Estimating and Correcting the Effects of Model Selection Uncertainty

Dissertation

Presented for the Degree of Doctor of Philosophy at the Faculty of Economics and Business Administration of the University of Göttingen

by

Georges Lucioni Edison Nguefack Tsague

from Fongo-Tongo, Cameroon.

Göttingen, 2005

First Examiner:	Prof. Dr. Walter Zucchini
Second Examiner:	Prof. Dr. Fred Böker
Date of oral exam:	03.02.2006

To Arend Pecresse.

Acknowledgements

I am most grateful to my supervisor, Prof. Dr. Walter Zucchini, for providing me with many hours of guidance and stimulating discussions, for his patience and encouragement. I would also like to thank him for introducing me to the interesting world of model selection uncertainty, especially for the good starting point.

I wish to thank Prof. Dr. Fred Böker and Prof. Dr. Stephan Klasen for accepting to be examiners for this thesis.

This work was completed within the Ph.D. program of the Center for Statistics, University of Göttingen. I wish to thank the members of the Center, especially the speaker, Prof. Dr. Manfred Denker for providing me with the financial support.

I gratefully acknowledge all faculty members, staff and Ph.D. students of the Institute of Statistics and Econometrics at University of Göttingen for the good working atmosphere during my stay.

Finally, I am grateful to my family, especially to my wife, Nicole, and our daugther Arend Pecresse, for their understanding and encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

AAW	Adjusted Akaike Weights
ABMA	Adjusted Bayesian Model Averaging
AIC	Akaike Information Criterion
ALW	Adjusted Likelihood Weights
ANOVA	Analysis Of Variance
BIC	Bayesian Information Criterion
BMA	Bayesian Model Averaging
BPMSE	Bayesian Post-Model-Selection Estimator
CDF	Cumulative Distribution Function
CI	Confidence Interval
EDA	Exploratory Data Analysis
FBMA	Fully Bayesian Model Averaging
FMA	Frequentist Model Averaging
IC	Information Criterion
HQ	Hannan and Quinn
MA	Model Averaging
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
OLS	Ordinary Least Square
PDF	Probability Density Function
PMF	Probability Mass Function
PMSE	Post-Model-Selection Estimator

Chapter 1

Introduction and Objective

1.1 Background and motivation

Many (possibly most) statistical analyses involve model selection, in a process referred to as model building. Often, selection is an iterative process, carried out by applying a series hypothesis tests. These are used to decide on the appropriate complexity of the model, whether certain covariates should be excluded, whether some of them should be transformed, whether interactions should be considered, and so on. A variety of additional methods have been specifically developed for model selection, both in the frequentist and the Bayesian frameworks. For an overview of model selection criteria, one may consult the monographs by Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Burnham and Anderson (2002) and the paper by Claeskens and Hjort (2003).

After a model has been selected, one usually proceeds with inference as if this model had been known in advance, ignoring the fact that model has been selected using the same data. Although it has been known for some time that this “double use” of the data leads to invalid inference, this fact is not taken into account in the vast majority of applications. A possible explanation is that the issue is seldom discussed in typical Statistics courses, especially in service courses offered to non-specialists. The problem is complex and not yet well understood; it is not clear, even to statisticians, how to carry out valid inference following model selection.

The bias due to not taking model selection into account is referred as *selection bias* (Miller, 1990; Zucchini, 2000) or *model selection bias* (Chatfield, 1995). The act of using the same data for model selection and for parameter estimation is referred as *model selection uncertainty* (Hjorth, 1994). We will use the term *model selection uncertainty* to refer to situations in which the true model is not known, where a model is selected using the data, and then the selected model is used to draw inferences, or to reach decisions.

A known consequence of ignoring model selection uncertainty is that, in general, the selected model appears to fit better than it does (*optimism principle*). For example, the estimated variance of estimator is likely to be too small, the confidence and prediction intervals are likely to be too narrow. Estimators obtained after a selection procedure has been performed are referred as *estimators-post-selection* (Hjort and Claeskens, 2003), or *post-model-selection estimators* (Leeb and Pötscher, 2005).

Since the problem is due to using the data twice, one could consider splitting the data into two sets; to use one set for model selection and the other for inference. Such a procedure has a serious drawback; it leads to a loss of information. This is undesirable, even unacceptable, especially when the sample size is small.

The severity and seriousness of the problem of model selection uncertainty can be appreciated by reading some of the remarks that have been written on the subject.

- Breiman (1992), p.738: "A **quiet scandal** in the statistical community."

- Chatfield (1995), p.421: "Statisticians admit this **privately**, but they(we) continue to ignore the difficulties because it is not clear what else could or should be done."

- Pötscher (1995), p.461: "This old and nagging problem."

- Buckland *et al.* (1997): "It seems surprising that more authors have not addressed this issue. In some fields, it would seem essential that the issue be addressed."

- Zucchini (2000), p.58: "The objectivity of formal model selection procedures and the ease with which they can be applied with increasing powerful computers on increasing complex problems has tended to obscure the fact that too much selection can do more harm than good. An overdose of selection manifests itself in a problem called selection bias which occurs when one uses the same data to select a model and also to carry out statistical inference [...] The solution is still being invented."

- Hjort and Claeskens, 2003, p.879: "There are at least two clear reasons fewer efforts have been devoted to these questions than to the primary ones related to finding 'one good model'. The first is that the selection strategies actually used by statisticians are difficult to describe accurately, as they involve many, partly nonformalized ingredients such as 'looking at residuals' and 'trying a suitable transformation'. The second is that these questions of estimator-post-selection behaviour simply are harder to formalize and analyse."

- Efron (2004), p.640: "Classical statistics as developed in the first half of the $20th$ century has two obvious deficiencies from practical applications: an overreliance on the normal distribution and failure to account for model selection. The first of these was dealt with in the century's second half [...] Model selection, the data-based choice [...] remains mostly *terra incognita* as far as statistical inference is concerned."

The above remarks summarize the motivation for the investigation described in this thesis. Our general objective is to contribute to an improved understanding of this problem. Our specific objectives are outlined in Section 1.3.

1.2 Related work

The literature that is relevant to this thesis can be divided into two categories: The first is concerned with the situation in which the data has been used to select a model and then to estimate some quantity of interest. The general aim of that literature has been to discover the properties of the post-model-selection estimators (PMSEs). The second category, model averaging, is about estimators that are not based on a single selected model, but rather on a weighted average of estimators from all the models under consideration.

In this section we briefly outline the main milestones; specific contributions will be acknowledged in the main text.

1.2.1 Post-model-selection estimators

Bancroft (1944) investigated the bias introduced by pre-testing the regression coefficients and the homogeneity of variance. A special case of Bancroft (1948) is given by Mosteller (1948) where the mean square error of pre-test estimator is found. This result was later extended by Huntsberger (1955). Sclove et al. (1972) pointed out the undesirable properties of pre-test estimators. The monograph of Judge and Bock (1978) discussed the pre-test properties in detail. Risk properties of pre-test can also be found in Lovell (1983), Roehrig (1984), Mittelhammer (1984), Judge and Bock (1983), Judge and Yancey (1986), Dijkstra (1988). These developments are summarised in Chatfield (1995), and Magnus and Durbin (1999). Danilov and Magnus (2004) gave the first and second moments of the pre-test estimators, and showed that the error of not reporting the correct moment can be large. A description of the pre-test problem is also given in Longford (2005).

Distributional properties of PMSEs are considered by Sen (1979), Sen and Saleh (1987), Dijkstra and Veldkam, Pötscher (1991), Giles and Srivastava (1993), Kabaila (1995,1998), Pötscher (1995), Pötscher and Novak (1998), Ahmed and Basu (2000), Kapetanios (2001), Dukić and Peña (2002), Hjort and Claeskens (2003), Leeb and Pötscher (2003, 2005), Bunea (2004).

1.2.2 Model averaging

Bernard (1963) mentioned model combination in the statistical literature in the framework of studying airline passenger data. Bates and Granger (1969) studied how to combine predictions from different forecasting models. Roberts (1965) suggested combining the opinions of experts in which the weights are the posterior probabilities of the models.

A formal Bayesian solution to model uncertainty dates to Leamer (1978) in which the posterior distribution was explicitly stated. This was the starting point for Bayesian model averaging (BMA). Madigan and Raftery (1994) introduced Occam's window method, to reduce the set of competing models. Draper (1995) advocated the same Bayesian model averaging methods with the idea of model expansion. Chatfield (1995), Kass and Raftery (1995) reviewed BMA, and the cost of ignoring model uncertainty. Raftery et al. (1997) studied BMA in the context of linear regression models. George (1999) discussed BMA in the framework of decision theory. Hoeting et al. (1999) described methods of implementing BMA, and gave practical applications. Merlise and George (2004) discussed general issues on model uncertainty.

In the classical literature, Akaike (1978) defined the concept of the *likelihood of a model* and proposed that this be used to determine the weights when selecting autoregressive models for time series. Leblanc and Tibshirani (1996) use likelihood weights in the context of linear regression. Buckland et al. (1997) proposed using Akaike weights and bootstrap weights as a method of incorporating model uncertainty. Strimmer and Rambaut (2001) used the bootstrap of the likelihood weights, and applied these to gene trees analysis. Candolo et al. (2003) accounted for model uncertainty using Akaike weights. Frequentist approach for model averaging is given in Hjort and Claeskens (2003). They give a general large sample theory for model averaging estimators, including PMSEs, together with their limiting distributions and risk properties.

1.3 Specific objectives

In this thesis we are mainly concerned with inference after model selection, that is, to understand how estimators behave if estimation is preceded by model selection based on the same data. Our objective is to examine the real effects of model selection uncertainty, and how these effects can be corrected. To achieve this we investigate a number of issues that seem not to have been fully investigated in the literature:

1. The frequency (or unconditional) performance of model averaging methods, in particular Bayesian model averaging (BMA); the Bayesian nature of Bayesian model averaging.

2. The differences and similarities between model averaging and model selection, and whether, in terms of a measure of risk, model averaging methods are a better alternative to model selection.

3. To describe a framework that connects model averaging and model selection, both in the frequentist framework and in the Bayesian.

4. To give simple examples in which the properties of PMSEs can be derived and compared analytically, not only under pre-test selection, but with any selection criterion.

5. To identify the key ingredients that complicate the model selection uncertainty problem, and to investigate whether the use of consistent selection criteria "solves" the problem.

6. To assess whether any specific model selection criterion can be generally recommended, i.e. leads to better post-model-selection estimation.

7. To investigate the extent to which Bayesian model selection can be affected by the model selection uncertainty problem.

8. To illustrate the model uncertainty problem in the framework of parameter estimation.

9. To assess whether bootstrap methods can be used to correct for model selection uncertainty.

1.4 Outline of the thesis

In Chapter 2 we consider the problem of *model uncertainty*. We study an approach, known as *model averaging*, that is intended to deal with the problem. The idea is to avoid the use of a single model to estimate the quantity of interest; instead one uses a weighted average of the estimates obtained using all the models under consideration. Model averaging can be carried out either in a Bayesian or in a frequentist setting. In this chapter we focus mainly on the former, and investigate its theoretical properties, specifically its conditional properties (given the data), its unconditional (frequentist) properties and its predictive performance. We argue that, regarded unconditionally, in general, it is hard to establish that current BMA estimators are truly Bayesian estimators. Therefore, their frequentist performances (e.g. admissibility, minimaxity) are likely to be unknown. We also argue that for model averaging in general, the properties of model averaging estimator cannot be assessed unless one assumes some underline model. However, there is uncertainty about the choice of this model and it is precisely this uncertainty that led to model averaging or model selection. Under such an assumption, one would simply use that model without applying model selection or model averaging. The same issue arises in the case of post-model-selection estimation to be discussed in Chapter 3, and also when assessing the properties of bootstrap-after-model-selection estimator discussed in Chapter 7. We provide an illustration of an alternative method of weighting that provides a *Fully Bayesian model averaging* (FBMA) approach when the quantity of interest is parametric.

