \propto \sqrt{\sigma^{-4}},$$

then

$$p(\phi,\sigma) \propto \sigma^{-2}$$

This prior distribution is refused by Jeffreys as mentioned before in §2.3, therefore, Jeffreys assumed independence between the autoregressive coefficient $\phi$ and the scale parameter $\sigma$. This assumption leads to the independence rule given by (3.2.8). Hence, the Jeffreys' prior distribution of $\phi$ and $\sigma$ is given by

$$p(\phi,\sigma) \propto \sigma^{-1} \qquad (5.3.3)$$

***Locally Uniform Prior***

Box and Taio (1973) considered the locally uniform prior based on the data translated likelihood concept. As discussed earlier in §2.4, the data translated likelihood is the one takes the form in (2.4.11) that is

$$l(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto g[\boldsymbol{\eta}(\boldsymbol{\theta}) - \mathrm{f}(\mathbf{y})],$$

where g(.) is a known function independent of $\mathbf{y}$, $\boldsymbol{\eta}' = (\eta_1\ \eta_2)$, is a vector of order 2 that is one-to-one transformation of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}' = (\phi\ \sigma)$, and $[\mathrm{f}(\boldsymbol{y})]' = [f_1(\boldsymbol{y})\ f_2(\boldsymbol{y})]$ is a vector of 2 functions of $\mathbf{y}$. The locally uniform distribution is taken as a noninformative prior for $\boldsymbol{\eta}$, then the corresponding noninformative prior of $\boldsymbol{\theta}$ is given by

$$p(\boldsymbol{\theta}) \propto |J|,$$

where

$$|J| = \begin{vmatrix} \dfrac{\partial \eta_1}{\partial \phi} & \dfrac{\partial \eta_1}{\partial \sigma} \\ \dfrac{\partial \eta_2}{\partial \phi} & \dfrac{\partial \eta_2}{\partial \sigma} \end{vmatrix}$$

The concern now is to try to rewrite the likelihood function, in (5.3.2), in the form of translated likelihood function given by (2.4.11). Thus, it will be helpful to consider the following quantity

$$\begin{aligned} \Sigma (y_t - \phi\, y_{t-1})^2 &= \Sigma [(y_t - \bar{y}) - (\phi\, y_{t-1} - \bar{y})]^2, \\ &= (T-1)s^2 + T(\phi\, y_{t-1} - \bar{y})^2, \end{aligned}$$

where $s^2 = \dfrac{\sum_{t=1}^{T} (y_t - \bar{y})^2}{T-1}$. Then the likelihood function could be expressed as

$$l(\phi, \sigma\,|\,\mathbf{y}) \propto \sigma^{-T} \exp\left\{ -\frac{T(\phi\, y_{t-1} - \bar{y})^2}{2\sigma^2} - \frac{(T-1)s^2}{2\sigma^2} \right\},$$

Multiplying the last form by $s^T$, where multiplication of likelihood by constant leaves it unchanged, then

$$l(\phi, \sigma\,|\,\mathbf{y}) \propto \left(\frac{s}{\sigma}\right)^T \exp\left\{ -\frac{T(\phi\, y_{t-1} - \bar{y})^2}{2s^2} \left(\frac{s^2}{\sigma^2}\right) - \frac{(T-1)s^2}{2\sigma^2} \right\},$$

can be written as

$$l(\phi, \sigma\,|\,\boldsymbol{y}) \propto \left(\frac{\sigma}{s}\right)^{-T} \exp\left\{ -\frac{T}{2} \left( \frac{\phi\, y_{t-1} - \bar{y}}{s} \right)^2 \left(\frac{\sigma}{s}\right)^{-2} \right\} \exp\left\{ -\frac{(T-1)}{2} \left(\frac{\sigma}{s}\right)^{-2} \right\},$$

$$l(\phi,\sigma|\,\mathbf{y}) \propto \exp\left\{-\frac{T}{2}\left(\frac{\phi\, y_{t-1}-\bar{y}}{s}\right)^2 \exp[-2\log\left(\frac{\sigma}{s}\right)]\right\} \times \exp\left\{-T\log\left(\frac{\sigma}{s}\right)-\left(\frac{T-1}{2}\right)\exp[-2\log\left(\frac{\sigma}{s}\right)]\right\}.$$

Eventually, the likelihood function can be given by the following form:

$$l(\phi,\sigma|\,\mathbf{y}) \propto \exp\left\{-\frac{T}{2}\left(\frac{\phi\, y_{t-1}-\bar{y}}{s}\right)^2 \exp[-2(\log\sigma-\log s)]\right\} \times$$
$$\exp\left\{-T(\log\sigma-\log s)-\left(\frac{T-1}{2}\right)\exp[-2(\log\sigma-\log s)]\right\}$$

This last form could be considered as a translation to the form in (2.4.11) such that $\boldsymbol{\eta}' \propto (\phi\, y_{t-1}\ \log\sigma)$ and $[\mathrm{f}(\boldsymbol{y})]' = [\bar{y}\ \log s]$. Then, one may take the locally uniform distribution as a noninformative prior for $\boldsymbol{\eta}$, hence the corresponding joint noninformative prior distribution for $\theta$ and $\sigma$ is given by

$$p(\phi,\sigma) \propto |J|,$$
$$\propto \begin{vmatrix} \frac{\partial}{\partial\phi}(\phi y_{t-1}) & \frac{\partial}{\partial\sigma}(\phi y_{t-1}) \\ \frac{\partial}{\partial\phi}(\log\sigma) & \frac{\partial}{\partial\sigma}(\log\sigma) \end{vmatrix},$$
$$\propto \begin{vmatrix} y_{t-1} & 0 \\ 0 & \sigma^{-1} \end{vmatrix},$$

then

$$p(\phi,\sigma) \propto \sigma^{-1},$$

This entirely agrees with the form of Jeffreys' prior given by (5.3.3).

***Maximal Data Information Prior***

Referring to §2.5, the MDIP for the parameters of AR(1) process could be given using the multiparameter version of equation (2.5.7), since the MDIP of $\boldsymbol{\theta}' = (\phi\ \sigma)$ depends on the quantity $I_{\mathbf{y}}(\boldsymbol{\theta})$ computed by (2.5.5). Then, it can be proved that,

$$I_{y_t}(\theta) = \int_{R_{y_t}} f(y_t|\,\theta)\ \ln f(y_t|\,\theta)\, dy_t,$$
$$= \log\left(\sigma^{-1}\right) - \log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}$$

Eventually, the MDIP, using the above measure of information in the sample and the form in (2.5.7), will be given as follows:

$$P^*(\phi,\sigma) \propto \exp\left\{\log\left(\sigma^{-1}\right) - \log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\right\}$$

Then, finally the joint MDIP of $\phi$ and $\sigma$ is in the following form

$$P^*(\phi,\sigma) \propto \sigma^{-1},$$

which is yet again the Jeffreys' prior given by (5.3.3). Thus, all approaches of noninformative priors studied by the current work have unanimity the form of Jeffreys' prior in case when no restriction assumed to the stationarity of AR(1) process. That result emphasizes the outstanding substance of Jeffreys' prior to be wide applicable. Moreover, different philosophies to noninformative elicitation in literature end up with the Jeffreys' prior.

***Posterior Analysis of AR(1)***

The joint posterior distribution of $\phi$ and $\sigma$ will be obtained by combining the prior distribution with the likelihood function. First, it will be helpful to simplify the quantity $\sum(y_t - \phi y_{t-1})^2$, in the exponent of the likelihood function, by completing the square with respect to $\phi$ then with respect to $y_t$. We obtain,

$$\sum(y_t - \phi y_{t-1})^2 = \sum y_t^2 + \sum y_{t-1}^2\left(\phi^2 - 2\phi\frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2}\right),$$

Consider $\hat{\beta} = \dfrac{\sum y_t y_{t-1}}{\sum y_{t-1}^2}$, which is the ordinary least square (OLS) estimate for the simple linear regression model since $y_{t-1}$ is viewed as a regressor for the dependent variable $y_t$. The above equation will be reduced to

$$\sum(y_t - \phi y_{t-1})^2 = \sum\left(y_t - \hat{\beta} y_{t-1}\right)^2 + \left(\phi - \hat{\beta}\right)^2 \sum y_{t-1}^2,$$

Consider $\nu q^2 = \sum\left(y_t - \hat{\beta} y_{t-1}\right)^2$, where $\nu = T - 1$, the likelihood function in (5.3.2) could be written as

$$l(\phi,\sigma|y_0,\boldsymbol{y}) \propto \sigma^{-T}\exp\left\{-\frac{1}{2\sigma^2}[\nu q^2 + \left(\phi - \hat{\beta}\right)^2 \sum y_{t-1}^2\right\} \tag{5.3.4}$$

Combining the likelihood function in (5.3.4) with the joint prior distribution in (5.3.3) will lead to the following joint posterior distribution of $\phi$ and $\sigma$

$$p(\phi,\sigma|y_0,\boldsymbol{y}) \propto \sigma^{-(T+1)}\exp\left\{-\frac{1}{2\sigma^2}[\nu q^2 + \left(\phi - \hat{\beta}\right)^2 \sum y_{t-1}^2]\right\} \tag{5.3.5a}$$

The above form is just the normal inverted-gamma distribution, which can also be written as

$$p(\phi,\sigma|y_0,\boldsymbol{y}) \propto \sigma^{-(\nu+1)}\exp\left\{-\frac{1}{2\sigma^2}\nu q^2\right\} \times \sigma^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\phi - \hat{\beta}\right)^2 \sum y_{t-1}^2\right\} \tag{5.3.5b}$$

Thus, to obtain the marginal posterior distribution of $\phi$, (5.3.5a) is integrated with respect to $\sigma$ gives the inverse of the kernel of I Γ-II (Inverted Gamma-2) distribution with parameters $\left(r = T, r\lambda^2 = \nu q^2 + (\phi - \hat{\beta})^2 \sum y_{t-1}^2\right)$. See Appendix-I for more details about that distribution. Eventually, the posterior p.d.f. of $\phi$ is given by

$$p(\phi | y_0, \boldsymbol{y}) \propto \left[1 + \frac{(\phi - \hat{\beta})^2 \sum y_{t-1}^2}{\nu q^2}\right]^{\frac{-(\nu+1)}{2}}, \quad \phi \in (-\infty, \infty) \tag{5.3.6a}$$

which is obviously the univariate t distribution with $\nu$ degrees of freedom. The posterior mean is given by $\hat{\beta}$ and the posterior variance equals to $\frac{\nu q^2}{(\nu-2)\sum y_{t-1}^2}$.

Similarly, the marginal posterior p.d.f. of $\sigma$ can be obtained by integrating (5.3.5b) with respect to $\phi$, that is integrating the normal distribution part in (5.3.5b) which gives constant. Hence, the marginal distribution of $\sigma$ is given by

$$p(\sigma | y_0, \boldsymbol{y}) \propto \sigma^{-(\nu+1)} \exp\left\{-\frac{1}{2\sigma^2}\nu q^2\right\}, \quad \sigma > 0 \tag{5.3.6b}$$

is the I Γ-II distribution with parameters $\left(r = \nu, \lambda^2 = q^2\right)$. The posterior mean and the posterior variance of $\sigma$ are given by $q\sqrt{\frac{\nu}{2}}\frac{(\frac{\nu}{2}-\frac{3}{2})!}{\Gamma(\frac{\nu}{2})}$ and $q^2\frac{\nu}{\nu-2} - \left(q\sqrt{\frac{\nu}{2}}\frac{(\frac{\nu}{2}-\frac{3}{2})!}{\Gamma(\frac{\nu}{2})}\right)^2$ respectively, where $\nu > 2$.

A considerable note that, the posterior analysis introduced above for AR(1) model, without restriction to stationarity, is identical to that of simple linear regression model.

### 5.3.2. Based on Informative Priors

In this section, the posterior analysis to the AR(1) model will be developed using the informative priors introduced in chapter 3. These informative priors are the Natural Conjugate prior and g-prior. The development will also be confined to the general case where no stationarity assumption is imposed. The posterior analysis here is similar to that introduced in chapter 4 to the GLM, however matrix representation will not be used and the derivations will be executed in terms of the standard deviation $\sigma$ not in terms of the precision $\tau = \frac{1}{\sigma^2}$.

### *1. Posterior Analysis of AR(1) Using Natural Conjugate Prior*

This approach involves starting with a model p.d.f., then selecting a prior distribution from a class leading to a posterior distribution belonging to the same class. The likelihood function of the AR(1) process which has been presented in (5.3.4) which could also be expressed by

$$l(\phi,\sigma|y_0,\mathbf{y}) \propto \sigma^{-(T-1)} \exp\left\{-\frac{\nu s^2}{2\sigma^2}\right\} \times \sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}\left(\phi-\hat{\beta}\right)^2 \sum y_{t-1}^2\right\}$$

This distribution belongs to the normal inverted-gamma class. Accordingly, the natural conjugate prior of the parameters $\phi$ and $\sigma$ is supposed to belong to the same class defined by

$$p(\sigma,\phi) \propto p_{IG}(\sigma) \times p_N(\phi|\sigma)$$

where,

$$p_{IG}(\sigma) \propto \sigma^{-(r+1)} \exp\left\{\frac{-r\lambda^2}{2\sigma^2}\right\}, \sigma, r, \lambda > 0$$

is the Inverted Gamma of type II with parameters $r$ and $\lambda$, as shown in Appendix-I, while

$$p_N(\phi|\sigma) \propto \sigma^{-1} \exp\left\{\frac{-h^2}{2\sigma^2}(\phi-\mu)^2\right\}, \phi, \mu \in (-\infty,\infty), \sigma, h > 0$$

is the Normal distribution with parameters $\mu$ and $\sigma^2 h^{-2}$. Hence, the joint prior distribution of $\phi$ and $\sigma$ given by

$$p(\phi,\sigma) \propto \sigma^{-(r+2)} \exp\left\{\frac{-1}{2\sigma^2}\left[r\lambda^2+(\phi-\mu)^2 h^2\right]\right\}, \phi, \mu \in (-\infty,\infty), \sigma > 0 \qquad (5.3.7)$$

Combining that joint prior distribution with the likelihood function, in (5.3.2), implies the following posterior distribution,

$$p(\phi,\sigma|y_0,\mathbf{y}) \propto \sigma^{-(T+r+2)} \exp\left\{-\frac{1}{2\sigma^2}\left(r\lambda^2+(\phi-\mu)^2 h^2+\sum_{t=1}^{T}(y_t-\phi y_{t-1})^2\right)\right\},$$

which can be written as

$$p(\phi,\sigma|y_0,\mathbf{y}) \propto \sigma^{-(T+r+2)} \exp\left\{-\frac{r\lambda^2}{2\sigma^2}\right\} \times \exp\left\{-\frac{1}{2\sigma^2}\left((\phi-\mu)^2 h^2+\sum_{t=1}^{T}(y_t-\phi y_{t-1})^2\right)\right\}$$

Consider the quadratic form in the second exponent of the above distribution denoted by *Q*, which can be simplified to

$$Q=(\phi-\mu)^2h^2+\sum_{t=1}^{T}(y_t-\phi y_{t-1})^2$$

$$=\mu^2h^2+\sum y_t^2-\left(\sum y_t y_{t-1}+\mu h^2\right)^2\left(\sum y_{t-1}^2+h^2\right)^{-1}+\left[\phi-\left(\sum y_t y_{t-1}+\mu h^2\right)\left(\sum y_{t-1}^2+h^2\right)^{-1}\right]^2\left(\sum y_{t-1}^2+h^2\right)$$

Consider $\hat{\phi}=\left(\sum y_t y_{t-1}+\mu h^2\right)\left(\sum y_{t-1}^2+h^2\right)^{-1}$, then the quantity $Q$ could be shortened to

$$Q=\mu^2h^2+\sum y_t^2-\hat{\phi}^2\left(\sum y_{t-1}^2+h^2\right)+\left(\phi-\hat{\phi}\right)^2\left(\sum y_{t-1}^2+h^2\right).$$

Then, the joint posterior distribution could finally be simplified to

$$p(\phi,\sigma|y_0,\boldsymbol{y})\propto\sigma^{-(T+r+2)}\exp\left\{-\frac{1}{2\sigma^2}\left[r\lambda^2+\sum y_t^2-\hat{\phi}^2\left(\sum y_{t-1}^2+h^2\right)+\mu^2h^2+\left(\phi-\hat{\phi}\right)^2\left(\sum y_{t-1}^2+h^2\right)\right]\right\} \quad (5.3.8a)$$

That can also be represented by

$$p(\phi,\sigma|y_0,\boldsymbol{y})\propto\sigma^{-(T+r+1)}\exp\left\{-\frac{1}{2\sigma^2}\left[r\lambda^2+\sum y_t^2-\hat{\phi}^2\left(\sum y_{t-1}^2+h^2\right)+\mu^2h^2\right]\right\}\times$$
$$\sigma^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\phi-\hat{\phi}\right)^2\left(\sum y_{t-1}^2+h^2\right)\right\} \quad (5.3.8b)$$

This is again, the normal inverted-gamma distribution. Obtaining the marginal distribution of $\phi$ requires integrating (5.3.8a) with respect to $\sigma$. That is integrating an Inverted-Gamma distribution of type II with parameters $r$ and $\lambda$ such that $\left(r=T+r+1,\ r\lambda^2=r\lambda^2+\sum y_t^2-\hat{\phi}^2\left(\sum y_{t-1}^2+h^2\right)+\mu^2h^2+\left(\phi-\hat{\phi}\right)^2\left(\sum y_{t-1}^2+h^2\right)\right)$. Hence, the marginal posterior distribution of $\phi$ is given by

$$p(\phi|\,y_0,\boldsymbol{y})\propto\left[1+\frac{\left(\phi-\hat{\phi}\right)^2\left(\sum y_{t-1}^2+h^2\right)}{\nu w^2}\right]^{-\frac{(\nu+1)}{2}},\ -\infty<\phi<\infty, \quad (5.3.9a)$$

which is the univariate t distribution with $\nu$ degrees of freedom where $\nu w^2=r\lambda^2+\sum y_t^2-\hat{\phi}^2\left(\sum y_{t-1}^2+h^2\right)+\mu^2h^2$ and $\nu=T+r$. Then, the posterior mean of $\phi$ is given by $\hat{\phi}$, whereas, the posterior variance is given by $\frac{\nu}{\nu-2}w^2\left(\sum y_{t-1}^2+h^2\right)^{-1}$.

However, the marginal posterior distribution of $\sigma$ could be obtained by integrating the second part of (5.3.8b). Consequently, the marginal posterior distribution of $\sigma$ represented by the following form:

$$p(\sigma|y_0,\boldsymbol{y})\propto\sigma^{-(\nu+1)}\exp\left\{-\frac{\nu w^2}{2\sigma^2}\right\},\ \sigma>0 \quad (5.3.9b)$$

It is the inverted-gamma distribution with parameters $\left(r=\nu,\lambda^2=w^2\right)$. Hence, the posterior mean is given by $w\sqrt{\frac{\nu}{2}}\frac{(\frac{\nu}{2}-\frac{3}{2})!}{\Gamma(\frac{\nu}{2})}$ and the posterior variance equals to $w^2\frac{\nu}{\nu-2}-\left(w\sqrt{\frac{\nu}{2}}\frac{(\frac{\nu}{2}-\frac{3}{2})!}{\Gamma(\frac{\nu}{2})}\right)^2$.

### 2. ***Posterior Analysis of AR(1) Using g-Prior***

As previously explained in §4.1., the posterior analysis AR(1) using g-prior could be developed analogous to the GLM. The joint g-prior distribution of $\phi$ and $\sigma$ suggested by Zellner is given by the following relation

$$p_g(\phi,\sigma)\propto p(\sigma)\times p(\phi|\sigma,g),$$

where $p(\sigma)\propto 1/\sigma$ and $p(\phi|\sigma,g)\propto\sigma^{-1}\exp\left\{-\frac{g}{2\sigma^2}(\phi-\mu)^2\sum y_{t-1}^2\right\}$. That is to take the prior distribution of $\sigma$ as the Jeffreys' prior assigned by rule given by (2.3.3). Whereas, the conditional prior of $\phi$ given $\sigma$ is taking to be Normal distribution with prior mean $\mu$ and prior dispersion proportional to $g\left(\sum y_{t-1}^2\right)$. This prior variance is simply a product of unknown scalar g and quadratic known value that is based on the observations. It is noticeable remark that, this quadratic term is considered as a main component of the variance of the OLS estimate. This is obviously the main motivation to the g-prior in comparison with the natural conjugate one. This motivation is more evident in multiparameter case (see §3.4, for the use of the g-prior in such a case). Accordingly, the joint g-prior of $\phi$ and $\sigma$ is given by the following form

$$p_g(\phi,\sigma)\propto\sigma^{-2}\exp\left\{-\frac{g}{2\sigma^2}(\phi-\mu)^2\sum y_{t-1}^2\right\}, \tag{5.3.10}$$

Combining this prior distribution with the likelihood function in (5.3.2) implies the following joint posterior distribution:

$$p(\phi,\sigma|y_0,\mathbf{y},g)\propto\sigma^{-(T+2)}\exp\left\{-\frac{1}{2\sigma^2}\left(g(\phi-\mu)^2\sum y_{t-1}^2+\sum_{t=1}^{T}(y_t-\phi y_{t-1})^2\right)\right\}$$

Consider the quadratic quantity $Q$ between the braces in the exponent of the above form. Then , it can be proved that,

$$Q = g(\phi-\mu)^2 \Sigma y_{t-1}^2 + \sum_{t=1}^{T}(y_t - \phi y_{t-1})^2$$
$$= g\mu^2 \Sigma y_{t-1}^2 + \Sigma y_t^2 + \left(\phi - \left(\hat{\beta} + g\mu\right)(1+g)^{-1}\right)^2 (1+g)\Sigma y_{t-1}^2 - \left(\hat{\beta} + g\mu\right)^2 (1+g)^{-1} \Sigma y_{t-1}^2$$

where $\hat{\beta} = \dfrac{\Sigma y_t y_{t-1}}{\Sigma y_{t-1}^2}$ is the OLS estimate. Consider $\bar{\phi} = \left(\hat{\beta} + g\mu\right)(1+g)^{-1}$ then the quantity $Q$ could be reduced to

$$Q = g\mu^2 \Sigma y_{t-1}^2 + \Sigma y_t^2 + \left(\phi - \bar{\phi}\right)^2 (1+g)\Sigma y_{t-1}^2 - \bar{\phi}^2 (1+g)\Sigma y_{t-1}^2$$
$$= \left(g\mu^2 - \bar{\phi}^2 (1+g)\right)\Sigma y_{t-1}^2 + \Sigma y_t^2 + \left(\phi - \bar{\phi}\right)^2 (1+g)\Sigma y_{t-1}^2.$$

Then, the joint posterior distribution is given by

$$p(\phi, \sigma | y_0, \mathbf{y}, g) \propto \sigma^{-(T+2)} \exp\left\{-\frac{1}{2\sigma^2}\left(\left(g\mu^2 - \bar{\phi}^2 (1+g)\right)\Sigma y_{t-1}^2 + \Sigma y_t^2 + \left(\phi - \bar{\phi}\right)^2 (1+g)\Sigma y_{t-1}^2\right)\right\} \tag{5.3.11a}$$

That could be written as

$$p(\phi, \sigma | y_0, \boldsymbol{y}, g) \propto \sigma^{-(T+1)} \exp\left\{-\frac{1}{2\sigma^2}\left(g\mu^2 - \bar{\phi}^2 (1+g)\right)\Sigma y_{t-1}^2 + \Sigma y_t^2\right\} \times$$
$$\sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}\left(\phi - \bar{\phi}\right)^2 (1+g)\Sigma y_{t-1}^2\right\} \tag{5.3.11b}$$

This is again, the normal inverted-gamma distribution. Hence, the marginal posterior distribution of $\phi$ is given by:

$$p(\phi | \sigma, y_0, \boldsymbol{y}) \propto \left[1 + \frac{\left(\phi - \bar{\phi}\right)^2 (1+g)\Sigma y_{t-1}^2}{\nu w^2}\right]^{-\frac{(\nu+1)}{2}}, \quad -\infty < \phi < \infty, \tag{5.3.12a}$$

which is the univariate t distribution with $\nu$ degrees of freedom where $\nu w^2 = \left(g\mu^2 - \bar{\phi}^2 (1+g)\right)\Sigma y_{t-1}^2 + \Sigma y_t^2$ and $\nu = T$. Then, the posterior mean of $\phi$ is given by $\bar{\phi} = \left(\hat{\beta} + g\mu\right)(1+g)^{-1}$ whereas, the posterior variance is given by $\dfrac{\nu}{\nu-2} w^2 \left((1+g)\Sigma y_{t-1}^2\right)^{-1}$.

However, the marginal posterior distribution of $\sigma$ could be obtained by integrating the second part of (5.3.8b). Consequently, the marginal posterior distribution of $\sigma$ is represented by the following form:

$$p(\sigma | y_0, \boldsymbol{y}) \propto \sigma^{-(\nu+1)} \exp\left\{-\frac{\nu w^2}{2\sigma^2}\right\}, \ \sigma > 0, \tag{5.3.12b}$$

which is the Inverted-Gamma distribution with parameters $\left(r=\nu,\lambda^2=w^2\right)$. Hence, the posterior mean is given by $w\sqrt{\frac{\nu}{2}}\frac{(\frac{\nu}{2}-\frac{3}{2})!}{\Gamma(\frac{\nu}{2})}$ and the posterior variance equals $w^2\frac{\nu}{\nu-2}-\left(w\sqrt{\frac{\nu}{2}}\frac{(\frac{\nu}{2}-\frac{3}{2})!}{\Gamma(\frac{\nu}{2})}\right)^2$.

One of the main objectives of the thesis is to compare the efficiency of the studied prior distributions for AR(1) process. So far, the study presents three candidate priors, which can be summarized throughout the following table:

**Table 5.1: Candidate Prior distributions for the AR(1) Process**

| Prior Name | Prior Distribution Form |
|---|---|
| **Jeffreys' Prior** | $P_J(\phi,\sigma)\propto 1/\sigma$, $\sigma>0$ *and* $\phi\in R$ |
| **Natural Conjugate Prior (NC)** | $p_{NC}(\phi,\sigma)\propto\sigma^{-(r+2)}\exp\left\{\frac{-1}{2\sigma^2}\left[r\lambda^2+(\phi-\mu)^2h^2\right]\right\}$, $\sigma,h>0$ *and* $\phi,\mu\in R$ |
| **g-Prior** | $p_g(\phi,\sigma)\propto\sigma^{-2}\exp\left\{-\frac{g}{2\sigma^2}(\phi-\mu)^2\Sigma y_{t-1}^2\right\}$, $\sigma,g>0$ *and* $\phi,\mu\in R$ |

## 5.4. Comparative Study

This section is devoted to investigate and compare the performance of the prior distributions introduced in table 5.1 that have been selected to implement the posterior analysis for the AR(1) process. Furthermore, the sensitivity of the posterior distribution to the change in the prior used is studied. The comparative study is implemented via some selected criteria.

A Computer program, using Matlab (version 7.1) software, is designed to figure out these results. A script that does such a task is presented in Appendix-II. The graphical and table presentation are done using Excel program.

### 5.4.1. Simulation Algorithm

The current study follows the simulation techniques used by Ismail (1994) and Soliman (1999). The current simulation study deals with data generated from the model AR(1) represented by (5.2.1). Ten cases of AR(1) model are considered, for which, the values of the parameter $\phi$ were $\pm 0.2$, $\pm 0.5$, $\pm 0.8$, $\pm 1$ and $\pm 1.5$ respectively. The current work aims to assign different values of the autoregressive parameter on a wide range within and outside the stationarity domain of the AR(1) model. For each model 500 samples were generated each of length 700. For each sample, the first 200 observations were dropped to eliminate the effect of the initial values. Five different time series lengths have been chosen to study the influence of the series length on the performance of different prior distributions. These lengths are 30, 50,100, 200 and 500. The comparative study depends on some criteria as will be shown in the following section:

### 5.4.2. Tools of Comparison

Various frequentist criteria are helpful to compare among prior distributions. The basic idea is to use the prior distribution to generate a posterior distribution, and investigate the frequentist properties of such resulted distribution. If the posterior outcomes resulted from one prior has substantially better properties than that resulting from another prior, then the latter prior is suspected (Yang, 1994).

An interesting tool was used to determine the reasonable prior distribution. It is just a percentage measure for the number of samples that satisfy some condition. The current study considers the following criterion:

***95% Highest Posterior Density Region (HPDR)*** that is defined as the region under the posterior density over the interval centered at the posterior mean with probability 95%. For each simulation, $n^*$ is defined to be the number of samples where the 95%HPDR contains the true value of the parameter. Then, the percentage $P^*$ is evaluated such that:

$$P^* = \frac{n^*}{500} \times 100 \qquad (5.4.1)$$

The performance of a prior is evaluated according to the value of $P^*$. That is, for a given prior, the greater percentage indicates a higher performance of the prior to guide to a posterior that presents powerfully the parameter.

### 5.4.3. Results and Discussion

Regarding the general case of AR(1) models, there is no restriction on the values of the autoregressive coefficient $\phi$. Thus, the posterior outputs of all of the proposed ten AR(1) models will be studied using the three prior distributions given in the first row of table 5.1 which are Jeffreys' prior, g-prior and the Natural Conjugate (NC) prior since $\phi$ may take any value over the real line. The algorithm of the comparative analysis was implemented according to the following outlines. For each of the 500 samples; the first 30 observations used to evaluate the posterior distribution of the parameter $\phi$ via the three candidate priors. The posterior mean and the posterior variance of $\phi$ were computed given each prior. Tracing the criterion mentioned above, an interval centered at the posterior mean with probability 0.95 was evaluated (this is simply the 95% HPDRs of $\phi$). For each model, the percentage of samples for which the actual parameter exists within the indicated interval was computed (as shown by (5.4.1)). This process is repeated for the first 50 observations (including the first 30). Similarly, the process is repeated for the first 100, 200 and, finally, for the 500 observations. A script written by Matlab program was designed to accomplish the task shown above. Such script is attached in Appendix-II.

The results for each of the ten models are summarized throughout ten figures; each figure consists of a table and a bar graph. These tables and graphs represent the percentage $P^*$ defined by (5.4.1) for each $n^*$. Each table consists of five rows and four columns. The first column represents the time series length, while each other column matches the used prior distribution. The values in the cells of the table denote the percentages $P^*$. The graph attached to each table describes a bar graph summary to the content of the table.