In Chapter 3 we consider the issue of model selection. As in Chapter 2, we assume that a set of alternative models is available, but that we will select a single model to carry out estimation. We also assume that the same data is used both for selecting the model and for estimation. Clearly, from a statistical point of view, this *post-model-selection estimation* approach is different from the model averaging approach considered in Chapter 2. The foundation of the problem is identified and formulated in a probability framework that allows us to investigate it theoretically. Properties of PMSEs are described for some simple cases, and various model selection criteria are compared. The issue of consistency in model selection is also discussed, and the effect of sample size is investigated.

Chapters 4 and 5 are about the issue of correcting for *model selection uncertainty*; the former discusses the problem from the frequentist point of view, and the latter from the Bayesian. We point out that, mathematically, post-model-selection estimation is simply a special case of model averaging, and so these two approaches can be compared within a single framework. Model selection and model averaging are compared, and an alternative scheme is proposed for dealing with model selection uncertainty. We define *Adjusted Akaike Weights* and *Adjusted Likelihood Weights.* These are introduced to take model selection into account in classical model averaging.

Chapter 5 investigates corrections for model selection uncertainty in a Bayesian framework. Conditional on the data, there is no model selection uncertainty problem, only model uncertainty. We point out that, if the estimators are viewed *unconditionally* and if a model is selected, then the problem of model selection uncertainty does arise. An alternative model weighting approach, which does take the selection procedure into account, is proposed. The approach, which is based on *prior model selection probabilities*, is illustrated using a simple example involving the estimation of proportions.

In Chapter 6 we investigate model selection uncertainty in the context of parameter estimation within a single parametric model family. This offers an alternative interpretation to a number of well-known distributional results. We illustrate that these can be regarded as solutions to the model selection uncertainty problem. In particular we show that profile likelihood, and nuisance parameter problems are interpretable in this framework.

Chapter 7 is concerned with the applicability of bootstrap methods to deal with model selection uncertainty. It is relatively easy to apply the bootstrap to assess the properties of PMSEs. However, by means of a concrete theoretical example, we illustrate that the resulting estimator can be poor. We identify the reason for this failure as the poor performance of the bootstrap in estimating model selection probabilities.

Chapter 8 summarises the main findings of the thesis and suggests possible extensions for future research work.

Chapter 2

Model Uncertainty and Model Averaging

2.1 Introduction

Consider a situation in which some quantity of interest, $\triangle$, is to be estimated from a sample of observations that can be regarded as realizations from some unknown probability distribution, and that in order to do so, it is necessary to specify a model for the distribution. There are usually many alternative plausible models available and, in general, they each lead to different estimates of $\triangle$. The term *model uncertainty* is used when it is not known which model correctly describes the probability distribution under consideration. A discussion on the issue of model uncertainty is given in, e.g., Clyde and George (2004).

In this chapter we will discuss a strategy, known as *model averaging*, that is used to deal with the problem of model uncertainty. The idea is to use a weighted average of the estimates of $\triangle$ obtained using each of the alternative models, rather than the estimate obtained using any single model. This is implemented both in the frequentist and in the Bayesian framework.

The main problem to be solved, when applying model averaging, is that of selecting the weights for the estimates obtained using the different models. Ideally one would wish to use weights which minimize some specified criterion, or "loss function". We point out that, in general, it is not feasible to determine optimal weights from the available information because these depend on the unknown true distribution, i.e. the distribution for the entire population, not just the sample.

We investigate the theoretical performance of the well-known Bayesian model averaging (BMA) from different points of view. We argue that some issues regarding BMA have not been clearly described in full.

We begin by considering BMA conditioned on the data and point out that the performance of the BMA estimate cannot be compared with that of any "single-model" estimate. Each of the latter has its own posterior and is optimal with respect to that. Similarly the BMA estimate is optimal with respect to its posterior, which is a weighted average of the posteriors of the individual single-model estimates.

We then consider *unconditional* performance, also called "long-run" or frequentist properties. By frequentist properties we mean the properties of an estimator over repeated sampling, and not those conditioned on a particular data set. If there are K models, each of which leads to an estimator of the quantity of interest, then the BMA estimator constitutes an additional estimator; i.e. one has to consider $K+1$ estimators. Even though the frequentist properties (e.g. admissibility, minimaxity) of each of the K individual models are known (since they are Bayes estimators), these do not suffice to determine the frequentist properties of the BMA estimator (except for simple parametric cases). The reason is twofold: no prior has been assigned to this $(K+1)$-st estimator, the BMA. Secondly, the distribution of the data under the BMA model hasn't been specified; only the posterior is known. Thus it is hard to show whether BMA is a fully Bayesian method; it is not based on a well-defined data-generating mechanism, i.e. a true model, which is required if one wishes to assess its frequentist performance. We will refer to it as quasi-Bayesian.

Thirdly we consider the predictive performance of BMA estimation. In the Bayesian literature this is often measured in terms of Good's (1952) "logarithm score rule" and is used to justify the use of BMA. We argue that, due to the non-negativity of the Kulback-Leibler information divergence, such a justification hinges critically on the assumption that the BMA model is the "true model". Assuming that any other model is true would automatically render the BMA non-optimal. In other words the *theoretical* justification for BMA in terms of its predictive performance is tautological. Its practical performance in applications and simulations have been reported to be favourable (see, e.g., Clyde (1999), Clyde and George (2000)).

We introduce a simple *fully Bayesian model averaging* (FBMA) approach based on a mixture of priors, and a mixture of parametric models instead of starting with posterior distribution, as in the case of BMA. This leads to a method that is Bayesian in the strict sense. The advantage of FBMA is that both its conditional (given the data) and unconditional (prior to seeing the data) performance are available, at least in theory, as is the case with standard Bayesian inference derived in the context of a fully Bayesian statistical model. We provide an illustration of a simple situation in which BMA is Bayesian, FBMA reducing to BMA.

To reduce the enormous computational effort required to apply BMA it has been suggested that some models be eliminated in a "preselection" step. Suggestions include Occam's window, Markov chain Monte Carlo model composition and stochastic search variable selection. We stress the fact that the long-run performance of BMA estimators will be affected if *data-based* model search methods are applied. This introduces an additional source of uncertainty which we call *model space selection uncertainty.* The application of preselection changes the estimator, and therefore its properties. It is necessary to take that source of additional uncertainty into account. For *posterior* analysis, i.e. conditioned on the data, such search strategies present no problem.

We next turn briefly to frequentist model averaging (FMA), in particular to the use of Akaike weights. We show how Akaike weights can be interpreted in the context of Akaike's (1978) predictive approach, and his concept of the "likelihood of a model". We illustrate how Akaike weights can be implemented in practice.

2.2 Model averaging and optimal weights

2.2.1 Model averaging

Suppose that the observations, x, have been generated by the model M_t. For example, x could be a random sample from a well-defined finite population. Then the M_t, the true model, is the distribution over the entire population.

Let $\mathcal{M} = (M_1, \ldots, M_K)$ be a set of K models and $\triangle$ the quantity of interest. Let $\hat{\triangle}_k$ be the estimator of $\triangle$ (using some specified estimation procedure) when model M_k is used. We will sometimes refer to the quantity $\triangle$ under model k as $\triangle_k$, $k = 1, 2, ..., K$.

The application of model averaging involves finding non-negative weights, $\pi = (\pi_1, \ldots, \pi_K)'$, that sum to one, and then estimating $\triangle$ by

$$\tilde{\triangle}_{\mathrm{MA}} = \sum_{k=1}^{K} \pi_k \hat{\triangle}_k. \tag{2.1}$$

The question that arises is whether one can select the weights so as to optimize the performance of this averaged estimator, in terms of some specified measure, say a loss function L. Finding the optimal weights involves solving the following optimisation problem over π:

$$\min_{\pi} \mathrm{E}_t L(\sum_{k=1}^{K} \pi_k \hat{\triangle}_k, \triangle_k), \quad \pi_k \geq 0, \quad \forall k; \quad \sum_{k=1}^{K} \pi_k = 1, \tag{2.2}$$

where the expectation is taken with respect to the true model, which may, or may not, be in the set of competing models, $\mathcal{M}$.

The expectation in (2.2) has to be taken with respect to the true model, M_t, which is unknown. One is therefore not in a position to obtain optimal weights; these have to be estimated and so the performance of the weighted estimator will depend on a variety of factors, such as how the weights are estimated.

2.2.2 Performance of model averaging estimators

One important problem associated with model averaging is that of evaluating the performance of the average estimator. Each estimator $\hat{\triangle}_k$ is derived under model M_k, therefore, the properties (e.g. mean, variance, MSE) of $\hat{\triangle}_k$ can be computed under this model. In general the weights are obtained using the data (i.e. they are estimated) and the model averaging estimator is

$$\hat{\triangle}_{\mathrm{MA}} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\triangle}_k. \tag{2.3}$$

The model M_{MA} from which the average estimator $\hat{\triangle}_{\mathrm{MA}}$ is derived is not known. To derive the properties of $\hat{\triangle}_{\mathrm{MA}}$, one needs to assume a model M_t, then obtains its properties under this model. There is no guarantee that the resulting weighted estimator will outperform every individual estimator $\hat{\triangle}_k$. However, there is uncertainty about the choice of this model and it is precisely this uncertainty that led to model averaging. The fact of not knowing M_t that generated $\hat{\triangle}_{\mathrm{MA}}$ leads to the difficulty of interpreting it. For instance, suppose that $\hat{\triangle}_k$ is the MLE of $\triangle_k$ for model M_k. The likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model M_k. The MLE is the parameter point for which the observed sample is most likely, that is the value of these parameters that maximizes the likelihood. Now, how can one interpret $\hat{\triangle}_{\mathrm{MA}}$ without the generating model? Therefore knowledge of the (long run) properties of model averaging estimators, even with an assumed true model is computationally difficult. Without knowing the generating model, the properties of $\hat{\triangle}_{\mathrm{MA}}$ are not defined. For example, one cannot compute the expectation of $\hat{\triangle}_{\mathrm{MA}}$ without specifying the distribution with respect to which this expectation is to be taken. To illustrate the point, we consider a mixture density problem.

Mixture models arise when an observation x is taken from a population composed of different subpopulations. The problem is that one does not know from which of these the observation is taken. Let K be the number of subpopulations, then X has a K-component mixture density

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x), \quad 0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1, \tag{2.4}$$

where π_k is the probability that x comes from the k^{th} subpopulation and $f_k(x)$ is the density of X in the k^{th} subpopulation. Let $\triangle$ the quantity of interest be the mean of the X. It is straightforward to see that under the mixture (2.4), the mean is given by

$$\mathrm{E}(X) = \sum_{k=1}^{K} \pi_k \mathrm{E}_k(X), \tag{2.5}$$

where the E stands for the expectation under the mixture (2.4) and E_k for the expectation under the k^{th} subpopulation. An estimator of the expectation of X is given by

$$\hat{\triangle} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\triangle}_k, \tag{2.6}$$

where $\hat{\triangle} = \hat{\mathrm{E}}(X)$ and $\hat{\triangle}_k = \hat{\mathrm{E}}_k(X)$.
Consider $\hat{\triangle}$ to be an average estimator. When the subpopulations have different parametric forms, methods exist (e.g. EM algorithm, Newton-type method) to compute $(\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\triangle}_1, \ldots, \hat{\triangle}_K)'$. The properties of the average estimator can be obtained under the mixture model (2.4). In this case, only the computation is the challenging issue. In repeating experiments, the data can be sampled from (2.4). Now, consider a weighted estimator given by a different weighting scheme $(\tilde{\pi}_1, \ldots, \tilde{\pi}_K)'$ as usual with model averaging,

$$\hat{\triangle}_{\mathrm{MA}} = \sum_{k=1}^{K} \tilde{\pi}_k \hat{\triangle}_k. \tag{2.7}$$

Model averaging estimator (2.7) looks similar to (2.6). However, in the case of model averaging, the properties of $\hat{\triangle}_{\mathrm{MA}}$ cannot be assessed unless one assumes some underline model.

The same issue arises in the case of post-model-selection estimation to be discussed in Chapter 3, and also when assessing the properties of bootstrap model selection discussed in Chapter 7. In fact, the properties of an estimator are well defined if one computes with respect to the model from which this estimator is derived.

2.3 Bayesian model averaging

2.3.1 Description

Consider a sample of data, x, and a set of K models $\mathcal{M} = (M_1, \ldots, M_K)$, which we will assume to contain the true model M_t. Each model M_k consists of a family of distributions $P(x|\theta_k, M_k)$, where θ_k represents a parameter (or vector of parameters).

To implement a BMA procedure we begin by assigning a prior probability, $P(M_k)$, to the event that model M_k is the true model, and a prior distribution, $P(\theta_k|M_k)$, to the parameters of model M_k, given that M_k is true, $k = 1, \ldots, K$. As outlined in Chipman, George and McCulloch (2001), the data generating process proceeds in the following three stages:

1. generate a model, M_k, from $P(M_1), \ldots, P(M_K)$,
2. generate a parameter, θ_k, from $P(\theta_k|M_k)$,
3. generate the data, x, from $P(x|\theta_k, M_k)$.

Conditioning on the data x and integrating out the parameter θ_k, one has posterior model probabilities:

$$P(M_k|x) = \frac{P(x|M_k)P(M_k)}{\Sigma_{j=1}^{K} P(x|M_j)P(M_j)}, \tag{2.8}$$

where

$$P(x|M_j) = \int_\Theta P(x|\theta_k, M_k)P(\theta_k|M_k)d\theta_k \tag{2.9}$$

is the integrated likelihood under model M_k. If $P(\theta_k|M_k)$ a discrete distribution, the integral in (5.18) is replaced by a sum.

Let $\triangle$ be a quantity of interest, for example a future observation from the same process that generated x. Then the posterior distribution of $\triangle$ is given by

$$P(\triangle|x) = \Sigma_{k=1}^{K} P(\triangle|x, M_k)P(M_k|x). \tag{2.10}$$

We note that $P(\triangle|x)$ is a weighted average of the posterior distributions $P(\triangle|M_k, x)$, $k = 1, ..., K$, where the k-th weight, $P(M_k|x)$, is the posterior probability that model M_k is the true model. The posterior distribution of $\triangle$, conditioned on model M_k being true, is given by

$$P(\triangle|x, M_k) = \int_\Theta P(\triangle|\theta_k, M_k)P(\theta_k|x, M_k)d\theta_k. \tag{2.11}$$

The posterior mean and posterior variance are given by

$$\begin{aligned}\hat{\triangle}_{\text{bma}} &= \mathrm{E}_{\text{bma}}(\triangle|x) = \Sigma_{k=1}^{K}\mathrm{E}(\triangle|x, M_k)P(M_k|x),\\ \mathrm{Var}_{\text{bma}}(\triangle|x) &= \Sigma_{k=1}^{K}[\mathrm{Var}(\triangle|x, M_k) + (\mathrm{E}(\triangle|x, M_k) - \hat{\triangle}_{\text{bma}})^2]P(M_k|x).\end{aligned} \tag{2.12}$$

Raftery et al. (1997) call this averaging scheme *Bayesian model averaging*. Learmer (1978) and Draper (1995) advocate the same idea. Madigan and Raftery (1994) note that BMA provides better predictive performance than any single model if the measure of performance is the logarithm score rule of Good (1952), under the posterior distribution of θ given x. Hoeting et al. (1999) give an extensive framework of BMA methodology and applications for different statistical models. Various real data and simulation studies (e.g. Clyde (1999), Clyde and George (2000)) have investigated the predictive performance of BMA.

Implementing BMA is demanding, especially the computation of the integrated likelihood. Software for BMA implementation, as well as some BMA papers, can be found at "http://www.research.att.com/∼volinsky/bma.html". For computations, Monte Carlo methods, or approximating methods, are used, Thus many BMA applications are based on the BIC, an asymptotic approximation of the log posterior odds when the prior odds are all equal. Another problem is the selection of priors both for models and parameters. In most cases, a uniform prior is used for each model, i.e. $P(M_k) = 1/K$, $k = 1, 2, ..., K$. When the number of models is large, model search strategies are sometimes used to reduce the set of models, by eliminating those that seem comparatively less compatible with the data. Of course, such data-based "preselection methods" are not strictly Bayesian, and secondly, the potential effects of preselection are ignored in BMA, at least as it is currently being implemented.

2.3.2 Theoretical performance of BMA

2.3.2.1 Conditioning on the data

For each model M_k, short run performance of an estimate $\delta_k(x)$ can be measured by the posterior expected loss

$$\rho(\delta_k(x)) = \mathrm{E}_k[L(\triangle, \delta_k(x))] = \int_{\Lambda} L(\triangle, \delta_k(x))P(\triangle|x, M_k)\, d\triangle,$$

where L is a loss function.
Since $\hat{\triangle}_k = \mathrm{E}(\triangle|x, M_k)$ is a Bayes estimate for M_k, it is the only decision rule with minimal posterior expected loss. BMA estimate $\hat{\triangle}_{\text{bma}}$ is the only decision

rule with minimum posterior expected loss with respect to the posterior distribution $P(\triangle|x)$ given in (2.10). This means that if one needs to compare the performances of BMA estimate with any of the estimate $\hat{\triangle}_k = \mathrm{E}(\triangle|x, M_k)$, one model should be used as reference. Since the true model is assumed to be one of the competing models, the comparison should be made with respect to the true model M_t. BMA performs better than any model M_k if the following holds

$$\rho(\hat{\triangle}_{\mathrm{bma}}) = \mathrm{E}_t[L(\triangle, \hat{\triangle}_{\mathrm{bma}})] \leq \rho(\hat{\triangle}_k) = \mathrm{E}_t[L(\triangle, \hat{\triangle}_k)], \tag{2.13}$$

for $k = 1, \ldots, K$, $k \neq t$. It is important to note that the expectation in (2.13) is taken with respect to the same model M_t.

2.3.2.2 Frequentist properties

The long run performance of each model M_k with an estimate $\delta_k(x)$ can be measured by the average loss (frequentist loss) given by

$$\mathrm{R}(\triangle, \delta_k(x)) = \mathrm{E}_k[L(\triangle, \delta_k(x))] = \int_{\mathcal{X}} L(\triangle, \delta_k(x)) P(x|\triangle, M_k)\, dx.$$

BMA is better than any single model M_k if

$$\mathrm{R}(\triangle, \hat{\triangle}_{\mathrm{bma}}) = \mathrm{E}_t[L(\triangle, \hat{\triangle}_{\mathrm{bma}})] \leq \mathrm{R}(\triangle, \hat{\triangle}_k) = \mathrm{E}_t[L(\triangle, \hat{\triangle}_k)], \tag{2.14}$$

for $k = 1, \ldots, K$, $k \neq t$.
The expectation in (2.14) is taken with respect to the model M_t. If one is able to find out the prior and the statistical model associated to BMA estimator, this will be a Bayes estimator. In this case, the long run performances are known.

2.3.2.3 Predictive performance

One measure of predictive performance is the Good's (1952) logarithm score rule. From the nonnegativity of Kullback-Leiber information divergence, it follows that if f and g two probabilities distribution functions,

$$\mathrm{E}_f(\log f(X)) \geq \mathrm{E}_f(\log g(X)).$$

Applying this to model averaging and model space, we have that

$$\mathrm{E}_{\mathrm{bma}}[\log P(\triangle|x)] \geq \mathrm{E}_{\mathrm{bma}}[\log P(\triangle|x, M_k)], \quad k = 1, \ldots, K \tag{2.15}$$

and it also holds that

$$\mathrm{E}_k[\log P(\triangle|x, M_k)] \geq \mathrm{E}_k[\log P(\triangle|x)], \quad k = 1, \ldots, K. \tag{2.16}$$

This means that, it is not possible to measure the performance of BMA using only (2.15). The expectation should be taken with respect to the true model. BMA will perform better than any single model if

$$\mathrm{E}_t[\log P(\triangle|x)] \geq \mathrm{E}_t[\log P(\triangle|x, M_k)], \quad k = 1, \ldots, K; \quad k \neq t. \tag{2.17}$$

These three measures of performance mean that one should measure the performance of BMA using an assumed true model. Therefore, there is no evidence that BMA outperforms any single competing model.

2.3.3 A fully Bayesian model averaging approach

2.3.3.1 Bayesian decision theory

There are three fundamental factors in Bayesian decision theory:

1. A distribution family of the observation, $f(x|\triangle)$,
2. a prior distribution for the parameter, $\pi(\triangle)$,
3. a loss function associated to a decision δ, $L(\triangle, \delta)$.

Using (1), (2) and from the Bayes rule, the posterior distribution of $\triangle$ is given by

$$\pi(\triangle|x) = \frac{f(x|\triangle)\pi(\triangle)}{\int_\Gamma f(x|\triangle)\pi(\triangle)d\triangle}. \tag{2.18}$$

Using the posterior distribution and (3) gives the optimal decision rule (Bayes rule) and the variance or risk. As long as the posterior distribution of $\triangle$ is available, one can perform Bayesian inference.

2.3.3.2 The Bayesian nature of BMA

In the BMA approach, one starts with the posterior given in Equation (2.10), given by the total law of probability. Subsequently, using a loss function, one computes the estimate and the associated variance. The question of what prior and statistical model are associated to this estimate remains. The priors and statistical model are only implicitly included in BMA estimates, through each competing model. For BMA method to be fully Bayesian method, one needs to know the prior $P(\triangle)$ and the statistical model $P(x|\triangle)$ from which the posterior $P(\triangle|x)$ is derived. This explains why Bayesian model averaging can't be considered as fully Bayesian approach, unless the associated prior and statistical model is known. The drawback of this quasi-Bayesian method is that, it is hard to know the long run performance of the resulting estimator (e.g. minimaxity, admissibility). The reason is that, the knowledge of frequentist performance of Bayesian rules involves that of the prior distribution and statistical model. The frequentist performance (e.g., average risk) of BMA can be evaluated by assuming a true statistical model $f_t(x|\triangle)$ that generated the data, without knowing the process that generated BMA estimates. Some frequentist performances of BMA estimator are given in Hjort and Claeskens (2003), followed with a discussion by Raftery and Zheng (2003).

2.3.3.3 Description of a fully BMA

The prior of the quantity of interest can be defined as

$$P_{\text{fbma}}(\triangle) = \Sigma_{k=1}^{K} P(\triangle|M_k)P(M_k). \tag{2.19}$$

The parametric statistical model $P_{\text{fbma}}(x|\triangle)$ can also be defined as

$$P_{\text{fbma}}(x|\triangle) = \Sigma_{k=1}^{K} P(x|\triangle, M_k)P(M_k). \tag{2.20}$$

$P(x|\triangle, M_k)$ is the parametric statistical model for model M_k. The use of Bayes rule leads to the posterior of the quantity of interest $P_{\text{fbma}}(\triangle|x)$ as

$$P_{\text{fbma}}(\triangle|x) = \frac{P_{\text{fbma}}(x|\triangle)P_{\text{fbma}}(\triangle)}{\int_\Gamma P_{\text{fbma}}(x|\triangle)P_{\text{fbma}}(\triangle)d\triangle}. \tag{2.21}$$

Defining a loss function, Bayesian estimates are then obtained.
This approach may be difficult to implement, but the long and short run properties are then known. That is, one can find conditions under which there are consistent, minimax and admissible. All the frequentist properties (minimaxity, admissibility, etc.) of Bayes rules now apply. We refer to this approach as *fully Bayesian model averaging* (FBMA).