**Figure 5.1**

$\phi$=0.2

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 94.4 | 95.4 | 97.2 |
| 50 | 94.4 | 94.4 | 95.6 |
| 100 | 93.8 | 94.0 | 94.4 |
| 200 | 96.0 | 95.8 | 95.8 |
| 500 | 95.6 | 94.4 | 94.2 |

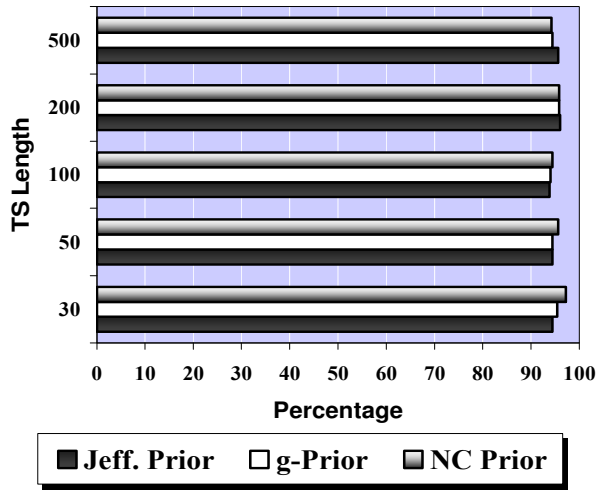



**Figure 5.2**

$\phi$=-0.2

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 94.8 | 94.4 | 97.6 |
| 50 | 95.2 | 94.8 | 97.2 |
| 100 | 94.0 | 94.8 | 95.2 |
| 200 | 97.0 | 96.8 | 96.8 |
| 500 | 95.4 | 95.4 | 95.2 |

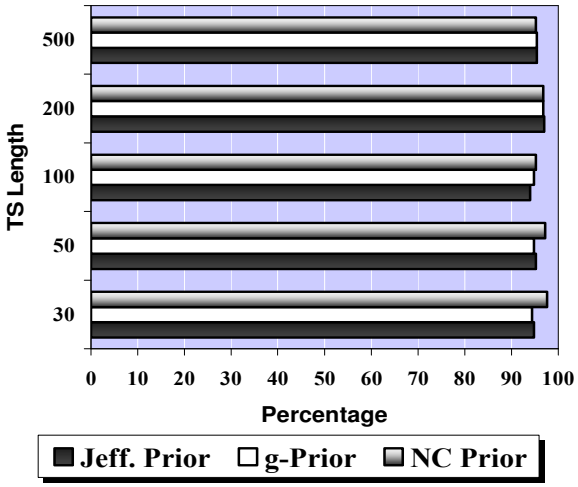



**Figure 5.3**

$\phi$=0.5

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 94.2 | 94.6 | 97.6 |
| 50 | 94.8 | 94.6 | 95.2 |
| 100 | 94.6 | 95.0 | 95.0 |
| 200 | 95.2 | 96.0 | 95.8 |
| 500 | 93.6 | 94.4 | 93.8 |

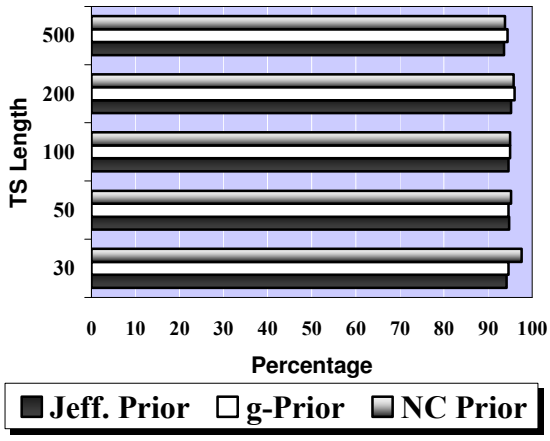



**Figure 5.4**

$\phi$=-0.5

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 95.2 | 94.6 | 97.4 |
| 50 | 96.2 | 96.0 | 96.8 |
| 100 | 95.2 | 94.0 | 94.2 |
| 200 | 95.8 | 96.8 | 96.4 |
| 500 | 96.0 | 95.4 | 95.2 |

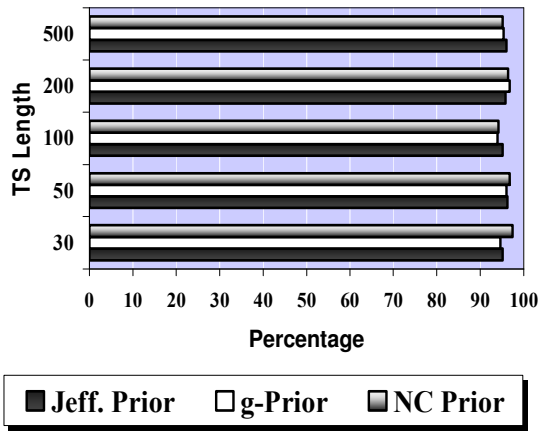


**Figure 5.5**

$\phi$=0.8

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 95.6 | 94.8 | 98.4 |
| 50 | 94.2 | 94.0 | 95.8 |
| 100 | 94.2 | 95.6 | 95.0 |
| 200 | 92.8 | 94.2 | 93.6 |
| 500 | 93.8 | 97.2 | 95.4 |

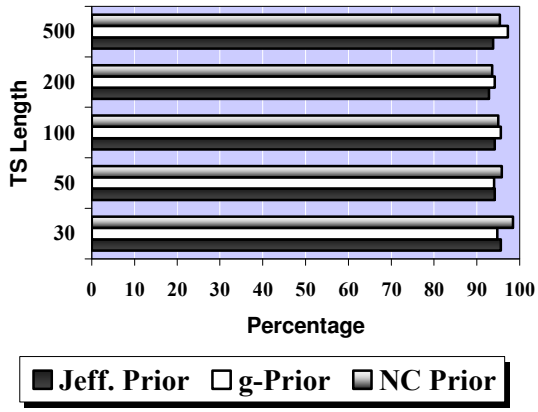



**Figure 5.6**

$\phi$=-0.8

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 95.6 | 95.8 | 97.8 |
| 50 | 94.4 | 95.6 | 96.8 |
| 100 | 94.6 | 94.8 | 94.6 |
| 200 | 95.4 | 96.2 | 95.8 |
| 500 | 96.8 | 98.2 | 97.0 |

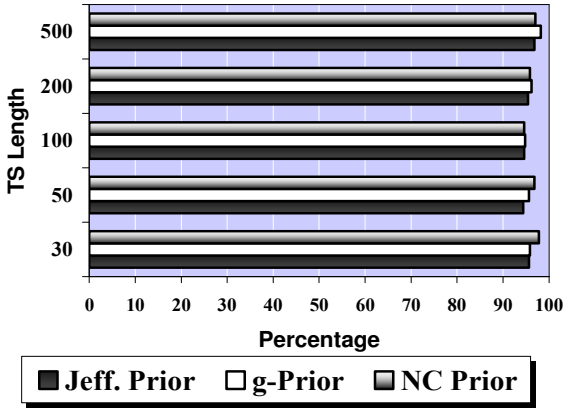



**Figure 5.7**

$\phi$=1

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 90.4 | 91.0 | 98.0 |
| 50 | 92.2 | 98.8 | 94.6 |
| 100 | 92.4 | 99.6 | 93.8 |
| 200 | 92.8 | 99.4 | 94.6 |
| 500 | 94.4 | 100.0 | 93.8 |

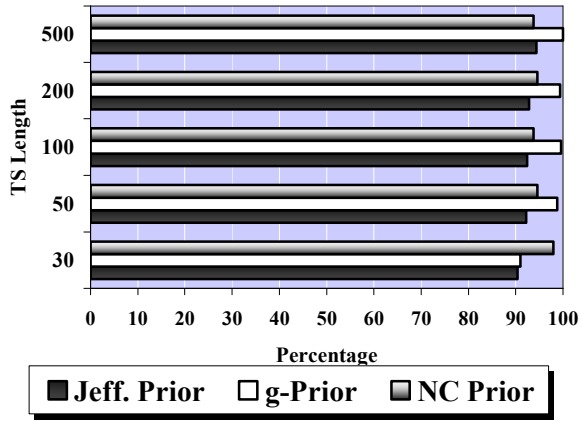



**Figure 5.8**

$\phi$=-1

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 95.6 | 96.0 | 99.4 |
| 50 | 94.4 | 99.6 | 96.8 |
| 100 | 95.4 | 99.6 | 96.4 |
| 200 | 97.2 | 99.6 | 96.2 |
| 500 | 95.6 | 100.0 | 95.6 |

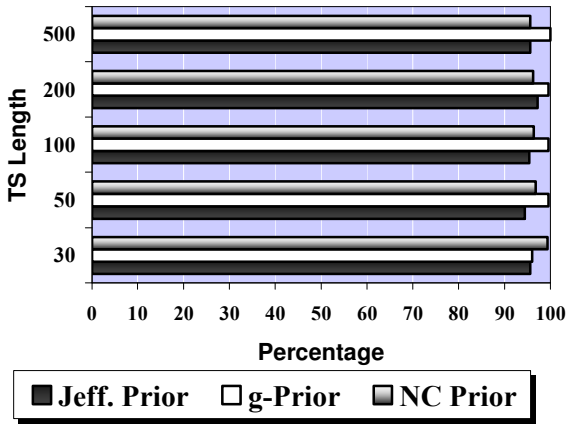


**Figure 5.9**

$\phi$=1.5

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 51.8 | 90.6 | 95.6 |
| 50 | 46.6 | 98.8 | 96.2 |
| 100 | 47.8 | 94.2 | 94.2 |
| 200 | 50.6 | 100.0 | 95.0 |
| 500 | 49.2 | 99.0 | 93.6 |

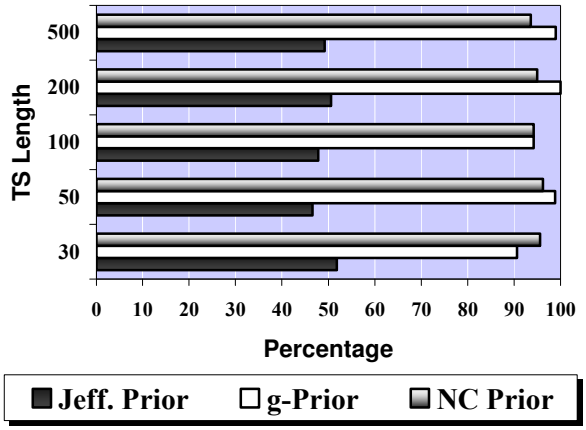
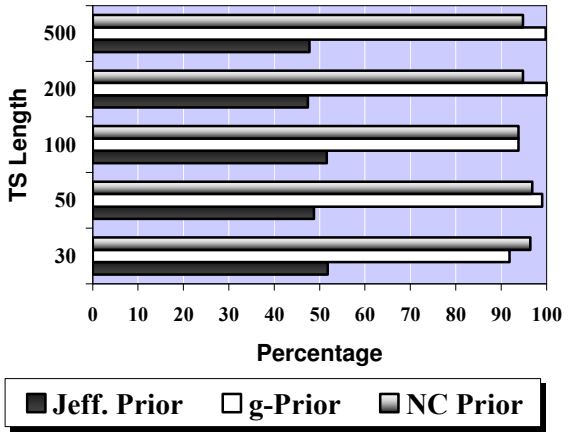



**Figure 5.10**

$\phi$=-1.5

| n | Jeff. Prior | g-Prior | NC Prior |
|---|---|---|---|
| 30 | 51.8 | 91.8 | 96.4 |
| 50 | 48.8 | 99.0 | 96.8 |
| 100 | 51.6 | 93.8 | 93.8 |
| 200 | 47.4 | 100.0 | 94.8 |
| 500 | 47.8 | 99.8 | 94.8 |




Regarding the above tables and graphs, we achieve the following conclusion:

1. Apart from the case $\phi = \pm 1.5$, all priors lead to consistent posterior, in the sense that the HPDR includes the parameter value in more than 90% of the cases at all time series lengths. There is no observable difference between the priors at each time series length.

2. For case $\phi = \pm 1.5$, the informative priors are highly better than the Jeffreys' prior which appears to be less consistent at all time series lengths.

3. The goodness of each prior is not sensitive to the increase of the time series length.

The above results support the use of Jeffreys' prior if there is an evident that $\phi \leq 1$, since it has approximately the same efficiency as informative priors and it avoids the problem of estimating the hyperparameters as well.

Nevertheless, if there is an evident that $\phi > 1$, it would be appropriate to select an informative prior because the lack of efficiency of the Jeffreys' prior. The NC appears to be a good choice for time series length below 50. However, the g-prior is better for longer time series.

## 5.5. Case Study

To illustrate the achieved results of the simulation study in section 5.4, three real life time series examples are considered. The data sets are the stock prices for some different firms. A graphical representation using Minitab package is enclosed to describe these data through a descriptive summary and time plot for each example. Moreover, the ACF and the PACF plot are displayed to check the possibility of modeling these data sets by AR(1) processes.

**Csae Study-I:** Weekly Average Closing Prices of Al-Watany Bank of Egypt from 1/1/1995 to 25/5/1997

**Csae Study-II:** Weekly Average Closing Prices of CIB from 1/1/1995 to 25/5/1997

**Csae Study-III:** Weekly Average Closing Prices of Kabo Company for Clothes from 1/2/1995 to 21/5/1997

**Figure 5.11**

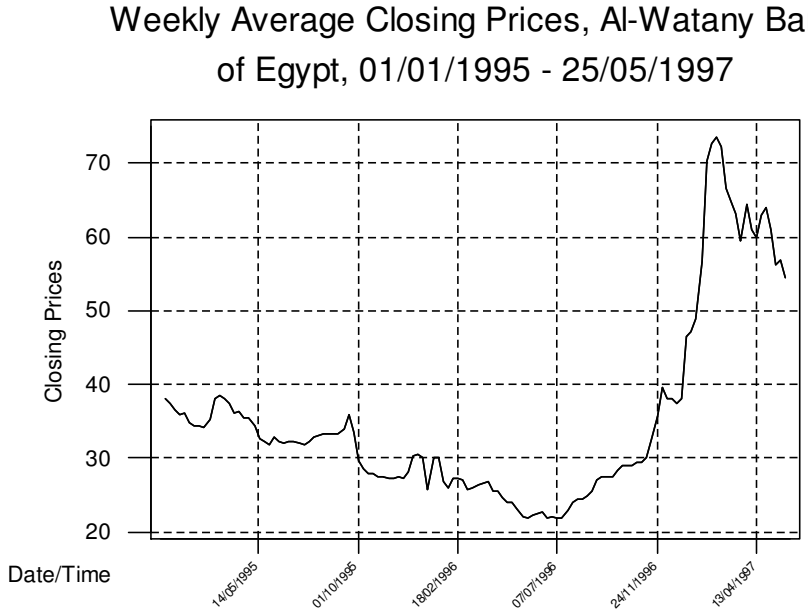
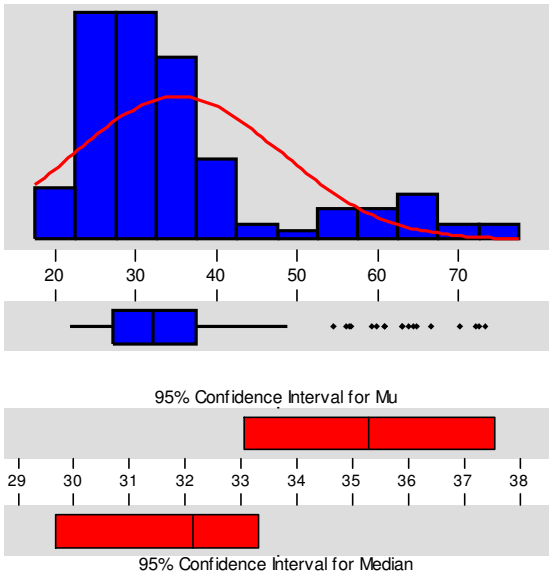


**Figure 5.12**

Descriptive Statistics: Weekly Average Closing Prices, Al-Watany Bank of Egypt, 01/01/1995 - 25/05/1997




Variable: AVERAGE

Anderson-Darling Normality Test

| | |
|---|---|
| A-Squared: | 9.576 |
| P-Value: | 0.000 |
| Mean | 35.2951 |
| StDev | 12.7873 |
| Variance | 163.516 |
| Skew ness | 1.55643 |
| Kurtosis | 1.53086 |
| N | 126 |
| Minimum | 21.8428 |
| 1st Quartile | 27.2222 |
| Median | 32.1315 |
| 3rd Quartile | 37.5089 |
| Maximum | 73.4243 |