Proposition 2.3.1 *Under (2.19) and (2.20), assuming that for all* k *and* j, $k \neq j$, $h_{kj}(x) = \int_\Gamma P(x|\triangle, M_k)P(\triangle|M_j)d\triangle < \infty$, *the posterior of the quantity of interest in (2.21) is given by* $P_{\text{fbma}}(\triangle|x) =$

$$\frac{\Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P^2(M_k) + \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(x|\triangle, M_k)P(\triangle|M_j)P(M_k)P(M_j)}{\Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) + \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} h_{kj}(x)P(M_k)P(M_j)}. \tag{2.22}$$

Proof. $P_{\text{fbma}}(x|\triangle)P_{\text{fbma}}(\triangle) = \{\Sigma_{k=1}^{K} P(x|\triangle, M_k)P(M_k)\}\{\Sigma_{k=1}^{K} P(\triangle|M_k)P(M_k)\}$

$= \Sigma_{k=1}^{K} P(x|\triangle, M_k)P(M_k)P(\triangle|M_k)P(M_k)$
$+ \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(x|\triangle, M_k)P(M_k)P(\triangle|M_j)P(M_j)$

$= \Sigma_{k=1}^{K} P^2(M_k)P(\triangle|M_k)P(x|\triangle, M_k)+\Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(M_k)P(M_j)P(x|\triangle, M_k)P(\triangle|M_j).$

Since, $P(\triangle|M_k)P(x|\triangle, M_k) = P(x|M_k)P(\triangle|x, M_k)$ by Bayes rule,
$P_{\text{fbma}}(x|\triangle)P_{\text{fbma}}(\triangle) = \Sigma_{k=1}^{K} P^2(M_k)P(x|M_k)P(\triangle|x, M_k)$

$+ \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(M_k)P(M_j)P(x|\triangle, M_k)P(\triangle|M_j)$. (1)

$P_{\text{fbma}}(x) = \int_\Gamma P_{\text{fbma}}(x|\triangle)P_{\text{fbma}}(\triangle)d\,\triangle$

$= \Sigma_{k=1}^{K} P^2(M_k) \int_\Gamma P(\triangle|M_k)P(x|\triangle, M_k)d\,\triangle$

$+ \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(M_k)P(M_j) \int_\Gamma P(x|\triangle, M_k)P(\triangle|M_j)d\triangle$

$= \Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) + \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} h_{kj}(x)P(M_k)P(M_j)$. (2)
Dividing (1) by (2) yields the result.

The use of direct BMA yields

$$P(\triangle|x) = \Sigma_{k=1}^{K} P(\triangle|x, M_k)P(M_k|x), \tag{2.23}$$

where

$$P(M_k|x) = \frac{P(x|M_k)P(M_k)}{\Sigma_{j=1}^{K} P(x|M_j)P(M_j)}. \tag{2.24}$$

This means that in general, BMA and FBMA are different. It will be hard to find the prior and statistical model associated to BMA.

Corollary 2.3.1 *Suppose that all the models have the same parametric statistical model, that is $P(x|\triangle, M_k) = P(x|\triangle, M_j)$ for all k and j, then FBMA reduces to BMA.*

Proof. In the numerator of (2.22), $\Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(x|\triangle, M_k)P(\triangle|M_j)P(M_k)P(M_j)$

$= \Sigma_{k=1}^{K} P(\triangle|M_k)P(x|\triangle, M_k)P(M_k)\Sigma_{j=1;j\neq k}^{K} P(M_j)$

$= \Sigma_{k=1}^{K} P(\triangle|M_k)P(x|\triangle, M_k)P(M_k)(1 - P(M_k))$, $\Sigma_{j=1;j\neq k}^{K} P(M_j) = 1 - P(M_k)$,

$= \Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P(M_k)(1 - P(M_k)).$

The numerator of (2.22) is therefore
$= \Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P^2(M_k)+\Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P(M_k)(1-P(M_k))$

$= \Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P^2(M_k) - \Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P^2(M_k)$

$+ \Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P(M_k)$

$= \Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x, M_k)P(M_k)$.

$h_{kj}(x) = \int_\Gamma P(x|\triangle, M_k)P(\triangle|M_j)d\triangle = \int_\Gamma P(x|\triangle, M_k)P(\triangle|M_k)d\triangle = P(x|M_k)$.
Therefore, the denominator of (2.22),
$P_{\text{fbma}}(x) = \Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) + \Sigma_{k=1}^{K}\Sigma_{j=1;j\neq k}^{K} P(x|M_k)P(M_k)P(M_j)$

$= \Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) + \Sigma_{k=1}^{K} P(x|M_k)P(M_k)\Sigma_{j=1;j\neq k}^{K} P(M_j)$

$= \Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) + \Sigma_{k=1}^{K} P(x|M_k)P(M_k)(1 - P(M_k))$
$= \Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) - \Sigma_{k=1}^{K} P(x|M_k)P^2(M_k) + \Sigma_{k=1}^{K} P(x|M_k)P(M_k)$

$= \Sigma_{k=1}^{K} P(x|M_k)P(M_k)$, a mixture of marginal distributions.
Therefore $P_{\text{fbma}}(\triangle|x) = \frac{\Sigma_{k=1}^{K} P(x|M_k)P(\triangle|x,M_k)P(M_k)}{\Sigma_{j=1}^{K} P(x|M_j)P(M_j)}$

$= \Sigma_{k=1}^{K} \{\frac{P(x|M_k)P(M_k)}{\Sigma_{j=1}^{K} P(x|M_j)P(M_j)}\} P(\triangle|x, M_k)$

$= \Sigma_{k=1}^{K} P(\triangle|x, M_k)P(M_k|x) = P(\triangle|x)$.
This means that in this special case, BMA is a fully Bayesian.
In this special case, the posterior mean and variance of the weighted estimate $\hat{\triangle}_{\text{fbma}}$ are those of BMA and given by

$$\begin{aligned} \hat{\triangle}_{\text{fbma}} &= \text{E}_{\text{fbma}}(\triangle|x) = \Sigma_{k=1}^{K} \text{E}(\triangle|x, M_k)P_{\text{fbma}}(M_k|x), \\ \text{Var}_{\text{fbma}}(\triangle|x) &= \Sigma_{k=1}^{K} [\text{Var}(\triangle|x, M_k) + (\text{E}(\triangle|x, M_k) - \hat{\triangle}_{\text{fbma}})^2] P_{\text{fbma}}(M_k|x). \end{aligned} \tag{2.25}$$

In general, as Bayes estimate, the form of the posterior mean and variance for FBMA are not given in advance. We can't expect this form to be simple as (2.25) for complex situations. It is important to note that model averaging in the framework of FBMA is with respect to mixture of both the priors and the parametric models. For BMA, model averaging is only with respect to mixture of posterior distribution. FBMA estimates may be computationally demanding (but feasible) since the posterior $P_{\text{fbma}}(\triangle|x)$ involves many sums, specially if the number K of models is large.

2.3.3.4 Illustration of a fully BMA method

When a coin is spun on its edge, instead of being thrown in the air, the long-run frequencies of heads is rarely 0.5. This means that one cannot model the process by uniform prior (that is beta (1,1)). A mixture of priors may be appropriate. Consider $X_1, \ldots, X_n$ be n independent Bernouilli trials, that is $X_i \sim$ Bernouilli($\triangle$), $Y = \sum_{i=1}^{n} X_i$ is the number of success. Y is a binomial(n, $\triangle$), $\triangle$ unknown. We will base inference on Y, since the likelihood function of the X_i's is $\triangle^Y (1-\triangle)^{n-Y}$ and involves the sufficient statistic Y. $f(y|\triangle) = \binom{n}{y} \triangle^y (1-\triangle)^{n-y}$, $y = 0, 1, \ldots, n$, is the probability mass function (PMF) of Y. Our quantity of interest is the unknown $\triangle$. Note that similar illustrations can be made with other examples of exponential families and conjugate prior distributions such as (normal, normal), (Poisson, gamma), (gamma, gamma), (negative binomial, beta), (multinomial, Dirichlet), (normal, gamma) etc.

$Y|\triangle \sim$ binomial$(n, \triangle)$, $\triangle \sim$ beta(α, β), then $\triangle|y \sim$ beta$(y+\alpha, n-y+\beta)$, therefore

$$\hat{\triangle} = \mathrm{E}(\triangle|y) = \tfrac{y+\alpha}{\alpha+\beta+n} \quad \text{and} \quad \mathrm{Var}(\triangle|y) = \tfrac{(y+\alpha)(n-y+\beta)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \tag{2.26}$$

are the Bayes estimates of $\triangle$ and its variance respectively.

The marginal distribution of Y is the beta-binomial(n, α, β), whose PDF is given by

$$f(y) = P(y|M_k) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)} = \binom{n}{y} \frac{B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)},$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(c) = \int_0^\infty t^{c-1}\exp(-t)dt$.

We consider 3 models, $n = 10$, $y = 3$. Table (2.1) shows the priors and the posteriors of the quantity of interest for three models for FBMA.

Figures (2.1), (2.2) and (2.3) compare the posterior distribution of the FBMA method. They are compared to each of the models viewed as true model. The prior for each model and the mixed prior for FBMA are also compared.

Model	$P(M)$	$P_{\text{fbma}}(M\|x)$	$P(\triangle\|M)$	$P(\triangle\|M, x)$	$\mathrm{E}(\triangle\|x, M)$	$\mathrm{Var}(\triangle\|x, M)$
M_1	0.5	0.774	beta(10,20)	beta(13,27)	0.325	0.005
M_2	0.2	0.169	beta(15,15)	beta(18,22)	0.450	0.006
M_3	0.3	0.057	beta(20,10)	beta(23,14)	0.575	0.006
FBMA	-	-	-	-	0.3625	0.0104

Table 2.1: Priors, posteriors, weights, mean and variance of each model.

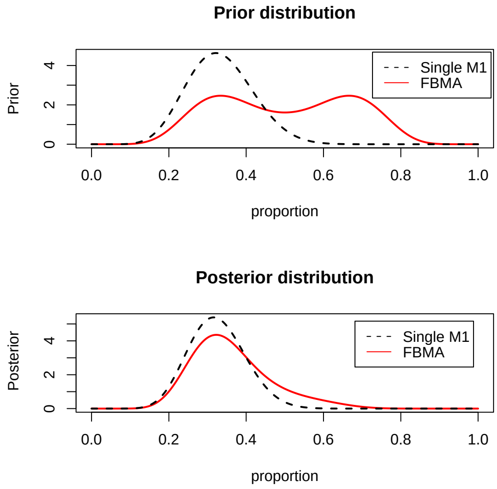

Figure 2.1: Prior and posterior distribution compared to model M_1 as a function of proportion.