95% Confidence Interval for Mu

| | |
|---|---|
| 33.0405 | 37.5497 |

95% Confidence Interval for Sigma

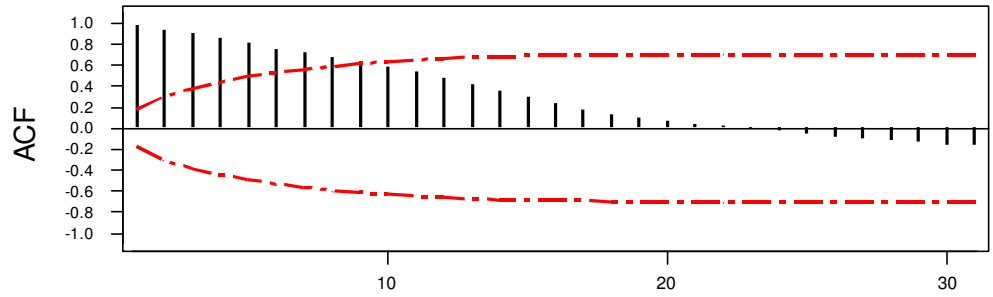
| | |
|---|---|
| 11.3797 | 14.5956 |

95% Confidence Interval for Median

| | |
|---|---|
| 29.6530 | 33.3105 |

**Figure 5.13**

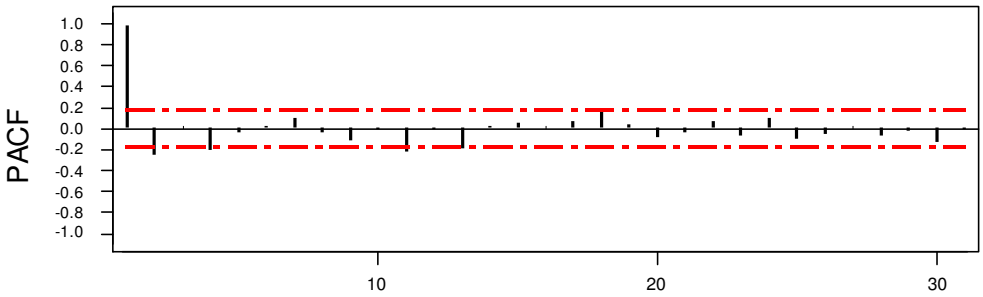
Weekly Average Closing Prices, Al-Watany Bank of Egypt




| Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.98 | 10.95 | 122.81 | 10 | 0.58 | 1.82 | 837.94 | 19 | 0.10 | 0.30 | 980.80 | 28 | -0.12 | -0.34 | 987.59 |
| 2 | 0.94 | 6.19 | 237.68 | 11 | 0.53 | 1.62 | 877.60 | 20 | 0.07 | 0.21 | 981.63 | 29 | -0.14 | -0.39 | 990.69 |
| 3 | 0.90 | 4.69 | 344.45 | 12 | 0.48 | 1.42 | 909.74 | 21 | 0.04 | 0.12 | 981.92 | 30 | -0.15 | -0.43 | 994.56 |
| 4 | 0.86 | 3.83 | 441.46 | 13 | 0.41 | 1.22 | 934.27 | 22 | 0.02 | 0.06 | 981.99 | 31 | -0.17 | -0.47 | 999.28 |
| 5 | 0.81 | 3.25 | 528.20 | 14 | 0.35 | 1.01 | 951.90 | 23 | -0.00 | -0.01 | 981.99 | | | | |
| 6 | 0.76 | 2.83 | 605.46 | 15 | 0.29 | 0.83 | 964.02 | 24 | -0.03 | -0.08 | 982.12 | | | | |
| 7 | 0.71 | 2.51 | 674.69 | 16 | 0.23 | 0.66 | 971.78 | 25 | -0.05 | -0.15 | 982.57 | | | | |
| 8 | 0.67 | 2.25 | 736.50 | 17 | 0.18 | 0.50 | 976.40 | 26 | -0.08 | -0.22 | 983.58 | | | | |
| 9 | 0.63 | 2.02 | 790.70 | 18 | 0.14 | 0.39 | 979.17 | 27 | -0.10 | -0.29 | 985.22 | | | | |

**Figure 5.14**

Weekly Average Closing Prices, Al-Watany Bank of Egypt




| Lag | PAC | T | Lag | PAC | T | Lag | PAC | T | Lag | PAC | T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.98 | 10.95 | 10 | -0.01 | -0.08 | 19 | 0.03 | 0.34 | 28 | -0.07 | -0.76 |
| 2 | -0.25 | -2.80 | 11 | -0.22 | -2.52 | 20 | -0.09 | -1.00 | 29 | -0.02 | -0.21 |
| 3 | 0.00 | 0.04 | 12 | -0.01 | -0.13 | 21 | -0.04 | -0.40 | 30 | -0.13 | -1.42 |
| 4 | -0.20 | -2.27 | 13 | -0.19 | -2.12 | 22 | 0.07 | 0.77 | 31 | -0.01 | -0.08 |
| 5 | -0.03 | -0.38 | 14 | 0.02 | 0.17 | 23 | -0.07 | -0.77 | | | |
| 6 | 0.02 | 0.23 | 15 | 0.05 | 0.54 | 24 | 0.09 | 1.03 | | | |
| 7 | 0.10 | 1.09 | 16 | 0.01 | 0.10 | 25 | -0.10 | -1.10 | | | |
| 8 | -0.03 | -0.37 | 17 | 0.07 | 0.75 | 26 | -0.05 | -0.57 | | | |
| 9 | -0.11 | -1.24 | 18 | 0.15 | 1.72 | 27 | 0.01 | 0.09 | | | |

**Figure 5.15**

Weekly Average Closing Prices, CIB,
Egypt, 01/01/1995 - 25/05/1997

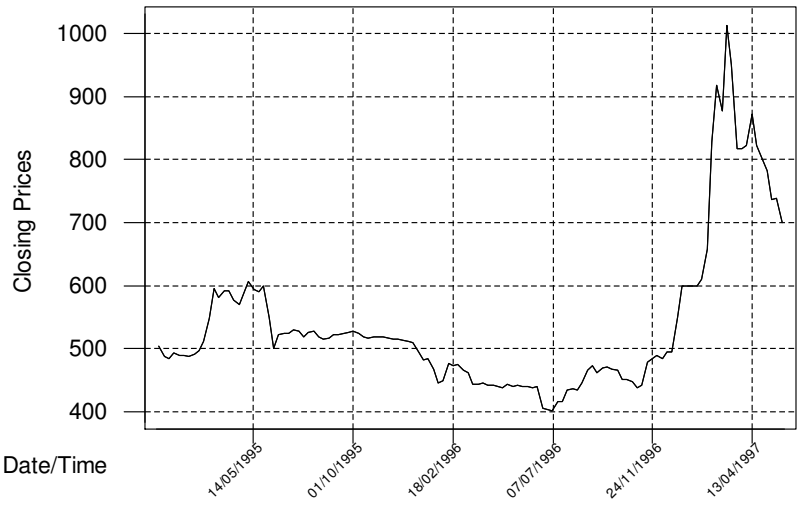
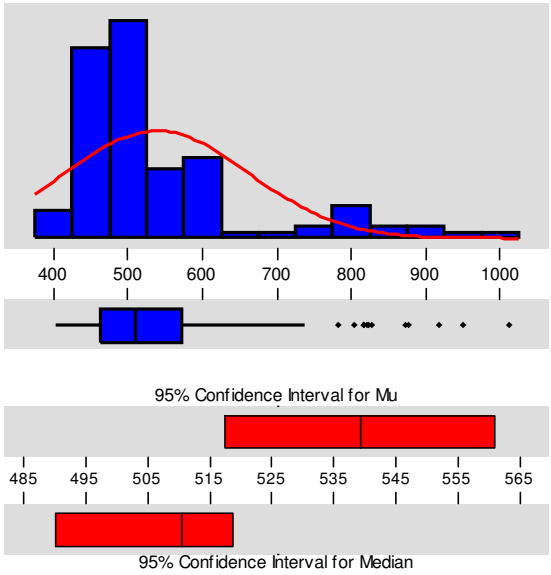


**Figure 5.16**

Descriptive Statistics: Weekly Average Closing Prices, CIB, Egypt, 01/01/1995 - 25/05/1997

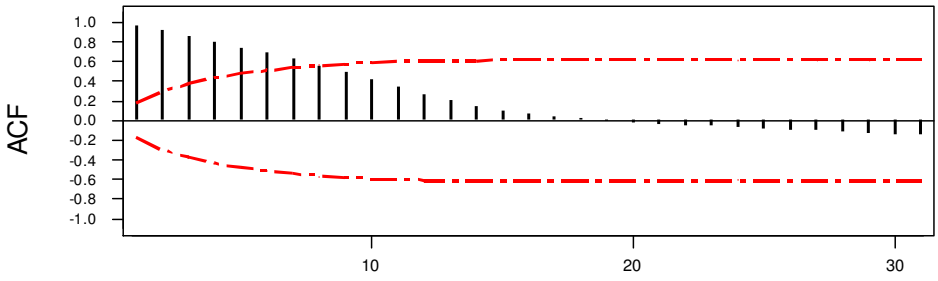



Variable: AVERAGE

| Anderson-Darling Normality Test | |
|---|---|
| A-Squared: | 9.890 |
| P-Value: | 0.000 |
| Mean | 539.190 |
| StDev | 123.208 |
| Variance | 15180.1 |
| Skew ness | 1.89066 |
| Kurtosis | 3.24422 |
| N | 126 |
| Minimum | 402.29 |
| 1st Quartile | 463.05 |
| Median | 510.51 |
| 3rd Quartile | 572.98 |
| Maximum | 1012.63 |
| 95% Confidence Interval for Mu | |
| 517.47 | 560.91 |
| 95% Confidence Interval for Sigma | |
| 109.64 | 140.63 |
| 95% Confidence Interval for Median | |
| 489.95 | 518.76 |

**Figure 5.17**

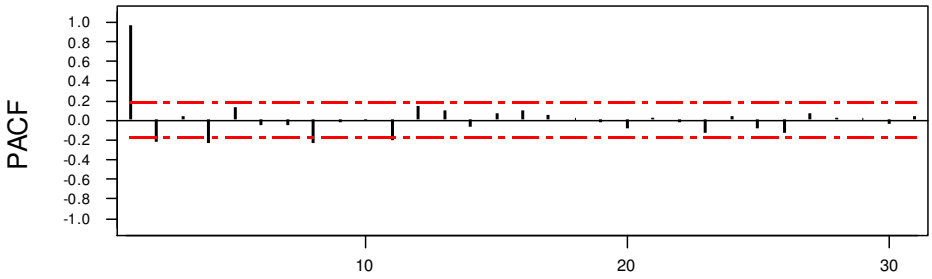
Weekly Average Closing Prices, CIB




| Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.96 | 10.81 | 119.63 | 10 | 0.42 | 1.40 | 698.51 | 19 | -0.00 | -0.01 | 735.84 | 28 | -0.12 | -0.37 | 744.07 |
| 2 | 0.91 | 6.05 | 227.60 | 11 | 0.33 | 1.11 | 714.22 | 20 | -0.02 | -0.07 | 735.90 | 29 | -0.13 | -0.40 | 746.67 |
| 3 | 0.86 | 4.55 | 324.86 | 12 | 0.26 | 0.86 | 723.89 | 21 | -0.04 | -0.12 | 736.12 | 30 | -0.14 | -0.44 | 749.83 |
| 4 | 0.80 | 3.66 | 409.10 | 13 | 0.21 | 0.67 | 730.04 | 22 | -0.05 | -0.16 | 736.51 | 31 | -0.15 | -0.48 | 753.63 |
| 5 | 0.74 | 3.08 | 482.10 | 14 | 0.15 | 0.49 | 733.32 | 23 | -0.06 | -0.19 | 737.05 | | | | |
| 6 | 0.69 | 2.66 | 545.42 | 15 | 0.10 | 0.33 | 734.84 | 24 | -0.07 | -0.22 | 737.80 | | | | |
| 7 | 0.63 | 2.31 | 598.95 | 16 | 0.07 | 0.22 | 735.55 | 25 | -0.08 | -0.25 | 738.79 | | | | |
| 8 | 0.56 | 1.98 | 641.74 | 17 | 0.04 | 0.13 | 735.80 | 26 | -0.09 | -0.29 | 740.13 | | | | |
| 9 | 0.49 | 1.67 | 674.39 | 18 | 0.02 | 0.05 | 735.84 | 27 | -0.10 | -0.33 | 741.89 | | | | |

**Figure 5.18**

Weekly Average Closing Prices, CIB




| Lag | PAC | T | Lag | PAC | T | Lag | PAC | T | Lag | PAC | T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.96 | 10.81 | 10 | -0.01 | -0.10 | 19 | -0.03 | -0.33 | 28 | 0.02 | 0.20 |
| 2 | -0.22 | -2.48 | 11 | -0.20 | -2.24 | 20 | -0.08 | -0.94 | 29 | 0.00 | 0.05 |
| 3 | 0.04 | 0.46 | 12 | 0.14 | 1.58 | 21 | 0.02 | 0.20 | 30 | -0.04 | -0.44 |
| 4 | -0.24 | -2.66 | 13 | 0.10 | 1.08 | 22 | -0.02 | -0.17 | 31 | 0.03 | 0.37 |
| 5 | 0.13 | 1.44 | 14 | -0.07 | -0.84 | 23 | -0.13 | -1.48 | | | |
| 6 | -0.05 | -0.58 | 15 | 0.07 | 0.75 | 24 | 0.03 | 0.35 | | | |
| 7 | -0.05 | -0.52 | 16 | 0.10 | 1.10 | 25 | -0.08 | -0.90 | | | |
| 8 | -0.23 | -2.61 | 17 | 0.05 | 0.52 | 26 | -0.13 | -1.42 | | | |
| 9 | -0.03 | -0.29 | 18 | 0.01 | 0.15 | 27 | 0.08 | 0.85 | | | |

**Figure 5.19**

Weekly Average Closing Prices, Kabo Company for Clothes, Egypt, 01/02/1995 - 21/05/1997

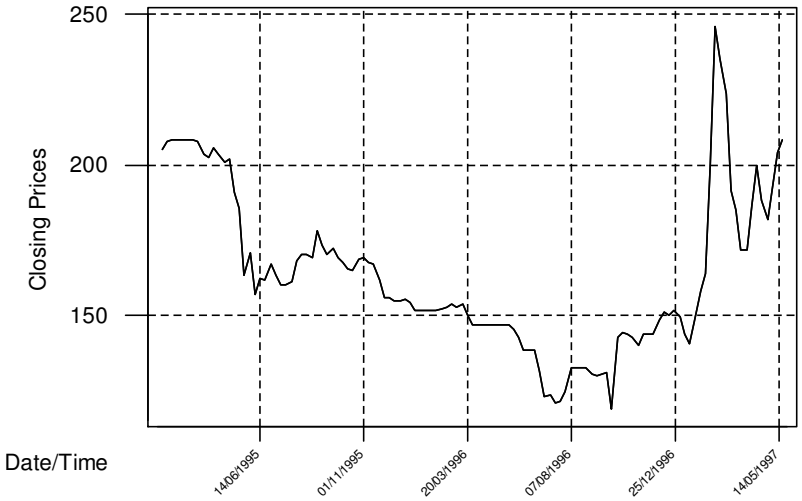
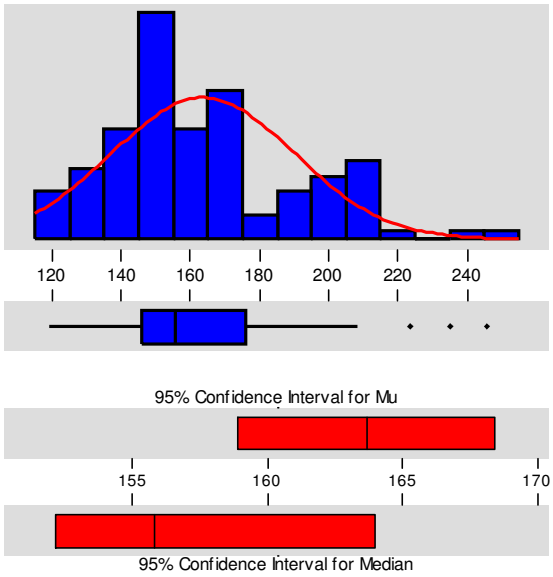


**Figure 5.20**

Descriptive Statistics: Daily Closing Prices, Kabo Company for Clothes, Egypt, 01/02/1995 - 21/05/1997




Variable: AVERAGE

Anderson-Darling Normality Test

| | |
|---|---|
| A-Squared: | 2.657 |
| P-Value: | 0.000 |
| Mean | 163.656 |
| StDev | 26.441 |
| Variance | 699.105 |
| Skew ness | 0.743041 |
| Kurtosis | 4.70E-02 |
| N | 121 |
| Minimum | 119.143 |
| 1st Quartile | 146.143 |
| Median | 155.800 |
| 3rd Quartile | 175.936 |
| Maximum | 245.940 |