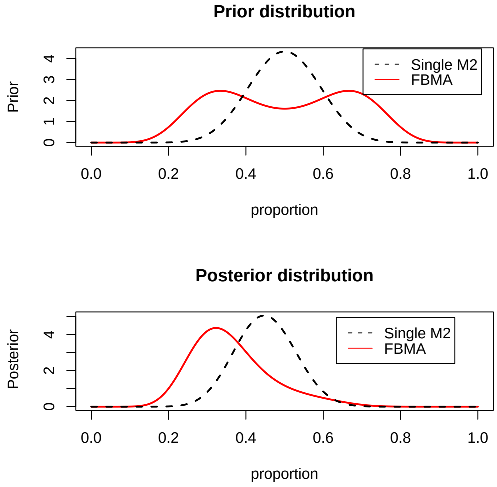

Figure 2.2: Prior and posterior distribution compared to model M_2 as a function of proportion.

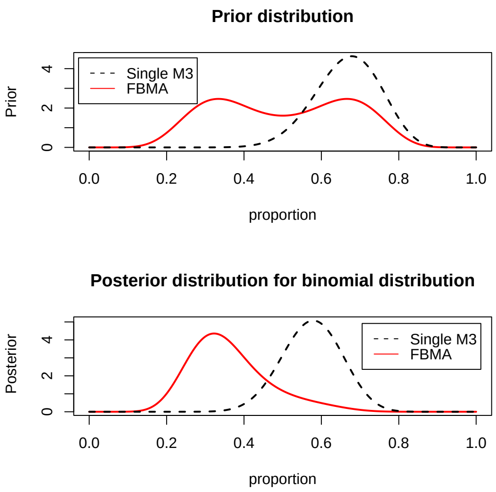

Figure 2.3: Prior and posterior distribution compared to model M_3 as a function of proportion.

2.3.4 Model search strategies

The implementation of BMA is demanding when the number of models considered is large. To ease the computational burden some authors have suggested methods for reducing the number of models that are included in the BMA. Three approaches are outlined. We then point out that, for the assessment of frequentist properties of BMA estimators, it is necessary to take account of such preselection. The general issue of post-model-selection estimation is discussed in Chapter 3.

2.3.4.1 Occam's window method

The method is described in Madigan and Rafterey (1994) and the two basic principles are:

1. If a model predicts the data much worse than the best model, then it should be dropped,

2. models that predict the data less well than their nested submodels should be dropped.

Based on (1) and (2), the set of considered models is reduced. Now, the point is to find class of models to be averaged. At each step, the method compares two models, and rejects one of them. The process is repeated until a set of acceptable models is found.

2.3.4.2 Markov chain Monte Carlo (MCMC) model decomposition

The method is described in (Madigan and York, 1995). A Markov chain is built on the model space with stationary distribution $P(M_k|x)$. If the chain is actually at state M_j, then a neighborhood model M_i should be accepted with probability that is the minimum between 1 and their respective Bayes factor. Otherwise, the chain will stay on M_j.

2.3.4.3 Other methods

The stochastic search variable selection (George and McCulloch, 1993) involves not removing a predictor from the full model, but are set to be nearly 0 with high probability. Then a Markov chain moves through all the models and parameters. Volinsky, Madigan, Raftery and Kronmal (1997) use the "leaps and bound" algorithm of Furnival and Wilson (1974) to identify models to be averaged. Clyde, DeSimone and Parmigiani (1996) use importance sampling methods to identify best subset of models.

2.3.4.4 Model space selection uncertainty in BMA

Suppose that using a model search strategy S, one obtains a subset $\hat{\mathcal{M}}$ of $\mathcal{M}$ with dimemsion $\hat{K}$. The weighted estimator is now given by

$$\tilde{\triangle}_{\text{bma}}(S) = \mathrm{E}(\triangle|x, \hat{\mathcal{M}}) = \Sigma_{k=1}^{\hat{K}} \mathrm{E}(\triangle_k|x, M_k) P(M_k|x). \tag{2.27}$$

Posterior variance and posterior risk of the estimates are computed as if $\hat{\mathcal{M}}$ was not data dependent, like those of $\hat{\triangle}_{\text{bma}}$. There is no problem since the data are held fixed. However, suppose that one wants to measure the performance of the averaged estimator using the averaged risk (frequentist risk) defined by

$$\mathrm{R}(\triangle, \tilde{\triangle}_{\text{bma}}(S)) = \int_{\mathcal{X}} L(\triangle, \tilde{\triangle}_{\text{bma}}(S)) P(x|\triangle, M_t)\, dx. \tag{2.28}$$

Expression (2.28) is now difficult to evaluate as it includes the model search strategy S. The fact of reducing the subset of models by model search strategy S using data introduces an additional source of uncertainty. We refer to this as *model space selection uncertainty*. The difference with model selection uncertainty is that here a subset is selected instead of one model. This fact should be taken into account in computing the frequentist risk of the averaged estimator after a model search strategy has been performed. Otherwise variance and these risks are likely to be underestimated.

2.4 Frequentist model averaging

2.4.1 Akaike weights and likelihood weights

Suppose that each model M_k is parametric with probability density function $f_k(x|\theta_k)$ with the quantity of interest $\triangle_k$ as a function of θ_k (e.g. $\triangle_k = g_k(\theta_k)$). Let $\hat{\theta}_k$ be the maximum likelihood estimator of θ_k for model M_k, L_k the likelihood and q_k the number of parameters for model M_k.

Because the optimal weights are not feasible, in general, the idea is to penalise the likelihood. Different penalties are possible. Buckland et al. (1997) define Akaike weights:

$$W_k = \frac{\exp(-s_k/2)L_k}{\Sigma_{i=1}^K \exp(-s_i/2)L_i} = \frac{\exp(-\frac{I_k}{2})}{\Sigma_{i=1}^K \exp(-\frac{I_i}{2})}, \tag{2.29}$$

where I_k is an information criterion of the form

$$I_k = -2\log L_k + s_k, \tag{2.30}$$

with s_k a penalty for model M_k. In particular, if the Akaike information criterion (AIC, $s_k = 2q_k$) is used, (2.29) becomes

$$W_{a_k} = \frac{\exp(-\frac{AIC_k}{2})}{\Sigma_{i=1}^K \exp(-\frac{AIC_i}{2})}. \tag{2.31}$$

This is a penalised likelihood where each model is penalised by the number of parameters. Extensive application of Akaike weights can be seen in Burnham and Anderson (2002). Hjort and Claeskens (2003) define similar weights with a smooth FIC (focused information criterion) and other model averaging schemes for estimators, known as *compromise estimators* together with their limiting distributions and risk properties. One way of interpreting Akaike weights comes from Akaike's (1978) predictive approach who defines the likelihood of a model to be asymptotically equivalent to $\exp(-\frac{AIC}{2})$ and uses for weights time series autoregressive models.

In the context of regression and classification, Leblanc and Tibshirani (1996) propose a simple likelihood given by

$$W_k = \frac{L_k}{\Sigma_{i=1}^K L_i}. \tag{2.32}$$

The expectation of W_k in (2.32) followed by bootstrap was proposed by Strimmer and Rambaut (2001) for gene trees analysis.

2.4.2 Likelihood of a model and Akaike weights interpretation

Viewed as a predictive approach, the likelihood of a model can be described as follows.
Consider the data set x as given and suppose that the purpose is to find the distribution of future observations x_+ given x. Let $f(x_+|x)$ and $f(x_+)$ be respectively the predictive density and the true density of x_+. We can think of $f(x_+|x)$ as an estimate of $f(x_+)$. Akaike (1978) defines the likelihood of a model to be $f(x_+|x)$. The goodness of this estimation can be measured by the entropy of $f(x_+)$ with respect to $f(x_+|x)$ by

$$K(f(.), f(.|x_+)) = -\int f(x_+) \log \frac{f(x_+)}{f(x_+|x)}\, dx_+. \tag{2.33}$$

Equation (2.33) can be written as

$$K(f(.), f(.|x_+)) = \int f(x_+) \log f(x_+|x)\, dx_+ - \int f(x_+) \log f(x_+)\, dx_+. \tag{2.34}$$

The first term on the right-hand side of Equation (2.34) is the expectation with respect to the true distribution of $\log f(x_+|x)$. It is not possible to evaluate Equation (2.34) since the true model is unknown. Suppose that we have a parametric family vector θ. If $f(.|x) = f(.|\theta)$, then $\log f(x|x) = \log f(x|\theta)$, but in general this does not hold. Akaike (1978) proposes to define the log likelihood of a model $f(.|x)$ by

$$l(f(.|x)) = \log f(x|x) + c, \tag{2.35}$$

where c is a constant for all possible $f(x)$. If attention is restricted to parametric family, we can write $f(x_+|x) = f(x_+|\hat{\theta}(x))$ and $\hat{\theta}(x)$ is the maximum likelihood estimate of θ. Akaike (1978) proves that asymptotically

$$l(f(.|\hat{\theta}(x))) = \log f(x|\hat{\theta}(x)) - p, \tag{2.36}$$

where p is the number of parameters, and then suggests $\exp(-\frac{AIC}{2})$ to be asymptotically a reasonable definition for the likelihood of a parametric model, for prediction purposes.

2.4.3 Illustrative example

A method for estimating design storms is described in Zucchini and Adamson (1984), Linhart and Zucchini (1986, p. 66) and is described as follows. The annual maximum storms are assumed to be independently and identically distributed with distribution function G. The distribution of the largest storm in h years is G^h. The storm associated with a design horizon of h years and probability (risk) of occurence ω is the solution of the equation

$$1-\omega = G^h(s).$$

Let F_θ be an approximating model for G, where θ is the parameter (possibly vector). Then the design storm, s, estimated by fitting G_θ is given by

$$\hat{s} = \hat{\triangle} = F^{-1}_{\hat{\theta}}((1-\omega)^{1/h}).$$

To estimate $\triangle$, the following six models were considered: gamma (α, β), normal (μ, σ^2), lognormal (μ, σ^2), exponential (ν), Weibull (ϱ, τ) and Gumbel (ζ, η). Data are the annual maximum 1 day storm depths (mm) at Vryheid for the period 1951-1980. The data are given in Table (2.2). Figure (2.5) displays the histogram and kernel density estimator.

Maximum likelihood estimates of the parameters, together with Akaike weights are given in table (2.3). The Gumbel distribution is the one with highest Akaike weight.

Estimates of $\triangle$, the design storms and the associated standard error and mean are given in Table (2.4), for each design horizon. The risk ω was chosen to be 0.2.

In this example, note that when the estimators for each competing model are assumed to be perfectly correlated, model averaging using Akaike weight has higher standard error than that of each competing model, except for the horizon 10 with exponential distribution. When they are assumed to be independent, except for horizon 1, Akaike weights model averaging has smaller standard error than all competing models. The variance formulae used are that of Buckland et al. (1997). This illustrates that there is an enormous discrepancy in the

Year	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Depth	45.2	66.5	142.0	83.9	61.1	60.6	84.5	80.0	79.0	137.5
Year	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Depth	52.5	50.0	170.0	62.0	43.5	60.0	60.0	53.5	58.0	93.0
Year	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Depth	84.5	74.5	94.0	80.0	74.0	64.0	60.0	51.5	58.5	88.0

Table 2.2: Annual maximum 1 day storm depths (mm) at Vryheid for the period 1951-1980.

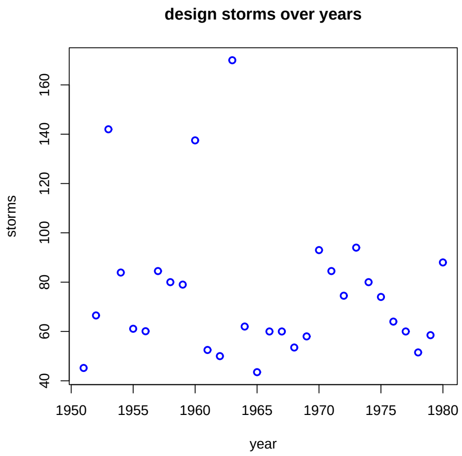

Figure 2.4: Plot of storms over year.

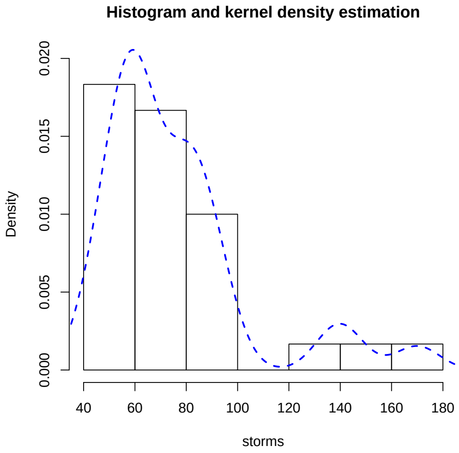

Figure 2.5: Histogram and kernel density estimation of storms.

estimates of standard errors of the model averaging estimator, depending of what is assumed. In general, this is a major issue when attempting to determine the properties of model averaging estimators (and post-model-selection estimators). The properties can only be determined under a specific assumption, e.g. that a given model generated the observations. Limiting distribution and risk properties of model averaging and post-model-selection estimators are given in Hjort and Claeskens (2003).

Estimate-weight	gamma	normal	lognormal	exponential	Weibull	Gumbel
$\hat{\lambda}_1$	8.98	75.71	4.27	0.013	2.68	64.08
$\hat{\lambda}_2$	0.12	28.65	0.32		85.08	17.86
W_{a_k}	0.059	0.000	0.333	0.000	0.001	**0.607**

Table 2.3: Maximum likelihood estimates of parameters and Akaike weights.

Distribution	Caracteristic	$h = 1$	$h = 5$	$h = 10$
gamma	Estimate	95.75	123.75	134.62
	Standard error	8.61	14.83	17.42
	Mean	94.92	121.95	132.45
normal	Estimate	99.82	124.70	133.38
	Standard error	9.51	14.42	16.16
	Mean	98.56	122.32	130.61
lognormal	Estimate	93.73	123.98	136.56
	Standard error	8.11	15.28	18.68
	Mean	93.06	122.31	134.60
exponential	Estimate	121.86	237.10	288.74
	Standard error	8.23	16.02	19.51
	Mean	121.56	236.52	288.04
Weibull	Estimate	101.60	130.23	140.17
	Standard error	9.17	15.12	17.39
	Mean	100.28	127.36	136.72
Gumbel	Estimate	90.87	119.62	132.00
	Standard error	7.72	13.12	15.49
	Mean	90.46	118.60	130.72
Model averaging	Estimate	92.13	121.30	133.68
	Standard error (perfect correlation)	39.52	19.30	18.81
	Standard error (independance)	28.91	12.89	11.86

Table 2.4: Estimated design storms, their standard errors and Akaike model averaging.

Chapter 3

Model Selection Uncertainty

3.1 Introduction

In most statistical modelling applications there are several models that are apriori plausible. It is then usual for the analyst to apply some model selection procedure, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), to select a single model. Although they are not generally interpreted as such, tests of statistical hypothesis also constitute selection procedures. For example in a regression analysis, the result of a test of the hypothesis that a certain regression coefficient is equal to zero is often used to decide whether the corresponding independent variable should be included or excluded from the regression. If so, one is applying hypothesis testing to decide between two regression models; one that includes the variable under consideration, and another without that variable. Similarly "iterative model building" is a systematic model selection procedure. In this thesis the term model selection procedure includes all formal or heuristic criteria (e.g. exploratory data analysis) or methodology that *use the data* to select one out of a set of competing models.

The fact that the selection was data-based is often ignored in the subsequent analysis. One proceeds as though one had chosen the model apriori, i.e. without reference to the data. In other words the model is regarded as having been known in advance, whereas in fact it was not. This leads to invalid inference due to *model selection uncertainty*. The important distinction here is between valid inference after model selection and a naive approach that does not take into account the selection step. The latter proceeds as though the selected model had been chosen in advance. We show that leads to a variety of problems. Since most statistical data analysis involve exploratory data analysis, the model selection uncertainty problem extends to any such analysis.

The selection procedure partitions the sample space into disjoints subsets that determine which model is selected. The selection of a particular model implies that the sample belongs only to that subset of the sample space which leds to its having been selected. In theory it is possible to apply a conditional analysis (e.g. Miller (2002)), but this is not easy. Secondly one would generally prefer to use unconditional analyses.

We will examine model selection uncertainty from the point of view of decision theory. We define a probability framework that clarifies the difficulties of deriving the properties of post-model-selection estimators (PMSEs), as well as the fact that a conditional analysis is incomplete; the natural approach in this framework is the unconditional. It also allows us to appreciate the difficulty of computing the distribution of post-model-selection estimators, the coverage probability of confidence intervals, the p-value of test statistics, and goodness of fits test after model selection.

Using a simple example of linear regression, we illustrate that no single post-model-selection estimator dominates all others in terms of risk (Nguefack and Zucchini, 2005). Thus it is not even possible to recommend any single criterion (e.g. AIC (Akaike, 1973), BIC (Schwarz, 1978), Cp (Mallows, 1973), Hypothesis testing, HQ (Hannan and Quin, 1979), etc.) for model selection that is to be followed by inference based on the selected model. In the case of hypothesis testing, there is no optimal level of significance level in the sense of having the smallest risk. This is not surprising when one notes that, in this framework, hypothesis testing, like the AIC, is a penalized likelihood criterion in which the penalty term is a function of the significance level. The example also illustrates the potential multimodal nature of post-model-selection estimators, their bias, variance and risk. We identify the model selection probabilities as the key quantities in this context. We also illustrate the behaviour of the post-model-selection estimators as the sample size increases.

Perhaps surprisingly, the use of *consistent* model selection procedures does not solve the problem. Thus, if the true model is one of the competing models, and if one uses a consistent criterion which will asymptotically choose this true model, the problem is still not solved. In this case the distribution of the PMSE does converge to the distribution of the true model, but convergence is only pointwise. This can be seen using an asymptotic efficiency approach which shows that the normalized risk continues to grow with increasing sample size. Thus even under such ideal assumptions the problem of model selection uncertainty remains.

3.2 Decision theory approach

Let $\mathcal{A}$ be a set of actions from which a statistician will take one action or a group of actions after observing a "fact". We refer to the fact as data, denoted by x. As long as this fact is not observed, the data are random and are denoted by X. The elements of $\mathcal{A}$ can be random or nonstochastic. An action is random if it depends on the X. Whether actions are random or not, the action $a(X)$ to be taken is a function of the data, $a(X) \notin \mathcal{A}$. The properties of $a(X)$ will therefore be different from that of the elements of $\mathcal{A}$. After the data have been observed, the decision $a(x)$ now belongs to the action space. One has to find properties of this estimate. For example, let sample action $\mathcal{A} = \{a_1, a_2\}$, only 2 actions. Suppose that action a_1 will be taken if it rains and action a_2 will be taken otherwise. Here $\mathcal{X} = \{\text{rain,sun}\}$ is the sample space, the random action $a(X) \notin \mathcal{A}$. Now suppose it rains, observed data=x=rain, then $a(x) = a_1 \in \mathcal{A}$. The properties of $a(x)$ are not that of action a_1. This means that as long as you do not have realisation of the random process X, the decision is not known and is not in the sample action $\mathcal{A}$. One can only say that this decision is a mixture of elements of $\mathcal{A}$. Suppose that the sample action is itself random, $\mathcal{A}(X) = \{a_1(X), a_2(X)\}$. The decision $a(X) \neq a_1(X)$ and $a(X) \neq a_2(X)$, in fact two random variables Z and Y are equal if $P(Z = Y) = 1$.

For example, suppose that a lecturer wants to report student overall performance to the university administration and he decides to report the mean if the mean is greater than the median and the median otherwise. As long as he did not have the grade, the decision is neither the mean nor the median. In general, the random decision is not in $\mathcal{A}$ because $P(a(X) = a) \neq 1$, $a \in \mathcal{A}$.

The order statistic of the random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order and they are denoted by $X_{(1)}, \ldots, X_{(n)}$. The order statistics are then random variables that satisfies $X_{(1)} \leq \ldots \leq X_{(n)}$. That is $X_{(r)}$ is the r^{th} smallest X_i. Suppose that $X_1, \ldots, X_n$ come from a continuous population with cdf $F(x)$ and pdf $f(x)$. A well known result in statistics of the pdf of $X_{(r)}$ is given by

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!} f_X(x)[F_X(x)]^{r-1}[1 - F_X(x)]^{n-r}. \tag{3.1}$$

From Equation (3.1), one can see that although for a realisation of the random sample, any $x_{(r)}$ is an element of the realised sample, as random variable, its density is completely different of the common density $f_X(x)$. Even asymptotically, the 2 distributions are not similar. For example, from extreme value theory, it is well known that, under some conditions, the asymptotic distribution of maxima is extreme value distribution (Frechet, Weibull or Gumbel distributions). In this example, we consider a selection procedure being that of finding a particular order statistics.

Example. Let $X_1, \dots, X_n$ be iid uniform(0,1), then $f(x) = 1$ and $F(x) = x$ for $x \in (0, 1)$, E(X)=1/2 and Var(X)=1/12.
The PDF of the r^{th} order statistic is

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r} = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)} x^{r-1}(1-x)^{(n-r+1)-1}.$$

This means that $X_{(r)} \sim \text{beta}(r, n-r+1)$ distribution completely different from uniform(0,1)=beta(1,1), $\text{E}X_{(r)} = \frac{r}{n+1}$ and $\text{Var}X_{(r)} = \frac{r(n-r+1)}{(n+1)^2(n+2)}$.

3.3 Problem, concepts and definitions

Let $x = (x_1, \dots, x_n)$ be n realisations (data) of the random variables $X = (X_1, \dots, X_n)$, and let M_t be the unknown true model that generated this process. Suppose that an approximating (parametric) model M_a is assumed, that is $X_i \sim f_a(x_i|\theta)$. Let $\triangle$ be the quantity of interest ($\triangle$ is a function of the parameter θ, $\triangle = \text{h}(\theta)$). *Model uncertainty* refers to the fact that the true model M_t is not known. Let $\hat{\triangle}$ be the estimator of $\triangle$ based on M_a. The bias of $\hat{\triangle}$ is given by $\text{Bias}_\theta(\hat{\triangle}) = \text{E}_\theta(\hat{\triangle}) - \triangle$.

Now, suppose that one selects from a set of K models $\mathcal{M} = (M_1, \dots, M_K)$ and let S be a selection procedure. Let $\tilde{M}(X|S, \mathcal{M})$ be the selected model and $\tilde{\triangle}(X|S, \mathcal{M})$ the corresponding estimator of $\triangle$.
These estimators are defined by

$$\tilde{M}(X|S, \mathcal{M}) = \sum_{k=1}^{K} I_k(X|S, \mathcal{M}) M_k, \tag{3.2}$$

$$\tilde{\triangle}(X|S, \mathcal{M}) = \sum_{k=1}^{K} I_k(X|S, \mathcal{M}) \hat{\triangle}_k, \tag{3.3}$$

where $I_k(X|S, \mathcal{M}) = 1$ if M_k is selected by S and 0 otherwise, $\hat{\triangle}_k$ is the estimator of $\triangle$ under model M_k. The estimator $\tilde{M}(X|S, \mathcal{M})$ depends on the data and the selection procedure through the random quantity $I_k(X|S, \mathcal{M})$, therefore is now random. We want to stress the dependence of selected model and PMSEs on S and $\mathcal{M}$. The random quantity $I_k(X|S, \mathcal{M})$ is a 0-1 weight, so that $\tilde{\triangle}(X|S, \mathcal{M})$ is a weighted estimator. We refer to $\tilde{\triangle}(X|S, \mathcal{M})$ as the *post-model-selection* *estimator* (PMSE). As random quantities, $\tilde{M}(X|S,\mathcal{M}) \notin \mathcal{M}$ and $\tilde{\triangle}(X|S,\mathcal{M}) \notin \{\hat{\triangle}_1,\ldots,\hat{\triangle}_K\}$, therefore their properties are different of those of any competing model in the set $\mathcal{M}$. The reason is that there is no evidence that

$$P_\theta(\tilde{M}(X|S,\mathcal{M}) = M_k) = P_\theta(\tilde{\triangle}(X|S,\mathcal{M}) = \hat{\triangle}_k) = 1, \quad \text{for} \quad k = 1,\ldots,K.$$

These probabilities depend on the true parameter θ.