95% Confidence Interval for Mu

| | |
|---|---|
| 158.897 | 168.415 |

95% Confidence Interval for Sigma

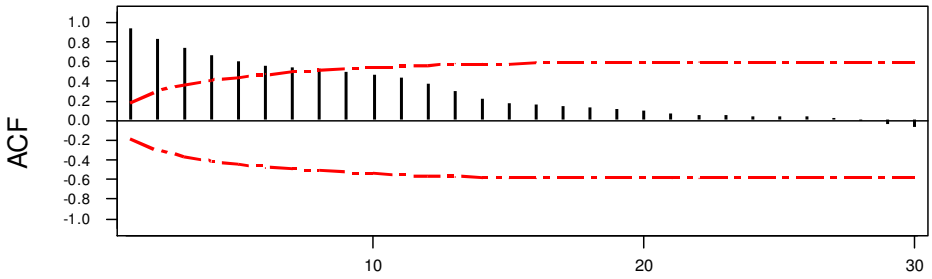
| | |
|---|---|
| 23.477 | 30.268 |

95% Confidence Interval for Median

| | |
|---|---|
| 152.090 | 163.969 |

**Figure 5.21**

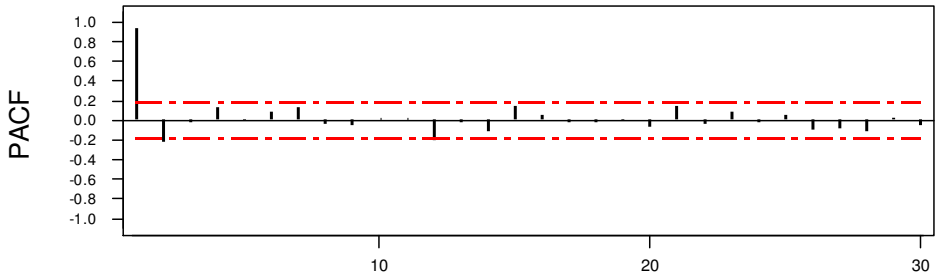
Weekly Average Daily Closing Prices, Kabo Company for Clothes




| Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.93 | 10.21 | 106.93 | 10 | 0.46 | 1.68 | 540.04 | 19 | 0.12 | 0.39 | 617.36 | 28 | -0.01 | -0.04 | 621.05 |
| 2 | 0.83 | 5.54 | 193.46 | 11 | 0.43 | 1.52 | 564.51 | 20 | 0.09 | 0.31 | 618.63 | 29 | -0.04 | -0.13 | 621.30 |
| 3 | 0.74 | 3.99 | 261.61 | 12 | 0.37 | 1.29 | 582.93 | 21 | 0.08 | 0.26 | 619.48 | 30 | -0.07 | -0.23 | 622.05 |
| 4 | 0.66 | 3.20 | 317.54 | 13 | 0.30 | 1.04 | 595.39 | 22 | 0.06 | 0.19 | 619.95 | | | | |
| 5 | 0.60 | 2.69 | 364.23 | 14 | 0.22 | 0.77 | 602.36 | 23 | 0.05 | 0.16 | 620.30 | | | | |
| 6 | 0.56 | 2.37 | 404.98 | 15 | 0.18 | 0.61 | 606.80 | 24 | 0.04 | 0.14 | 620.57 | | | | |
| 7 | 0.54 | 2.18 | 443.20 | 16 | 0.15 | 0.52 | 610.12 | 25 | 0.04 | 0.13 | 620.81 | | | | |
| 8 | 0.52 | 2.03 | 479.20 | 17 | 0.14 | 0.48 | 612.96 | 26 | 0.03 | 0.11 | 620.99 | | | | |
| 9 | 0.49 | 1.86 | 511.72 | 18 | 0.13 | 0.44 | 615.38 | 27 | 0.02 | 0.06 | 621.04 | | | | |

**Figure 5.22**

Weekly Average Daily Closing Prices, Kabo Company for Clothes




| Lag | PAC | T | Lag | PAC | T | Lag | PAC | T | Lag | PAC | T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.93 | 10.21 | 10 | 0.01 | 0.09 | 19 | -0.01 | -0.07 | 28 | -0.12 | -1.30 |
| 2 | -0.22 | -2.42 | 11 | 0.01 | 0.09 | 20 | -0.07 | -0.72 | 29 | 0.02 | 0.20 |
| 3 | -0.02 | -0.25 | 12 | -0.21 | -2.32 | 21 | 0.14 | 1.56 | 30 | -0.05 | -0.60 |
| 4 | 0.13 | 1.41 | 13 | -0.03 | -0.31 | 22 | -0.04 | -0.49 | | | |
| 5 | -0.01 | -0.16 | 14 | -0.11 | -1.26 | 23 | 0.08 | 0.86 | | | |
| 6 | 0.08 | 0.84 | 15 | 0.14 | 1.53 | 24 | -0.02 | -0.18 | | | |
| 7 | 0.13 | 1.40 | 16 | 0.05 | 0.52 | 25 | 0.06 | 0.63 | | | |
| 8 | -0.04 | -0.48 | 17 | -0.02 | -0.22 | 26 | -0.10 | -1.06 | | | |
| 9 | -0.06 | -0.63 | 18 | -0.02 | -0.27 | 27 | -0.08 | -0.91 | | | |

Regarding above figures, the time series plot for all the data sets shows that none of them is stationary. Such result is confirmed, as well, by the ACF charts, since the ACF for all data sets is decaying slowly. Moreover, it is entirely evident as shown by figures 5.14, 5.18 and 5.22 that the PACF's of all data sets are cutting off after the first lag. This result emphasizes that all data can be modeled by AR(1) processes, according to Box-Jenkins criteria.

The concern now is to demonstrate the results of the previous section, §5.4, via these three examples. Thus, following the procedure of section 5.4, the posterior analysis was accomplished and compared over the three candidate priors; Jeffreys' prior, g-prior and NC prior. For each data set, the posterior mean, the posterior variance and the 95% highest posterior density region (HPDR) centered at the posterior mean are evaluated with respect to the three proposed prior distributions. The results of such posterior analysis are summarized through the following table (table 5.2). A matlab script is designed to employ such calculations. It is attached in Appendix-III.

**Table 5.2**

**Posterior Mean, Posterior Variance of ϕ and the 95% HPDRs centered at the posterior mean by Prior Distribution for Different Data Sets**

**[a] Case Study-I: Weekly Average Closing Prices of Al-Watany Bank of Egypt (n=126)**

| Prior | Postrior Mean | Posterior Variance | 95% HPDRs |
|---|---|---|---|
| Jeffreys' Prior | 1.0026 | 0.0000 | [ 0.9920 , 1.0131] |
| g-Prior | 1.0029 | 0.0000 | [ 0.9814 , 1.0143] |
| NC Prior | 1.0029 | 0.0000 | [ 0.9914 , 1.0144] |

**[b] Case Study-II: Weekly Average Closing Prices of CIB (n=126)**

| Prior | Postrior Mean | Posterior Variance | 95% HPDRs |
|---|---|---|---|
| Jeffreys' Prior | 1.0016 | 0.0000 | [ 0.9919 , 1.0113] |
| g-Prior | 1.0007 | 0.0000 | [ 0.9904 , 1.0111] |
| NC Prior | 1.0111 | 0.0001 | [ 0.9830 , 1.0182] |

**[c] Case Study-III: Weekly Average Closing Prices of Kabo Company for Clothes (n=121)**

| Prior | Postrior Mean | Posterior Variance | 95% HPDRs |
|---|---|---|---|
| Jeffreys' Prior | 0.9989 | 0.0000 | [ 0.9898 , 1.0080] |
| g-Prior | 0.9990 | 0.0000 | [ 0.9887 , 1.0094] |
| NC Prior | 0.9991 | 0.0000 | [ 0.9887 , 1.0094] |

Examining the above results shows similar conclusions for the posterior analysis. The performance of the three priors is almost the same since they all lead to the same posterior mean values with very small posterior variance. The unique difference is shown through the 95% HPDRs that supposed to give a probability 0.95 with shortest interval. Therefore, the length of the computed interval is taken as a powerful tool to compare the performance of the priors.

Regarding case study-I, table 5.2a shows that Jeffreys' prior and the NC prior gave the shortest interval with length 0.02. In addition, g-prior leads to a bit similar value with length 0.03. Concerning the posterior analysis of case-II, table 5.2b shows similar outcomes, since both g-prior and Jeffreys' prior guide to 95% HPDRs with shortest length, 0.02. However, NC prior leads to posterior interval with length 0.04. On the other hand, case-III gives entirely similar results, since all priors guide to interval with length 0.02 (see table 5.2c).

# Chapter 6

# Conclusion and Future Work

This study is interested in the problem of prior selection in Bayesian analysis. To achieve the goals of the study, several well known priors in the literature were discussed and explained. The priors were divided according to their nature into informative priors and noninformative priors.

Among noninformative priors the study considered, the Jeffreys' prior, the locally uniform prior and the maximal data information prior. Whereas, among informative priors, the study was interested in the natural conjugate prior and the g-prior.

For each prior, the basic idea was explained, the derivation was given, the main properties were discussed and some theoretical examples were shown.

Some applications of the problem of prior selection were given. The posterior analysis of the general linear model was employed using informative priors.

A comprehensive application to, the well known time series model, AR(1) was done. The posterior analysis of AR(1) was employed. The three noninformative techniques implied the same form except for Jeffreys' prior where it assumed the independence rule. While, posterior analysis of AR(1), using informative approaches, showed same results as for the GLM since AR(1) is often considered as a special case of the GLM.

Simulation was used to check the efficiency of the priors to achieve a consistent posterior distribution for the coefficient $\phi$ of the AR(1) models. Several simulation studies were employed assuming different values for $\phi$ and different time series lengths. Some criteria were used to indicate the goodness of the priors.

In the simulation study, for $\phi$ takes values within the stationary limits, all the priors lead to consistent posterior. Nevertheless, for $\phi$ takes values outside the stationary

limits, the informative priors only were efficient. Furthermore, a recommendation was given for the g-prior for long time series.

Finally, the study considered some real time series examples to illustrate the process of prior selection in the posterior analysis in real life. All the time series considered follow the AR(1) processes. Posterior analysis of the real data examples showed similar results for all priors.

**Future Work**

This study can be extended in different aspects to enclose further points of future research. In further details, the following points are some examples of future research:

1. The study can be extended to involve many other types of prior distributions that may be informative or noninformative priors.

2. Moreover, the application of the prior selection problem can be extended to further models such as; multivariate GLM, bivariate AR(1), bivariate autoregressive models, multivariate AR(1) and multivariate autoregressive models.

# *Bibliography*


1. Barnett, V. (1973). *Comparative Statistical Inference*. John Wiley & Sons, London.
2. Bayes, T.R. (1763). Essay Towards Solving a Problem in the Doctrine of Chances. Reprinted in *Biometrika*. **45** (1958), pp. 243-315.
3. Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.
4. Berger, J.O. and Berliner L. M. (1983). Robust Bayes and Empirical Bayes Analysis with ε-Contaminated Priors. *Technical Report # 83-35*, Dep. Of Statistics, Purdue University, West Lafayette.
5. Berger, J.O. and Bernardo J.M. (1989). Estimating a Product of Means: Bayesian Analysis with Reference Priors. *Journal of the American Statistical Association*. **84**, pp. 200-207.
6. Berger, J.O. and Bernardo J.M. (1992). Ordered Group Reference Priors with Application to the Multinomial Problem. *Biometrika*. **79**, pp. 25-37.
7. Berger, J.O. and Pericchi, L.R. (1996). The Intrinsic Bayes Factor for Model Selection. *J Amer. Statist. Assoc*. **91**, pp. 109-122.
8. Berger, J. O. and Pericchi, L.R. (2004). Training Samples in Objective Bayesian Model Selection. *Annals of Statistics*. **32**, No. 3, pp. 841-869.
9. Berger, J.O. and Strawderman, W.E. (1993). Choice of Hierarchical Priors: Admissibility in Estimation of Normal Means. *Tech. Report*, Purdue University, USA.
10. Berger, J.O. and Yang R. (1996). A Catalog of Noninformative Priors. *Tech. Report*, Purdue University, USA.
11. Berger, J.O., Bernardo, J.M. and Sun, D. (2007). The Formal Definition of Reference Priors. *Tech. Rep*. Universidad de Valencia, Spain.
12. Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*. John Wiley & Sons, Inc., New York, NY.

13. Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts.

14. Broemeling, L.D. (1985). *Bayesian Analysis of Linear Models*. Marcel Dekker Inc., New York.

15. Clyd, M. (2003). Gaussian Hyper-G Priors for Variable Selection. *Statistics Seminars*, Dep. Of Statistics. University of Pennsylvania.

16. Datta, G.S. and Ghosh, M. (1995). Some Remarks on Noninformative Priors. *Journal of the American Statistical Association*. **90**, pp. 1357-1363.

17. Datta, G.S. and Ghosh, M. (1996). On the Invariance of Noninformative Priors. *Ann. Statist*., **24**, pp. 141-159.

18. DeGroot, M.H. (1970). *Optimal Statistical Decision*. McGraw-Hill, New York.

19. Diaz, J. and Farah, J. L. (1981). Bayesian Identification of Autoregressive Processes. *22nd NBER-NSF Seminar on Bayesian Inference in Econometrics.*

20. Fernandez G., Eduardo Ley E. and Steel M. F. J. (1998). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*.

21. Foster, D.P. and George, E.I. (1994), The Risk Inflation Criterion for multiple regression. *The Annals of Statistics*, **22**, pp. 1947-1975.

22. Ghosh, M. and Heo, J. (2000). Default Bayesian Priors for Regression Models with First-Order Autoregressive Residuals. *Journal of Time Series Analysis*, **24**, (no. 3), pp. 269-282.

23. Geisser, S. (1980). A predictive Primer. In *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.), **ch24**, North-Holland, Amsterdam.

24. Geisser, S. (1990). Predictive approaches to discordance testing. In *Bayesian and Likelihood Methods in Statistics and Econometrics* (S. Geisser, J.S. Hodges, S.J. Press, A. Zellner, eds.), 321-335. North-Holland, Amsterdam.

25. Gelman, A. (2002). Prior Distributions. Items for *Encyclopedia of Environmetrics*. Columbia University, New York.

26. Golan, A. (2002). Information and entropy econometrics – Editor's view. *Journal of Econometrics*, **107**, pp.1-15.

27. Good, I.J. (1956). Some Terminology and Notation in Information Theory. *Proc. of the Institute of Electrical Engineers*, C, **103**, pp. 200-204.

28. Good, I.J. (1980). The Contribution of Jeffreys to Bayesian Statistics. In *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.), **ch4**, North-Holland, Amsterdam.

29. Good, I.J. (1983a). *Good Thinking: The Foundation of Probability and Its Application*. Minneapolis: University of Minnesota Press.

30. Good, I.J. (1983b). The Robustness of Hierarchical Model for Multinomials and Contingency Tables. In *Scientific Inference, Data Analysis, and Robustness.* G.EP. Box, T. Leonard and C.F. Wu (Eds.). Academic Press, New York.

31. Hahn, E.D. (2006). Re-examining informative prior elicitation through the lens of Markov chain Monte Carlo methods (MCMC). *Journal of the Royal Statistical Society*. Ser. A, **169**, part 1, pp. 37-48

32. Hartigan, J. (1964). Invariant Prior Distributions. *Annals of Mathematical Statistic*, **35**, pp. 836-845.

33. Huzurbazar, V.S. (1980). Bayesian Inference and Invariant Prior Probabilities. In *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.), **ch28**, North-Holland, Amsterdam.

34. Irony, T.Z. and Singpurwalla, N.D. (1996). Noninformative Priors Do Not Exist: A Discussion with Jos´e M. Bernardo. *J. Statist. Planning and Inference*.

35. Ismail, M.A. (1994). *Bayesian Forecasting for Nonlinear Series*. Unpublished Ph.D., University of Wales, UK.

36. Jaynes, E.T. (1957). Information Theory and Statistical Mechanics. I, II. *Physical Review,* **106**, pp. 620-630; **108**, pp. 171-190.

37. Jaynes, E.T. (1968). Prior Probabilities. I, II. *IEEE Transactions on System Science and Cybernetics*, SSC-4, 227-241.

38. Jaynes, E.T. (1980). Marginalization and Prior Probabilities. *Bayesian Analysis in Econometrics and Statistics*, (A. Zellner ed.), North Holland, Amsterdam.

39. Jaynes, E.T. (1982). On the Rationale of Maximum Entropy Methods. *Proc. of IEEE*, **70**, pp. 939-952.

40. Jaynes, E.T. (1983). *Papers on Probability, Statistics and Statistical Physics*. (R. Rosen Krantz ed.). Dordrecht: D. Reidel.

41. Jaynes, E.T. (1985). *Where Do We Go From Here*, in: Maximum Entropy and Bayesian Methods in Inverse Problems, ed. C Ray Smith, pp.21-58, D. Reidel,

42. Jeffreys, H. (1931). *Scientific Inference*. Cambridge: Cambridge University Press.

43. Jeffreys, H. (1939). *Theory of Probability* (1st Ed.). Oxford University Press, London.

44. Jeffreys, H. (1948). *Theory of Probability* (2nd Ed.). Oxford University Press, London.

45. Jeffreys, H. (1961). *Theory of Probability* (3rd Ed.). Oxford University Press, London.

46. Jörnsten, R. and Yu, B. (2002) Multiterminal Estimation: A New and Improved Upper Bound on Estimation Efficiency Using Compressed Information. *Proceedings of the ISIT*.

47. Kadane, J.B. (1980). Predictive and Structural Methods for Eliciting Prior Distributions. In *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.). North-Holland, Amsterdam.

48. Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S. and Peters, S.C. (1980). Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association*. **75**, pp. 845-854.

49. Karlsson, S. (2001). *Bayesian Methods in Econometrics: Linear Regression*. (A course to a graduate level class), Stockholm School of Economics, Swedish School of Economics and Business Administration (Helsinki).

50. Kass, R.E. (1982). A Comment on "Is Jeffreys a Nicessarist?" *The American Statistician*., **36**, No. 4, pp.390-391.

51. Kass, R.E. (1990). Data-Translated Likelihood and Jeffreys' Rule. *Biometrika*. **77**, pp. 107-114. *Journal of Economic Survey*, **8**, No. 1, pp. 1-34

52. Kass, R.E. and Wasserman, L. (1995), A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion, *Journal of the American Statistical Association*, **90**, 928-934.

53. Kass, R.E. and Wasserman, L. (1996). "The Selection of Prior Distributions by Formal Rules". *Journal of the American Statistical Association*. **91**, pp. 1343-1370.

54. Koop, G. (1994). Recent Progress in Applied Bayesian Econometrics. *Journal of Economic Survey*s, Vol. 8, No. 1, 1-34.

55. Koop, G., Osiewalski, J. and Steel, M.F.J. (1995). Bayesian Long-Run Prediction in Time Series Models. *Journal of Econometrics*, **69**, pp. 61-80.

56. Koop, G. and Potter, S. (2003). Forecasting Dynamic Factor Models Using Bayesian Model Averaging. *Econometrics Journal*, Vol. 7, pp. 550-565.

57. Lahiff, M. (1980). Time series forecasting with informative prior distribution. *Technical Report,* No. 111, Department of Statistics, University of Chicago.

58. Laplace, P.S. (1812). *Theorie Analytique des Probabilities*. Courcier, Paris.

59. Lee, P.M. (1989). *Bayesian Statistics: An Introduction*. Oxford University Press, New York.

60. Lempers, F.B. (1971). *Posterior Probabilities of alternative Linear Models*. Rotterdam University Press.

61. Lindley, D.V. (1956). On a Measure of the Information Provided by an Experiment. Part 1, Probability. *Annals of Mathematical Statistic*, **27**, pp. 986-1005.

62. Lindley, D.V. (1965a). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Part 1, Probability. Cambridge, University of Press, UK.

63. Lindley, D.V. (1965b). *Introduction to Probability and Statistics from a Bayesian* Viewpoint. Part 2, Inference. Cambridge University of Press.

64. Lindley, D.V. (1980). Jeffreys' Contribution to Modern Statistical Thought. In *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.), **ch4**, North-Holland, Amsterdam.

65. Lindley, D.V. and Smith A. F .M. (1972). Bayes Estimates of the Linear Model. *Journal of the Royal Statistical Association*, Ser B, **34**, pp. 1-18.

66. Litterman, R. (1980). *A Bayesian Procedure for Forecasting with Vector Autoregression*. Manuscript dated September 1980 and Ph.D. Thesis, University of Minnesota, 1979.

67. Maddala, G.S. (1988). *Introduction to Econometric*. New York: Macmillan.

68. Mills, T.C. (1990). *Time Series Techniques for Economists*. New York: Cambridge University Press.

69. Muth, J. (1961). Rational Expectations and the Theory of Price Movements. *Econometrica,* pp. 315-335.

70. O'Hagan, A. (1994). *Kandall's Advanced Theory of Statistics*. Vol 2B-*Bayesian Inference*, London: Arnold.

71. Packiorek, C.J. (2006). Misinformation in the Conjugate Prior for the Linear Model with Implication for Free-Knot Spline Modelling. *Bayesian Analysis*, **1**, No. 2, pp. 375-383.

72. Peers, H.W. (1965). On Confidence Points and Bayesian Probability Points in the Case of Several Parameters. Journal of the Royal Statistical *Society*. Ser. B, **27**, pp. 9-16.

73. Peers, H.W. (1968). Confidence Properties of Bayesian Interval Estimates. *Journal of the Royal Statistical Society*. Ser. B, **30**, pp. 535-544.

74. Pericchi, L.R. (1998). Sets of Priors Probabilities and Bayesian Robustness. A Contribution to the Documentation Section on the website of the *Society for Imprecise Probability Theory and Application* (SIPTA): http://sipta.org.

75. Phillips, P.C.B. (1991). To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends. *Journal of Applied Econometrics*. **6**, pp. 333-364.

76. Press, S.J. (1989). *Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, New York.

77. Pérez, J.M. and Berger, J.O. (2002). Expected Posterior Prior Distributions for Model Selection. *Biometrika*. **89**, pp. 491-512.

78. Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston.

79. Rao, C.R. (1987). Differential Metrics in Probability Spaces. in *Differential Geometry in Statistical Inference*, Vol. 10, Ch. 5, Institute of Mathematical Statistics, Hayward, pp. 217-240.

80. Robbins, H. (1956). An Empirical Bayes Approach to Statistics. *Proceeding of the Third Berkeley Symposium on Mathematical Statistics*, volume 1, pp. 157-163, University of California Press, Berkeley.

81. Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Tech. J.* **27**, pp. 379-423.

82. Sinha, S.K. and Zellner, A. (1990). A Note on the Prior Distributions of Weibull Parameters, *SCIMA*, **19**, pp. 5-13.

83. Soliman, E.E.A. (1999). *On Bayesian Time Series Analysis*. Unpublished Ph.D., Faculty of Economics and Political Sciences, Cairo University, Egypt.

84. Soofi, E.S. (1994). Capturing the Intangible Concept of Information. *Journal of the American Statistical Association*. **89**, no. 428, pp. 1243-1254.

85. Sun, D. and Ye, K. (1995). Reference Prior Bayesian Analysis for Normal Mean Product. *Journal of the American Statistical Association*. **90**, pp. 589-597.

86. Villegas, C. (1977). On the Representation of Ignorance. *Journal of the American Statistical Association*. **72**, pp. 651-654.

87. Villegas, C. (1981). Inner Statistical Inference, II. *Ann. Statist*. **9**, pp. 768-776.

88. Villegas, C. (1984). On a Group Structural Approach to Bayesian Inference. *Tech. Report* 84-11, Department of Mathematics, Simon Fraser University, Burnaby.

89. Welch, B.L. and Peers, H.W. (1963). On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods. *Journal of the Royal Statistical Society*. Ser. B, **25**, pp. 318-329.

90. West, M., MÄuller, P. and Escobar, M.D. (1994). *Hierarchical Priors and Mixture Models, with Applications in Regression and Density Estimation*. (In: Aspects of Uncertainty: A Tribute to D.V. Lindley (eds: P Freeman et al)), p363-386.

91. Winkler, R.L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, **67**, pp. 776-800.

92. Winkler, R.L. (1977). Prior Distribution and Model-Building in Regression Analysis. in A. Aykac and C. Brumat, eds., *New Developments in the Applications of Bayesian Methods*. ( North-Holland, Amsterdam.), pp. 233-242.

93. Winkler, R.L. (1980). Prior Information, Predictive Distributions, and Bayesian Model-Building. In *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.). North-Holland, Amsterdam.

94. Wrinch, D. and Jeffreys, H. (1919). On Some Aspects of the Theory of Probability. *Philosophical Magazine*, **38**, pp. 715-731.

95. Wrinch, D. and Jeffreys, H. (1921). On Certain Fundamental Principles of Scientific Inquiry (Two Papers), Philosophical Magazine, **42**, pp. 369-390.

96. Wrinch, D. and Jeffreys, H. (1923) On Certain Fundamental Principles of Scientific Inquiry. *Philosophical Magazine*, **45**, pp. 368-374.

97. Yang, R. (1994). *Development of Noninformative Priors for Bayesian Analysis*. Unpublished Ph.D. Purdue University, USA.

98. Ye, K. (1990). *Noninformative Priors in Bayesian Analysis*. Unpublished Ph.D., Purdue University, USA.

99. Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, New York.

100. Zellner, A. (1977). Maximal Data Information Prior Distributions. Basic Issues in Econometrics, pp. 201-215, University of Chicago Press, London.

101. Zellner, A. (1980). Bayesian Analysis in Economics and Statistics. Essays in honor of H. Jeffreys. (Vol. 1), North-Holland, Amsterdam. New York.

102. Zellner, A. (1982a). "Is Jeffreys a Nicessarist?" *The American Statistician*., 36, No. 4, pp. 28-30.

103. Zellner, A. (1982b). Reply to a Comment on "Is Jeffreys a Nicessarist?" *The American Statistician*., 36, No. 4, pp.392-393.

104. Zellner, A. (1983). Application of Bayesian Analysis and Econometrics. Statistician, 32, pp. 23-34.

105. Zellner, A. (1984). Basic issues in econometrics. University of Chicago Press, Chicago.

106. Zellner, A. (1985). On Assessing Informative Prior Distributions for Regression Coefficient. In Bernardo et al., eds., Bayesian Statistics, 2 (North-Holland Publishing Co., Amsterdam, pp. 571-581.

107. Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior distributions. Bayesian inference and Decision Techniques. Essays in Honor of Bruno de Finetti, P.d.f. Goel and A. Zellner, eds., North-Holland, pp. 233-243, New York.

108. Zellner, A. (1991). Bayesian Methods and Entropy in Economics and Econometrics. (W.T. Grandy and L.H. Schick eds.), Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht. pp. 17-31.

109. Zellner, A. (1995). Past and Recent Results on Maximal Data Information Priors. Manuscript, H.G.B. Alexander Research Foundation, University of Chicago, Chicago.

110. Zellner, A. (1996). Models, prior information, and Bayesian analysis. Journal of Econometrics, 75 (no.1), pp. 51-68.

111. Zellner, A. and Highfield, R.A. (1988). Calculation of Maximum Entropy Distributions and Approximation of Marginal Posterior Distributions. Journal of Econometrics, 37, pp. 195-210.

112. Zellner, A. and Min, C-K. (1993). Bayesian Analysis, Model Selection and Prediction. Invited paper presented in W.T. Grandy and P.W. Milonni (eds), Physics and Probability: Essays in Honor of Edwin T. Jaynes, Cambridge: Cambridge University Press, pp. 195-206.

113. Zellner, A. and Tiao, G.C. (1964). Bayesian Analysis of the Regression Model with Autocorrelated Errors. *Journal of the American Statistical Association*, **59**, pp. 763-778.

# *Appendix-I*

## Forms for some standard distribution used in the thesis

**I. Gamma-I:** (Raiffa and Schlaifer, 1961, Part III, CH. 7, pp. 225)

A continuous random variable $X$ is said to have a Gamma of type I distribution with parameters (r,λ) if the p.d.f. of $X$ is defined by:

$$f(x|r,\lambda)=\frac{(\lambda x)^{r-1}\exp\{-\lambda x\}}{\Gamma(r)}\lambda,\quad x\geq 0,\ r,\lambda>0$$

and $\mu_x=\frac{r}{\lambda}$, and $\sigma_x^2=\frac{r}{\lambda^2}$.

**II. Gamma-II:** (Raiffa and Schlaifer, 1961, Part III, CH. 7, pp. 226)

A continuous random variable $X$ is said to have a Gamma of type II distribution with parameters (r,λ) if the p.d.f. of $X$ is defined by:

$$f(x|r,\lambda)=\frac{1}{\Gamma(\frac{r}{2})}\left(\frac{r\lambda x}{2}\right)^{r/2-1}\left(\frac{r\lambda}{2}\right)\exp\left\{-\frac{r\lambda x}{2}\right\},\quad x\geq 0,\ r,\lambda>0$$

and $\mu_x=\frac{1}{\lambda}$, and $\sigma_x^2=\frac{2}{r\lambda^2}$.

**III. Inverted-Gamma-I:** (Raiffa and Schlaifer, 1961, Part III, CH. 7, pp. 227)

If a continuous random variable X has a Gamma-I distribution with parameters (r,λ), then the inverse transformation $\boldsymbol{Y}=\boldsymbol{X}^{-1}$ is said to have an Inverted-Gamma of type I distribution with p.d.f. is defined by:

$$f(y|r,\lambda)=\frac{(\lambda/y)^{r+1}\exp\{-\lambda/y\}}{\lambda\,\Gamma(r)},\quad y\geq 0,\ r,\lambda>0$$

and $\mu_y=\lambda(r-1)^{-1},\quad r>1,$ and $\sigma_y^2=\lambda^2(r-1)^{-2}(r-2),\quad r>2.$

**IV. Inverted-Gamma-II:** (Raiffa and Schlaifer, 1961, Part III, CH. 7, pp. 228)

A continuous random variable $X$ is said to have an Inverted Gamma of type II distribution with parameters (r,λ) if the p.d.f. of $X$ is defined by:

$$f(x|r,\lambda)=\frac{2\,(r\lambda^2/2)^{r/2}}{\Gamma(r/2)}x^{-(r+1)}\exp\{-r\lambda^2/2x^2\}$$

and $\mu_x = \lambda\sqrt{\frac{r}{2}}\frac{(\frac{r}{2}-\frac{3}{2})!}{\Gamma(\frac{r}{2})}, \quad r>1,$ and $\sigma_x^2 = \lambda^2\frac{r}{r-2} - \mu_x^2, \quad r>2.$

**V. Univariate t Distribution:** (Raiffa and Schlaifer, 1961, Part III, CH. 7, pp. 232)

A continuous random variable X is said to have a t distribution, with $\nu$ degrees of freedom, location $\mu$ and precision $p$, if the p.d.f. of X is defined by:

$$f(x|\mu,p,\nu) = \frac{p^{\frac{1}{2}}}{\nu^{\frac{1}{2}}B(\frac{1}{2},\frac{\nu}{2})}\left(1+\frac{p(x-\mu)^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}, \quad x,\mu\in R \text{ and } p,\nu>0,$$

and $\mu_x = \mu$ and $\sigma_x^2 = \frac{\nu}{\nu-2}p^{-1} \quad \nu>2,$ where $\mu$ is the location.