Each estimator $\hat{\triangle}_k$ is derived from the model M_k. The model from which the PMSE is derived is not known. It is important to note that for computing the properties of PMSE (e.g. distribution, mean, variance, MSE), one needs to assume a true model M_t. Therefore, these properties depend on the true parameter θ. There is still uncertainty about the choice of this model. It is precisely that uncertainty that led to perform model selection. This means that the problem involved in model selection is not only that of obtaining the properties, but also the choice of an assumed true model from which to get these properties.

After the data have been observed, $\tilde{M}(x|S,\mathcal{M}) \in \mathcal{M}$ and $\tilde{\triangle}(x|S,\mathcal{M}) \in \{\hat{\triangle}_1,\ldots,\hat{\triangle}_K\}$. The point is to study estimators from Equation(3.3), not those of the realisation $X = x$, corresponding to a fixed model. For example, even if the $\hat{\triangle}_k$'s are unbiased for $\triangle$, this does not guarantee unbiasness of $\tilde{\triangle}(X|S,\mathcal{M})$. If one does not take into account the selected procedure, we will call the subsequent inference the *naive inference*, the selected model is then called the *naive model* denoted M_{k^*}. Let $\hat{\triangle}_{k^*}$ be an estimator of $\triangle$ based on the naive model. We define the following quantity:

model selection difference bias=$\text{Bias}_\theta(\tilde{\triangle}(X|S,\mathcal{M}))$-$\text{Bias}_\theta(\hat{\triangle}_{k^*})$,

model selection variance difference=$\text{Var}_\theta(\tilde{\theta}(X|S,\mathcal{M}))$-$\text{Var}_\theta(\hat{\triangle}_{k^*})$,

model selection risk difference=$\text{MSE}_\theta(\tilde{\triangle}(X|S,\mathcal{M}))$-$\text{MSE}_\theta(\hat{\triangle}_{k^*})$.

Model selection density difference and *model selection distribution difference* are defined to be respectively the difference between the density and distribution of PMSE and that of the naive estimator. The sign of these quantities is not a priori known.

More generally, we refer to *model selection difference inference* as the difference inference between post-model-selection inference and the naive inference. *Model selection inference* is likely to be present because $P_\theta(\tilde{M}(X|S,\mathcal{M}) = M_{k^*})$ is likely to be less than 1, since one does not know the true parameter θ. But our interest is not of reducing the model selection inference, since we do not consider naive inference to be a valid inference. Instead, we are interested in looking for better estimators than PMSEs (improving upon PMSE) or computing exact

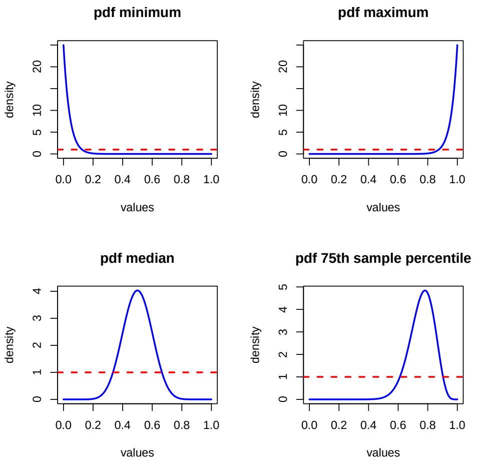

Figure 3.1: Densities comparing order statistic (solid line) with a naive distribution (dotted line).

properties PMSEs. The magnitude of model selection inference depends on the selection procedure S. Under the naive approach, all the model selection difference quantities are assumed to be 0. A valid computation of these quantities can be performed under the model $\tilde{M}(X|S,\mathcal{M})$. When a close form exists for these quantities, this will be a function of θ and one can estimate these quantities by using a plug-in the estimator of θ. To illustrate the fact that PMSE is different from any competing estimator, consider the following example of order statistic. The naive inference consists of working with one X_i, that is, making inference with uniform distribution. Model selection difference bias and variance are derived as follow

model selection bias difference=$\mathrm{E}X_{(r)} - 1/2$,
model selection variance difference=$\mathrm{Var}X_{(r)} - 1/12$.

Graphical illustrations are given for $n = 25$ in Figure (3.1) and Figure (3.2). For Figure (3.1), one can see the difference between the naive distribution (uniform) and the right distribution (beta-distribution) of order statistic, here viewed as PMSEs. Similar remarks apply for the bias, and the variance. Model selection bias difference and model selection variance difference shown in (3.2) are not equal to 0.

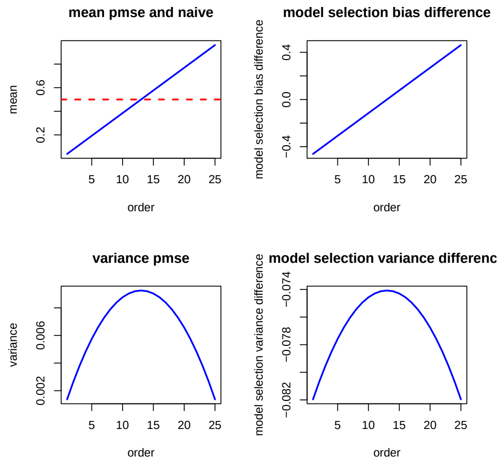

Figure 3.2: Model selection bias difference and model selection variance difference as a function of order statistic.

3.4 Graphical representation and partition

We refer to the term selection procedure as any method leading to the choice of a model. Although, for computation reasons, we restrict our attention to parsimonious selection criteria like AIC and also hypothesis testing, model selection itself is complex and can be sometimes be performed as an iterative process which is difficult to formalise. Usually, in practice, with data at hand and knowledge of statistical theory, explorative data analysis (EDA) is performed. One can select a model directly or iteratively.

3.4.1 Direct selection

A set of models is considered; one model is selected using a formal selection criterion, inference about a quantity of interest is performed.

3.4.2 Iterative approach to model building

1. Model identification: the data and any available information are used to suggest a class of parsimonious models to consider. One model is selected using a selection criterion.

2. Using the selected model, parameters or any quantity of interest are estimated, conditioning on the adequacy of the selected model.

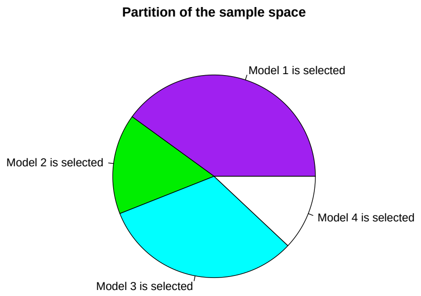

Figure 3.3: A selection procedure partitions the sample space.

3. Diagnostic checking: the fitted model is then checked to assess whether it is really adequate (for e.g. by analysing the residuals). If it is not, one returns to step (1). This process continues until a "good model" is found, then inference is then made for any quantity of interest.

From Equation (3.3), one can see that model selection partitions the sample space (for the data) $\mathcal{X}$ into disjoints subsets. Let $\mathcal{X}_k(S, \mathcal{M})$ be the subset for which model M_k is selected, namely,

$$\mathcal{X}_k(S, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : M_k \quad \text{is selected}\},$$

then

$$\mathcal{X} = \bigcup_{k=1}^{K} \mathcal{X}_k(S, \mathcal{M}), \quad \mathcal{X}_k(S, \mathcal{M}) \cap \mathcal{X}_l(S, \mathcal{M}) = \emptyset, \quad \forall k, l, \qquad k \neq l. \tag{3.4}$$

Let $\mathcal{X}_{k^*}(S, \mathcal{M})$ be the set for which the naive model M_{k^*} is selected. Model selection inference is likely to be present because $\mathcal{X}_{k^*}(S, \mathcal{M})$ is likely to be different of $\mathcal{X}$. For AIC as information criterion, the partition is

$$\mathcal{X}_k(AIC, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : AIC_k = \min_k(AIC_1, \ldots, AIC_K)\}.$$

Graphically, Figure (3.3) shows how the sample space has been partioned, for $K = 4$ models, where each model is selected on a unique subset.

3.5 Comparing classical and model selection approaches for data analysis

The aim of this section is to illustrate that in applied statistical analysis, the use of a single model after a preliminary analysis like exploratory data analysis (e.g. histogram or any useful plot, etc.) is also an inference after (informal) model selection procedure. For both formal and informal selection procedures, the model selection uncertainty problem is hard to formalise.

3.5.1 Frequentist approach to statistical analysis

The frequentist approach to statistical analysis can be sketched as follows:

1. Quantity of interest $\triangle$.

2. Data $x = (x_1, \ldots, x_n)$.

3. Use x for preliminary analysis, e.g. exploratory data analysis (histogram, any useful plots).

4. Regard the data as a realisation of a random process $X = (X_1, \ldots, X_n)$, $X \sim M_t$ (true unknown model).

5. From (3), assume $X \sim \hat{M}_\theta$ ($\triangle = h(\theta)$): there is model uncertainty, since the true model is unknown, and model selection uncertainty since $\hat{M}_\theta$ depends on the data.

6. Use an estimation method, e.g. MLE, to get $\hat{\theta}(X)$, therefore $\hat{\triangle}(X)$.

7. Find the properties of $\hat{\triangle}$, e.g., $\mathrm{E}_\theta(\hat{\triangle}(X))$, $\mathrm{Var}_\theta(\hat{\triangle}(X))$, $\mathrm{MSE}_\theta(\hat{\triangle}(X))$.

8. Use the data x to compute: $\hat{\triangle}(x)$, $\hat{\mathrm{E}}_{\hat{\theta}(x)}$, $\hat{\mathrm{Var}}_{\hat{\theta}(x)}$, $\hat{\mathrm{MSE}}_{\hat{\theta}(x)}$, confidence interval or other quantities.

From the above, the data are used in steps (3) and (8). From Step (4) to step (7), the data are viewed as random (unconditional analysis). The step (3) is viewed as an informal model selection procedure leading to a choice of the model $\hat{M}_\theta$. If step (3) is not performed, the only problem is model uncertainty since the data are used once (only in step (8)). However, in statistical data analysis, one typically does exploratory data analysis.

3.5.2 Frequentist model selection approach

1. Quantity of interest $\triangle$.

2. Data $x = (x_1, \ldots, x_n)$.

3. Use x for preliminary analysis, e.g. exploratory data analysis (histogram, any useful plots).

4. Regard the data as realisation of a random process $X = (X_1, \ldots, X_n)$, $X \sim M_t$ (true unknown model).

5. From (3), postulate $\hat{\mathcal{M}} = (M_1, \ldots, M_K)$ alternative plausible (parametric θ) models, $\triangle = h(\theta)$.

6. Use <u>any</u> model selection criteria and data x to select a model (Model Uncertainty) $\hat{M}(x) = M_{\hat{k}(x)} \in \mathcal{M}$, $\hat{k}(x) \in \{1, \ldots, K\}$ and perform inference (step 7-step 9) ignoring model selection: model selection uncertainty.