**VI. Multivariate t Distribution:** (Raiffa and Schlaifer, 1961, Part III, CH. 7, pp. 256)

Let $\boldsymbol{X}$ be a $k\times 1$ real random vector, then $\boldsymbol{X}$ is said to have a multivariate t distribution, with $\nu$ degrees of freedom, location k-vector $\boldsymbol{\mu}$ and positive definite precision matrix $P$, if the p.d.f. of $\boldsymbol{X}$ is defined by:

$$f(\boldsymbol{X}|\boldsymbol{\mu},P,\nu) = \frac{|P|^{\frac{1}{2}}}{\nu^{\frac{k}{2}}B(\frac{k}{2},\frac{\nu}{2})}\left(1+\frac{(\boldsymbol{X}-\boldsymbol{\mu})'P(\boldsymbol{X}-\boldsymbol{\mu})}{\nu}\right)^{\frac{-(\nu+k)}{2}}, \quad \boldsymbol{X},\boldsymbol{\mu}\in R^k \text{ and } \nu>0,$$

and $\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{\mu}$ and $\boldsymbol{V}(\boldsymbol{X}) = \frac{\nu}{\nu-2}P^{-1} \quad \nu>2.$

**VII. Normal Gamma-I:** (Broamling, 1980, App., pp. 442)

Let X be a real random variable and Y a positive random variable, then X and Y are said to have a Normal Gamma of type I if the density of X and Y is:

$$l(x,y|\mu,\tau,r,\lambda) = \frac{\lambda}{\Gamma(r)}(\lambda y)^{r-1}\exp\{-\lambda y\}\times (2\pi)^{-\frac{1}{2}}(y)^{\frac{1}{2}}\exp\left\{-\tfrac{1}{2}y(x-\mu)^2\right\} \text{ where } x,\mu\in \text{R and r, } \lambda, y>0.$$

such that $x|y \sim \text{Normal}(\mu, y)$ and $y \sim \text{Gamma - I}(r,\lambda).$

**VIII. Normal Inverted-Gamma-II:**

Let X be a real random variable and Y a positive random variable, then X and Y are said to have a Normal Inverted Gamma of type II if the density of X and Y is:

$$l(x,y|\mu,\tau,r,\lambda) = \frac{2(r\lambda^2/2)^{r/2}}{\Gamma(r/2)}y^{-(r+1)}\exp\left\{-r\lambda^2/2y^2\right\}\times (2\pi)^{-\frac{1}{2}}y^{-1}\exp\left\{-\frac{1}{2y^2}(x-\mu)^2\right\},$$

such that $x|y \sim \text{Normal}(\mu, y)$ and $y \sim \text{Inverted Gamma - II}(r,\lambda),$ where $x,\mu\in \text{R and r, } \lambda, y>0.$

# *Appendix-II*

## A Matlab script to simulate from AR(1) for eliciting a reasonable prior distribution

```
% ------------------------------------------------------ +
% Script M-file for the application part of MS.c.        |
% Thesis Title: On the Prior Selection in Bayesian Analysis |
% Created By: Niveen El-Zayat                            |
% First Created Date: 28 Jan. 2007 - 9:00 pm             |
% Last Updated Date:  17 April 2007 - 10:15 am            |
% ------------------------------------------------------ +
cd('D:\Yarab\Thesis Work\Computer Part')
clear
clc
close all
% -------------------------------------------------+
% [1]-Setting the Parameters Values of AR(1) Model |
% -------------------------------------------------+
prompt={'Enter the Sample Size (T):','Enter the Number of Simulated Samples
(N):',...
    'Enter White Noise Variance (Sigma^2):','Enter AR(1) Coefficient (Phi):',...
    'Set the initial value of y0 as:'};
   name='Input for Parameters of AR(1) Model';
   numlines=1;
   defaultanswer={'700','500','1','.5','0'};
   Entry1=inputdlg(prompt,name,numlines,defaultanswer);
%
T=str2num(Entry1{1});
N=str2num(Entry1{2});
Sigm_Sq=str2num(Entry1{3});
phi=str2num(Entry1{4});
y0=str2num(Entry1{5});
%
smpl_length={'30','50','100','200','500'};
Entry2=listdlg('name','Input for Sample Lengths','promptstring',...
          'Enter Sample Length Values','liststring',smpl_length);
for i=1:length(Entry2)
    n_length(i)=str2num(smpl_length{Entry2(i)});
end
% TrSmpl=[.2 .2 .3 .2 .1];
TrSmpl=[.1 .1 .1 .1 .1];
```

```
% ---------------------------------------+
% [2]- Data Generation from AR(1) Process |
% ---------------------------------------+
% state the initial seed of simulation from Normal dist.
randn('state',0);
e=sqrt(Sigm_Sq)*wgn(T,N,2);
%
% set the 1st raw of data equal to the initial values
y(1,:)=y0*ones(1,N);
% shift the white noise matrix 1 raw down that correspond the initial values
e=[zeros(1,N);e];
% Generating the AR(1) Process
for i=1:T
    y(i+1,:)=phi*y(i,:)+e(i+1,:);
end
% Defining the intial values to be the first valye of y's(to be used in
% posterior analysis for stationary AR(1)
Y0=y(2,:);
% supressing the first 200 values to eliminate initial assumption
y(1:200,:)=[];
% ------------------------------------------------------------+
% [3]- Posterior Analysis to AR(1)- Using NonInformative Prior|
% -----------------------------------+----------------------+
% [3-1]- General Case (Jeffreys' Prior)|
% -----------------------------------+
for i=1:length(n_length)

B_J(i,:)=(sum(y(1:n_length(i),:).*y(2:n_length(i)+1,:))./sum(y(1:n_length(i),:).^2
);
    B_rep=repmat(B_J(i,:),n_length(i),1);
    VB_J(i,:)=(sum((y(2:n_length(i)+1,:)-
B_rep.*y(1:n_length(i),:)).^2))./((n_length(i)-3)...

*sum(y(1:n_length(i),:).^2));
end
if n_length<=30 % tabulated value for t-dist
    BU=B_J+2.045*sqrt(abs(VB_J));
    BL=B_J-2.045*sqrt(abs(VB_J));
else           % tabulated value for Normal-dist
    BU=B_J+1.96*sqrt(VB_J);
    BL=B_J-1.96*sqrt(VB_J);
end
Jeff=double(phi>=BL&phi<=BU);
Jeff=100*sum(Jeff,2)/size(Jeff,2);
```

```
%
% ------------------------------------------------------------+
% [2]- Posterior Analysis to AR(1)- Using Informative Prior   |
% ------------------------------------+----------------------+
% [2-1]- General Case (g-Prior)       |
% ------------------------------------+
% Estimation of Hyperparameters Using a Training Sample (35% of the Actual
% Sample)
for i=1:length(n_length)
    n0(i)=floor(.1*n_length(i));
%     g0(i)=log(n_length(i)).^-3;
%     g0(i)=(n_length(i)^(-1/2));
%     g0(i)=(n_length(i)^(-1));
    Mu0(i,:)=(sum(y(1:n0(i),:).*y(2:n0(i)+1,:))./sum(y(1:n0(i),:).^2);
%   Posterior analysis besed on the remaining sample (n-n0)
    n_n0(i)=n_length(i)-n0(i);
%     g0(i)=(n_n0(i)^(-1/2));
    g0(i)=(n_n0(i).^(-1));
%     g0(i)=log(n_n0(i))^-3

B(i,:)=(sum(y(n0(i)+1:n_length(i),:).*y(n0(i)+2:n_length(i)+1,:)))./sum(y(n0(i)+1:n
_length(i),:).^2);
    B_g(i,:)=(B(i,:)+(g0(i).*Mu0(i,:))).*((1+g0(i)).^-1);
    s=((g0(i).*(Mu0(i,:).^2))-
(B_g(i,:).^2).*(1+g0(i))).*sum(y(n0(i)+1:n_length(i),:).^2)+sum(y(n0+2:n_length(i)+
1,:).^2);
    gg=1+g0(i);
    VB_g(i,:)=s./((n_n0(i)-2)*(gg.*sum(y(n0+1:n_length(i),:).^2)));
end

if n_n0<=30 % tabulated value for t-dist
    BU=B_g+2.045*sqrt(VB_g);
    BL=B_g-2.045*sqrt(VB_g);
else           % tabulated value for Normal-dist
    BU=B_g+1.96*sqrt(abs(VB_g));
    BL=B_g-1.96*sqrt(abs(VB_g));
end
g_Prior=double(phi>=BL&phi<=BU);
g_Prior=100*sum(g_Prior,2)/size(g_Prior,2);
% ------------------------------------------------------------+
% [2]- Posterior Analysis to AR(1)- Using Informative Prior   |
% ------------------------------------------------+-----------+
% [2-2]- General Case (Natural Conjugate Prior)|
% ------------------------------------------------+
% Estimation of Hyperparameters Using a Training Sample (10% of the Actual
```

```
% Sample)
for i=1:length(n_length)
    n0(i)=floor(.1*n_length(i));
    aaa=mod(n0(i),2);
    if aaa==1
        n0(i)=n0(i)+1;
    end
    Mu0(i,:)=(sum(y(1:n0(i),:).*y(2:n0(i)+1,:))./sum(y(1:n0(i),:).^2);
    Mu0_rep=repmat(Mu0(i,:),n0(i),1);
    V0(i,:)=(sum((y(2:n0(i)+1,:)-Mu0_rep.*y(1:n0(i),:)).^2))./((n0(i)-3)...
                                                        *sum(y(1:n0(i),:).^2));
    s=sqrt((sum((y(2:n0(i)+1,:)-Mu0_rep.*y(1:n0(i),:)).^2))./(n0(i)-1));
    Esgm0=(s.*sqrt((n0(i)-1)./2).*factorial((n0(i)-1)/2-1.5))./(gamma((n0(i)-
1)./2));
    Vsgm0=((s.^2).*((n0(i)-1)./(n0(i)-3)))-Esgm0.^2;
    r0=3;
    Lmda0=sqrt((r0-2)./r0).*(Vsgm0+Esgm0.^2);
%   Posterior analysis besed on the remaining sample (n-n0)
    n_n0(i)=n_length(i)-n0(i);

B_NC(i,:)=(sum(y(n0(i)+1:n_length(i),:).*y(n0(i)+2:n_length(i)+1,:))+Mu0(i,:).*V0(i
,:))./...

(sum(y(n0(i)+1:n_length(i),:).^2)+V0(i,:));
    VB_NC(i,:)=((r0.*Lmda0.^2)+sum(y(n0(i)+2:n_length(i)+1,:).^2)-...
                (B_NC(i,:).^2).*(sum(y(n0(i)+1:n_length(i),:).^2)+V0(i,:))+...
                ((Mu0(i,:).^2).*V0(i,:))./((n_n0(i)-2).*...
                (sum(y(n0(i)+1:n_length(i),:).^2)+V0(i,:)));
end
if n_n0<=30 % tabulated value for t-dist
    BU=B_NC+2.045*sqrt(abs(VB_NC));
    BL=B_NC-2.045*sqrt(abs(VB_NC));
else           % tabulated value for Normal-dist
    BU=B_NC+1.96*sqrt(abs(VB_NC));
    BL=B_NC-1.96*sqrt(abs(VB_NC));
end
NC_Prior=double(phi>=BL&phi<=BU);
NC_Prior=100*sum(NC_Prior,2)/size(NC_Prior,2);

save Post_AR1
phi
Sigm_Sq
Criterion1=[Jeff g_Prior NC_Prior]
```

# *Appendix-III*

## A Matlab script for obtaining the posterior analysis for some real time series data sets fitted by AR(1) models

```
cd('D:\Yarab\Thesis Work\Computer Part')
clear
clc
close all

% y=xlsread('KABO.xls','D2:D122');
% y=xlsread('CMRBNK.xls','D2:D127');
% y=xlsread('WATNY.xls','E2:E127');

% figure
% subplot(2,1,1)
% autocorr(y,40)
% subplot(2,1,2)
% parcorr(y,40)
%
y0=y(1);
T=length(y)-1; % The 1st observation will be taken as y0 so the whole
               % sample used as data is of size (T-1)
v=T-1;
% ------------------------------------------------------------+
% [1]- Posterior Analysis to AR(1)- Using NonInformative Prior|
% ------------------------------------+----------------------+
% [1-1]- General Case (Jeffreys' Prior)|
% ------------------------------------+
B_J=(sum(y(1:T).*y(2:T+1)))./sum(y(1:T).^2);
B_rep=repmat(B_J,T,1);
VB_J=(sum((y(2:T+1)-B_rep.*y(1:T)).^2))./((v-2)*sum(y(1:T).^2));

% The HDRs
if T<=30 % tabulated value for t-dist
    BU1=B_J+2.045*sqrt(VB_J);
    BL1=B_J-2.045*sqrt(VB_J);
else           % tabulated value for Normal-dist
    BU1=B_J+1.96*sqrt(VB_J);
    BL1=B_J-1.96*sqrt(VB_J);
end
```

```
Jeff_pr=[B_J VB_J]
DHRs_Jeff=[BL1 BU1]
% ------------------------------------------------------------+
% [2]- Posterior Analysis to AR(1)- Using Informative Prior   |
% ------------------------------------+----------------------+
% [2-1]-   General Case (g-Prior)      |
% ------------------------------------+
% Estimation of Hyperparameters Using a Training Sample (10% of the Actual
% Sample)
n0=floor(.1*T);
%     g0=log(n_length(i)).^-3;
Mu0=(sum(y(1:n0).*y(2:n0+1)))./sum(y(1:n0).^2);
%   Posterior analysis besed on the remaining sample (n-n0)
n_n0=T-n0;
% g0=(n_n0^(-1/2));
g0=n_n0^-1;
% g0=log(n_n0)^-3
B=(sum(y(n0+1:T).*y(n0+2:T+1)))./sum(y(n0+1:T).^2);
B_g=(B+g0*Mu0).*((1+g0).^-1);
s=((g0.*(Mu0.^2))-(B_g.^2).*(1+g0)).*sum(y(n0+1:T).^2)+sum(y(n0+2:T+1).^2);
gg=1+g0;
VB_g=s./((n_n0-2)*(gg.*sum(y(n0+1:T).^2)));

% The HDRs
if n_n0<=30 % tabulated value for t-dist
    BU2=B_g+2.045*sqrt(abs(VB_g));
    BL2=B_g-2.045*sqrt(abs(VB_g));
else        % tabulated value for Normal-dist
    BU2=B_g+1.96*sqrt(abs(VB_g));
    BL2=B_g-1.96*sqrt(abs(VB_g));
end
g_pr=[B_g VB_g]
DHRs_g=[BL2 BU2]

% ------------------------------------------------------------+
% [2]- Posterior Analysis to AR(1)- Using Informative Prior   |
% ---------------------------------------------+--------------+
% [2-2]- General Case (Natural Conjugate Prior)|
% ---------------------------------------------+
% Estimation of Hyperparameters Using a Training Sample (10% of the Actual
% Sample)

% Estimation of Hyperparameters Using a Training Sample (10% of the Actual
```

```
% Sample)
n0=floor(.1*T);
aaa=mod(n0,2);
if aaa==1
    n0=n0+1;
end
Mu0=(sum(y(1:n0).*y(2:n0+1)))./sum(y(1:n0).^2);
Mu0_rep=repmat(Mu0,n0,1);
V0=(sum((y(2:n0+1)-Mu0_rep.*y(1:n0)).^2))./((n0-3)*sum(y(1:n0).^2));
s=sqrt((sum((y(2:n0+1)-Mu0_rep.*y(1:n0)).^2))./(n0-1));
Esgm0=(s.*sqrt((n0-1)./2).*factorial((n0-1)/2-1.5))./(gamma((n0-1)./2));
Vsgm0=((s.^2).*((n0-1)./(n0-3)))-Esgm0.^2;
r0=3;
Lmda0=sqrt((r0-2)./r0).*(Vsgm0+Esgm0.^2);
%   Posterior analysisi besed on the remaining sample (n-n0)
n_n0=T-n0;
B_NC=(sum(y(n0+1:T).*y(n0+2:T+1))+Mu0.*V0)./(sum(y(n0+1:T).^2)+V0);
VB_NC=((r0*Lmda0^2)+sum(y(n0+2:T+1).^2)-(B_NC.^2).*(sum(y(n0+1:T).^2)+V0)+...
            ((Mu0^2)*V0))./((n_n0-2).*(sum(y(n0+1:T).^2)+V0));

% The HDRs
if n_n0<=30 % tabulated value for t-dist
    BU3=B_NC+2.045*sqrt(abs(VB_NC));
    BL3=B_NC-2.045*sqrt(abs(VB_NC));
else           % tabulated value for Normal-dist
    BU3=B_NC+1.96*sqrt(abs(VB_NC));
    BL3=B_NC-1.96*sqrt(abs(VB_NC));
end
NC_pr=[B_NC VB_NC]
DHRs_NC=[BL3 BU3]

save Post_AR1_CaseStudy
```

**جامعة القاهرة**
**كلية الاقتصاد والعلوم السياسية**
**قسم الإحصاء**

# حول اختيار التوزيعات القبلية فى التحليل البيزى

**رسالة مقدمة للحصول على درجة الماجيستير فى الإحصاء**

إعداد

**نيفين إبراهيم على الزيات**

تحت إشراف

**ا.د ليلى عثمان الزينى**
استاذ الإحصاء المساعد
قسم الإحصاء
كلية الاقتصاد والعلوم السياسية

**د. عماد الدين عبد السلام سليمان**
مدرس الإحصاء
قسم الإحصاء
كلية الاقتصاد والعلوم السياسية

**القاهرة (2007)**

# الإجــازة

أجازت لجنة المناقشة هذه الرسالة للحصول على درجة الماجيستير فى الإحصاء بتقدير ممتاز بتاريخ 2007/6/28.

بعد استيفاء جميع المتطلبات،

# اللجنـــــــة

| الاسم | الدرجة العلمية | التوقيع |
|---|---|---|
| 1. ا.د. نادية مكارى جرجس | أستاذ الإحصاء المتفرغ بكلية الاقتصاد والعلوم السياسية – جامعة القاهرة | |
| 2. ا.د. عبد الرؤوف عبد الرحمن عبد الواحد | أستاذ الإحصاء بكلية التجارة – جامعة طنطا | |
| 3. ا.د. ليلى عثمان الزيني | أستاذ الإحصاء المساعد بكلية الاقتصاد والعلوم السياسية – جامعة القاهرة | |

**الاسم:** نيفين إبراهيم على الزيات

**الجنسية:** مصرية

**تاريخ وجهة الميلاد:** 1973/9/3 – العجوزة – الجيزة

**الدرجة:** ماجيستير فى الإحصاء

**التخصص:** إحصاء

**المشرفان:**

| | |
|---|---|
| **ا.د ليلى عثمان الزينى** | **د. عماد الدين عبد السلام سليمان** |
| استاذ الإحصاء المساعد | مدرس الإحصاء |
| قسم الإحصاء | قسم الإحصاء |
| كلية الاقتصاد والعلوم السياسية | كلية الاقتصاد والعلوم السياسية |

**عنوان الرسالة:** حول اختيار التوزيعات القبلية فى التحليل البيزى

**ملخص الرسالة:** تعنى الرسالة باستعراض أهم وأبرز الطرق المتبعة فى الأدبيات لإختيار التوزيع القبلى للمعالم المجهولة للنموذج محل الدراسة بغرض إجراء التحليل البيزى. استعرضت الدراسة أهم تلك الطرق وهى الأساليب غير المعلوماتية والأساليب المعلوماتية. تناولت الدراسة شرح أبرز التوزيعات المعروفة فى الأدبيات لكل من الأساليب السابق ذكرها. كما أُبرزت الفلسفة الخاصة بكل توزيع وكيفية اشتقاقه وتطبيقه كما تم سرد عيوب ومزايا كل نوع بالاضافة الى إبراز أوجه الإختلاف والتشابه بين التوزيعات المختلفة إن وُجدت. تم تطبيق بعض التوزيعات القبلية التى تناولتها الدراسة بغرض إجراء التحليل البعدى لنموذج الإنحدار الخطى العام وكذلك لنماذج الإنحدار الذاتى من الرتبة الأولى. بالإضافة إلى ذلك تمت دراسة مدى كفاءة التوزيعات القبلية المختلفة فى التحليل البعدى لنماذج الإنحدار الذاتى من الرتبة الأولى، وتوصلت الى انه ليس هناك توزيع معين هو الأفضل بل الاختيار يتوقف على طول السلسلة الزمنية. علاوة على ذلك فقد تم تطبيق التوزيعات المختلفة لإجراء التحليل البعدى لبعض السلاسل الزمنية الحقيقية والتى تتبع نماذج الإنحدار الذاتى من الرتبة الأولى.

# مستخلـــــــــص

تعنى الرسالة باستعراض أهم وأبرز الطرق المتبعة فى الأدبيات لإختيار التوزيع القبلى للمعالم المجهولة للنموذج محل الدراسة بغرض إجراء التحليل البيزى. استعرضت الدراسة أهم تلك الطرق وهى الأساليب غير المعلوماتية والأساليب المعلوماتية. تناولت الدراسة شرح أبرز التوزيعات المعروفة فى الأدبيات لكل من الأساليب السابق ذكرها. كما أُبرزت الفلسفة الخاصة بكل توزيع وكيفية اشتقاقه وتطبيقه كما تم سرد عيوب ومزايا كل نوع بالاضافة الى إبراز أوجه الإختلاف والتشابه بين التوزيعات المختلفة إن وُجدت. تم تطبيق بعض التوزيعات القبلية التى تناولتها الدراسة بغرض إجراء التحليل البعدى لنموذج الإنحدار الخطى العام وكذلك لنماذج الإنحدار الذاتى من الرتبة الأولى. بالإضافة إلى ذلك تمت دراسة مدى كفاءة التوزيعات القبلية المختلفة فى التحليل البعدى لنماذج الإنحدار الذاتى من الرتبة الأولى، وتوصلت الى انه ليس هناك توزيع معين هو الأفضل بل الاختيار يتوقف على طول السلسلة الزمنية. علاوة على ذلك فقد تم تطبيق التوزيعات المختلفة لإجراء التحليل البعدى لبعض السلاسل الزمنية الحقيقية والتى تتبع نماذج الإنحدار الذاتى من الرتبة الأولى.

**الكلمات الدالة:** التحليل البيزى - التوزيع القبلى – التوزيع البعدى - التوزيعات القبلية غير المعلوماتية - توزيع چيفريز القبلى - Invariance - التوزيع المنتظم المحلى القبلى – دالة الإمكان المحولة للبيانات - التوزيع المعظم لمعلومات العينة القبلى - التوزيعات القبلية المعلوماتية – توزيع چى القبلى - توزيع Natural Conjugate القبلى - نموذج الإنحدار الخطى العام - نماذج الإنحدار الذاتى من الرتبة الأولى.

**المشرفان:**

**ا.د ليلى عثمان الزينى** — استاذ الإحصاء المساعد، قسم الإحصاء، كلية الاقتصاد والعلوم السياسية

**د. عماد الدين عبد السلام سليمان** — مدرس الإحصاء، قسم الإحصاء، كلية الاقتصاد والعلوم السياسية

# ملخـــــص الرســـــالة


تُعد مشكلة أختيار التوزيع القبلى للمعالم المجهولة من أبرز التحديات التى تواجه تطبيق التحليل البيزى فى كثير من المجالات التطبيقية. فأختيار التوزيع القبلى خطوة أساسية لإجراء التحليل البيزى حول المعالم المجهولة بهدف اتخاذ القرار المناسب. حيث يقوم التحليل البيزى بتحديث المعلومات القبلية المتاحة المتمثلة فى التوزيع القبلى فى ضوء المعلومات التى توفرها دالة الإمكان محل الدراسة وصولا الى ما يسمى بالتوزيع البعدى والذى يشمل كل المعلومات الممكن توافرها عن المعالم المجهولة، ومن ثم يستخدم التوزيع البعدى بغرض الاستدلال الإحصائى لمعالم المجتمع.

فيما سبق يتضح الدور الاساسى الذى يمثله اختيار التوزيع القبلى فى بنية التحليل البيزى وهذا ما يبرر ثراء الأدبيات الاحصائية بالعديد من الاساليب والطرق التى طُورت بهدف تحديد كيفية اختيار التوزيع القبلى. وتنقسم تلك الأساليب بشكل عام الى نوعين، اساليب التوزيعات القبلية غير المعلوماتية Noninformative prior approaches، وأساليب التوزيعات القبلية المعلوماتية Informative prior approaches.

وتستخدم التوزيعات القبلية غير المعلوماتية فى حالة عدم توفر او الإفتقار الى المعلومات القبلية عن معالم النموذج وقد لاقت هذه الأساليب قبولا إجماعيا فى الأدبيات حيث انها لا تتطلب آراء شخصية من قبل الباحث، بالإضافة إلى انها تقدم منهجاً تلقائياً لتحديد التوزيع القبلى بناءً على عدد من الخطوات المتعاقبة ومن ثم لا يختلف الباحثون فى النتائج البعدية اذا ما افترضوا نفس التوزيع القبلى. ولعل Jeffreys' prior من أبرز الأنواع التى لاقت شهرةً واستخداماً أكثر شمولاً فى الأدبيات لما يتسم تطبيقه من بساطة فى الاشتقاق وامكانية للتطبيق فى العديد من المجالات. ومن أهم سماته التى أعطته تلك الأهمية تحقيق مبدأ الـ invariance وهو الحصول على توزيعات بعدية متسقة اذا ما تم صياغة النموذج بدلالة دوال فى المعالم الاصلية. وبالرغم من ذلك فان Jeffreys' prior لا يصلح تطبيقه فى بعض المجالات التى تحتوى على معالم غير متجانسة الصفات أو اذا كان مدى المعالم غير مستقل. مما دفع الباحثين إلى البحث عن اساليب أخرى غير معلوماتية لتحديد التوزيع القبلى وتلك الاساليب تختلف فى الفلسفة التى يتبناها كل أسلوب ومن أبرز تلك الأنواع: التوزيع المنتظم المحلى Locally uniform prior والذى تم اقتراحه بواسطة Box and Taio (1973)، وتقوم فكرته على ايجاد توزيع قبلى يحافظ على خصائص دالة الامكان. كذلك من التوزيعات التى تم تطويرها فى الأدبيات ما يسمى بالتوزيع المعظم لمعلومات العينة Maximal data information prior والذى اقترحه Zellner (1977) وتقوم فكرته على إيجاد التوزيع القبلى الذى يحقق مبدأ تعظيم معلومات العينة بالنسبة للمعلومات القبلية مما دفعه لتطوير بعض المعايير الهامة لقياس المعلومات المرتبطة بتوزيع احتمالى معين.

وعلى الصعيد الآخر فإن التوزيعات القبلية المعلوماتية تُستخدم فى حالة توافر معلومات قبلية عن المعالم المجهولة ومن ثم انبثقت العديد من الطرق لتوصيف هذه المعلومات القبلية فى شكل توزيعات احتمالية ونتيجة لتطور الطرق الحسابية باستخدام برامج الحاسب الآلى إرتفعت كفاءة تلك الطرق مما أدى إلى قبولها بعد أن كانت

مرفوضة وغير مستحسنة فى الماضى. ولعل من أبرز الأساليب لتحديد التوزيعات القبلية المعلوماتية Natural Conjugate priors والذى اقترحه Raiffa and Schlaifer (1961)، وتقوم فكرة هذا الأسلوب على اختيار توزيع احتمالى قبلى له نفس خصائص دالة الإمكان ويؤدى الى توزيع إحتمالى بعدى يتسم أيضا بنفس الخصائص. ويتسم هذا النوع بالبساطة فى التحديد وإمكانية أوسع للتطبيق، إلا أن هذا النوع يواجه عائق وحيد هو كيفية تحديد وتقدير معالم التوزيع القبلي وهى ما تسمى بالـ hyperparameters. وعلى الرغم من ذلك طور الباحثون العديد من الطرق التى يمكن بها تقدير معالم التوزيع القبلي ولكنها تخضع للمقارنة واختبارات الكفاءة حتى يمكن اختيار أفضلها. ومن التوزيعات القبلية المعلوماتية التى تم تطويرها أيضا g-prior والذى تم تطويره بواسطة Zellner (1986) والذى قام بتطويره بغرض أجراء التحليل البيزى لنموذج الإنحدار الخطى العام، ويُعد هذا النوع كحالة خاصة من الـ Natural Conjugate priors ولكن يتطلب فقط تحديد معالم التوزيع القبلي الخاصة فقط بالمركز بينما يقترح استخدام مصفوفة التصميم لتقدير المعالم الخاصة بالتغاير.

ولما كان من أهداف الدراسة المقترحة البحث فى الأنواع المعروفة فى الأدبيات للتوزيعات القبلية المعلوماتية وغير المعلوماتية مع إبراز الإختلافات بينها من خلال بعض التطبيقات النظرية المعروفة مثل نموذج الانحدار الخطى العام ونماذج الإنحدار الخطى الذاتى من الرتبة الأولى. كما هدفت الى دراسة كفاءة هذه التوزيعات فى التحليل البعدى لنماذج الإنحدار الخطى الذاتى من الرتبة الأولى. من خلال دراسات المحاكاة. كذلك استعانت الدراسة ببعض السلاسل الزمنية الحقيقية لتوضيح كيفية تطبيق هذه العملية فى الحياة الواقعية.

ونخلص من هذا انه لا يوجد ما يسمى بالتوزيع الأفضل فى التطبيق ويحتاج الاختيار الى معرفة تامة لخصائص معالم النموذج محل الدراسة كما انه يتوقف أيضا على حجم العينة فطالما ان حجم العينة كبير كبراً كافيا فمن الأبسط استخدام Jeffreys' prior حيث ان التوزيعات القبلية المختلفة لن تؤدى الى اختلافات معنوية فى نتائج التوزيع البعدى المنبثق عن كل منها.

وقد انتظمت الدراسة فى ستة فصول نستعرضها فيما يلى:

**الفصل الأول: مقدمة**

ويشتمل على الاطار العام للدراسة ومدى أهمية موضوعها فى عملية التحليل البيزى بالاضافة إلى محتويات واهداف الدراسة المقدمة.

**الفصل الثانى: التوزيعات القبلية غير المعلوماتية**

ويتناول هذا الفصل استعراض لتعريفات التوزيعات القبلية غير المعلوماتية وأهم الدوافع والاسباب الأساسية لاستخدامها، كذلك يستعرض العديد من الأنواع لهذه التوزيعات كما تناولتها الأدبيات. وقد تم التركيز على دراسة ثلاثة أنواع معروفة جيداً فى الأدبيات هى Jeffreys' prior, Locally uniform prior and Maximal data information prior، وقد تم استعراض فلسفة كل نوع وكذلك طرق الاشتقاق كما استعرض الفصل اهم المزايا وأوجه القصور الخاصة بكل نوع على حدة. وكذلك تمت الاستعانة ببعض الأمثلة النظرية لتوزيعات

إحتمالية مختلفة لتوضيح كيفية الاشتقاق النظرى وتمت مقارنة النتائج النظرية الخاصة بكل نوع من الانواع التى تناولتها الدراسة.

**الفصل الثالث: التوزيعات القبلية المعلوماتية**

تناول هذا الفصل نظرة شاملة عن التوزيعات القبلية المعلوماتية وأهم الاسباب لإستخدامها والتعريفات التى تناولتها الأدبيات الخاصة بتلك التوزيعات، كما تناول أهم الآدبيات المكتوبة لتطوير انواع متعددة لهذه التوزيعات. وتركزت الدراسة فى هذا الفصل على تناول Natural conjugate prior and g-prior بشكل اكثر تعمقاً يشمل تعريف كل نوع وكيفية اشتقاقه وخصائصه.

**الفصل الرابع: التحليل البعدى لنماذج الانحدار الخطى العام**

تم إجراء التحليل البعدى لنموذج الإنحدار الخطى العام استناداً إلى التوزيعات القبلية المعلوماتية التى تناولها بالتحليل الفصل السابق. وتم تلخيص أبرز الفروق للنتائج النظرية المترتبة على كل نوع.

**الفصل الخامس: التحليل البيزى للسلاسل الزمنية لنماذج الإنحدار الذاتى من الرتبة الأولى**

تناول هذا الفصل مقدمة حول المفاهيم الأساسية لنماذج الإنحدار الذاتى من الرتبة الأولى، كما تم إجراء التحليل البعدى باستخدام بعض التوزيعات القبلية المعلوماتية وغير المعلوماتية التى تم تناولها بالدراسة الفصلين الثانى والثالث. بالإضافة إلى المقارنة النظرية فقد تم أختبار مدى كفاءة التوزيعات القبلية المختلفة فى التحليل البعدى لنماذج AR(1) باستخدام اسلوب المحاكاة، وتمت المقارنة باستخدام بعض معايير الكفاءة. كذلك تناول هذا الفصل تطبيق التوزيعات القبلية المختلفة لإجراء التحليل البعدى لثلاث سلاسل زمنية حقيقية خاصة بأسعار البورصة لبعض الشركات المصرية. واستعرض الفصل نتائج دراسة المحاكاة وكذلك نتائج تحليل السلاسل الزمنية الحقيقية المستخدمة ببعض الجداول والرسوم التى ساعدت فى تلخيص نتائج الدراسة.

**الفصل السادس: النتائج والدراسات المستقبلية**

ويتناول ملخص الدراسة والنتائج التى توصلت إليها، كما يُشار فيه إلى بعض نقاط البحث التى يمكن تناولها فى دراسات مستقبلية.

**الملاحق:**

ولقد ذُيلت الدراسة بثلاثة ملاحق: الأول يتناول عرض لأهم التوزيعات الإحتمالية التى أشارت اليها الدراسة والمستخدمة فى أغلب الاشتقاقات مثل توزيع جاما وتوزيع جاما العكسى بأنواعهما المختلفة وكذلك توزيع ت المتعدد وأيضا توزيع جاما المعتاد وتوزيع جاما المعتاد العكسى. بينما تضمنت الملاحق الثانى والثالث برامج Matlab التى أستخدمت لغرض أجراء دراسة المحاكاة وتحليل البيانات الواقعية اللتان تم الاشارة اليهن فى الفصل السابق.