7. Use an estimation method with the selected model, e.g. MLE, to get $\hat{\theta}(X)$, therefore $\hat{\triangle}(X)$.

8. Find the properties of $\hat{\triangle}$, e.g., $\mathrm{E}_\theta(\hat{\triangle}(X))$, $\mathrm{Var}_\theta(\hat{\triangle}(X))$, $\mathrm{MSE}_\theta(\hat{\triangle}(X))$.

9. Use the data x: $\hat{\triangle}(x)$, $\hat{\mathrm{E}}_{\hat{\theta}(x)}$, $\hat{\mathrm{Var}}_{\hat{\theta}(x)}$, $\hat{\mathrm{MSE}}_{\hat{\theta}(x)}$, confidence interval or other quantities.

Here, model selection procedures are involved in step 3 (informal) and step 6 (formal, e.g. AIC). The data are used in steps (3), (6) and (9). In steps (3) and (6), only <u>random</u> data X should be used, therefore $\hat{M}(X) = M_{\hat{k}(X)} \notin \hat{\mathcal{M}}$, $\hat{k}(X) \notin \{1, \ldots, K\}$. This is difficult to perform, due to the partition of the sample space.

One can see that even classical statistical data analysis involving exploratory data analysis suffers also of the model selection uncertainty problem. This renders the problem difficult to formalise.

3.6 Illustrative examples of PMSEs

3.6.1 Simple linear regression

Consider two models
$M_0 : Y_i = \beta_0 + \epsilon_i$

$M_1 : Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad$ where $\epsilon_i \sim N(0, \sigma^2)$.
Let $\theta = (\beta_0, \beta_1)'$. The objective is to estimate $\triangle = E(Y|x_+)$ for a particular x_+.
Selection:
use some criterion, e.g., AIC, BIC, Cp, HQ, hypothesis test, etc. Suppose, e.g., that AIC(M_0) > AIC(M_1), then M_1 is chosen.
Estimation:
$\tilde{\triangle}_{\text{pretest}} = \begin{cases} \hat{\beta}_0 & \text{if } M_0 \text{ is selected} \\ \hat{\beta}_0 + \hat{\beta}_1 x & \text{if } M_1 \text{ is selected.} \end{cases}$
The estimator and the prediction interval are computed assuming that the selected model is fixed, i.e. known in advance.
Problem: the same data are used to choose the model and to make inferences.
Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n, \tag{3.5}$$

where the $\epsilon_i \sim N(0, \sigma^2)$, σ known (for simplicity); the results do not change much for unknown σ.
The OLS estimators are given by
$\hat{\beta}_1 = \frac{\Sigma_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\Sigma_{i=1}^n (x_i - \overline{x})^2}$; $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$,
$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \overline{x}}{\Sigma_{i=1}^n (x_i - \overline{x})^2}$.
Now, let x_+ be a future value of the covariate. The aim is to estimate the mean $\triangle = E(Y|x_+)$.
Consider two models: $M_0 : \triangle = \beta_0$ and $M_1 : \triangle = \beta_0 + \beta_1 x_+$.

3.6.1.1 Pre-test estimators

One method of selecting between the 2 models is by testing
$H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 \neq 0$.
This means that the model selection method here is the pre-test and is given by

$$\tilde{\triangle}_{\text{pretest}} = \begin{cases} \hat{\beta}_0 & \frac{|\hat{\beta}_1|}{v_1^{1/2}} < z_{1-\frac{\alpha}{2}} \\ \hat{\beta}_0 + \hat{\beta}_1 x_+ & \frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq z_{1-\frac{\alpha}{2}}, \end{cases} \tag{3.6}$$

where $v_1 = Var(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma_{i=1}^n (x_i - \overline{x})^2}$, $\overline{x} = \frac{1}{n} \Sigma_{i=1}^n x_i$ and $z_{1-\frac{\alpha}{2}}$ the quantile of N(0,1).
The partition of the sample space is
$\mathcal{X}_0(S, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : \frac{|\hat{\beta}_1|}{v_1^{1/2}} < z_{1-\frac{\alpha}{2}}\}$,
$\mathcal{X}_1(S, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : \frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq z_{1-\frac{\alpha}{2}}\}$.

Properties of $\tilde{\triangle}_{\text{pretest}}$.
The pre-test estimator (3.6) can be written as

$$\tilde{\triangle}_{\text{pretest}} = \hat{\beta}_0 I_0(\frac{|\hat{\beta}_1|}{v_1^{1/2}} < z_{1-\frac{\alpha}{2}}) + (\hat{\beta}_0 + \hat{\beta}_1 x_+) I_1(\frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq z_{1-\frac{\alpha}{2}}), \tag{3.7}$$

where I_0 and I_1 are, respectively, indicator functions under M_0 and M_1 with $I_0 + I_1 = 1$. It follows that

$$\tilde{\triangle}_{\text{pretest}} = \hat{\beta}_0 + \hat{\beta}_1 x_+ I_1(\frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq z_{1-\frac{\alpha}{2}}). \tag{3.8}$$

For simplicity of computations, suppose $\overline{x} = 0$, without loss of generality, since linear regression model (3.5) can be parametrised as

$$Y_i = \lambda_0 + \lambda_1(x_i - \overline{x}) + \epsilon_i, i = 1, \ldots, n, \tag{3.9}$$

where $\lambda_0 = \beta_0 + \beta_1\overline{x}$ and $\lambda_1 = \beta_1$. This means that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$, also $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed, therefore $\hat{\beta}_0$ and $\hat{\beta}_1$ are independant.
Let $Z_1 = \frac{\hat{\beta}_1 - \beta_1}{v_1^{1/2}}$, then $Z_1 \sim N(0,1)$. It follows that $\hat{\beta}_1 = v_1^{1/2}(Z_1 + b_1)$.
$\tilde{\triangle}_{\text{pretest}} = \hat{\beta}_0 + x_+ v_1^{1/2}(Z_1 + b_1) I_1(|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}}) = \hat{\beta}_0 + x_+\tilde{\beta}_1$.

$$\tilde{\triangle}_{\text{pretest}} = \hat{\beta}_0 + x_+ \tilde{\beta}_{1\text{pretest}}, \tag{3.10}$$

where $\tilde{\beta}_{1\text{pretest}}$ is the pretest estimator of β_1 given by

$$\tilde{\beta}_{1\text{pretest}} = \begin{cases} 0 & \frac{|\hat{\beta}_1|}{v_1^{1/2}} < z_{1-\frac{\alpha}{2}} \\ \hat{\beta}_1 & \frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq z_{1-\frac{\alpha}{2}}. \end{cases} \tag{3.11}$$

That is

$$\tilde{\beta}_{1\text{pretest}} = v_1^{1/2}(Z_1 + b_1) I_1(|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}}) = v_1^{1/2} A_{\text{pretest}}, \tag{3.12}$$

where $A_{\text{pretest}} = (Z_1 + b_1) I_1(|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}})$.

Proposition 3.6.1 *Under the model (3.5), the mean, bias, variance and MSE of the pre-test estimator (3.7) are given by*

$$\begin{aligned}
&\text{E}(\tilde{\triangle}_{\text{pretest}}) = \beta_0 + x_+ v_1^{1/2}[b_1(\Phi(r) + 1 - \Phi(q)) - \phi(r) + \phi(q)], \\
&\text{Bias}(\tilde{\triangle}_{\text{pretest}}) = x_+ v_1^{1/2}[b_1(\Phi(r) - \Phi(q)) - \phi(r) + \phi(q)], \\
&\text{Var}(\tilde{\triangle}_{\text{pretest}}) = \tfrac{\sigma^2}{n} + x_+^2 v_1[(b_1^2 + 1)(\Phi(r) + 1 - \Phi(q)) + 2b_1(-\phi(r) \\
&+\phi(q)) - r\phi(r) + q\phi(q) - \text{E}^2(\tilde{\triangle}_{\text{pretest}})], \\
&\text{MSE}(\tilde{\triangle}_{\text{pretest}}) = \text{Bias}^2(\tilde{\triangle}_{\text{pretest}}) + \text{Var}(\tilde{\triangle}_{\text{pretest}}),
\end{aligned} \tag{3.13}$$

where $b_1 = \frac{\beta_1}{v_1^{1/2}}$ *(standardized slope),* $r = -z_{1-\frac{\alpha}{2}} - b_1$*,* $q = z_{1-\frac{\alpha}{2}} - b_1$*,* ϕ *and* Φ *are respectively the PDF and CDF of standard normal.*

Proof. The properties of the pre-test estimator $\tilde{\triangle}$ are given by that of the pretest of the slope (3.26). This pre-test of slope determines the behaviour of $\tilde{\triangle}$. We will then first compute the moment of $\tilde{\beta}_{1\text{pretest}}$. From (3.10), the moments of $\tilde{\triangle}$ are given by

$$\begin{aligned}
&\mathrm{E}(\tilde{\triangle}_{\text{pretest}}) = \mathrm{E}(\hat{\beta}_0) + x_+\mathrm{E}(\tilde{\beta}_{1\text{pretest}}) = \beta_0 + x_+\mathrm{E}(\tilde{\beta}_{1\text{pretest}}),\\
&\mathrm{Bias}(\tilde{\triangle}_{\text{pretest}}) = \mathrm{E}(\tilde{\triangle}_{\text{pretest}}) - \beta_0 - b_1 v_1^{1/2} x_+ = x_+(\mathrm{E}(\tilde{\beta}_{1\text{pretest}}) - \beta_1),\\
&\mathrm{Var}(\tilde{\triangle}_{\text{pretest}}) = \mathrm{V}(\hat{\beta}_0) + x_+^2\mathrm{Var}(\tilde{\beta}_{1\text{pretest}}) = \tfrac{\sigma^2}{n} + x_+^2\mathrm{Var}(\tilde{\beta}_{1\text{pretest}}).
\end{aligned} \tag{3.14}$$

$|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}}$ then $Z_1 \geq q$ or $Z_1 < r$.

$\mathrm{E}(A_{\text{pretest}}) = \int_{-\infty}^{r}(b_1 + z)\phi(z)dz + \int_{q}^{\infty}(b_1 + z)\phi(z)dz$

$= b_1 \int_{-\infty}^{r} \phi(z)dz + \int_{-\infty}^{r} z\phi(z)dz + b_1 \int_{q}^{\infty} \phi(z)dz + \int_{q}^{\infty} z\phi(z)dz$

$= b_1(\Phi(r) + 1 - \Phi(q)) + \int_{-\infty}^{r} z\phi(z)dz + \int_{q}^{\infty} z\phi(z)dz$.

Now $\int_{-\infty}^{r} z\phi(z)dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} ze^{-z^2/2}dz = -\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\Big|_{-\infty}^{r} = -\phi(r)$.

Similarly, $\int_{q}^{\infty} z\phi(z)dz = \phi(q)$.
Therefore, $\mathrm{E}(A_{\text{pretest}}) = b_1(\Phi(r) + 1 - \Phi(q)) - \phi(r) + \phi(q)$.

$\mathrm{Var}(A_{\text{pretest}}) = \mathrm{E}(A^2_{\text{pretest}}) - \mathrm{E}^2(A_{\text{pretest}}) = \int_{-\infty}^{r}(b_1+z)^2\phi(z)dz + \int_{q}^{\infty}(b_1+z)^2\phi(z)dz - \mathrm{E}^2(A_{\text{pretest}})$

$= b_1^2 \int_{-\infty}^{r} \phi(z)dz + 2b_1 \int_{-\infty}^{r} z\phi(z)dz + \int_{-\infty}^{r} z^2\phi(z)dz + b_1^2 \int_{q}^{\infty} \phi(z)dz + 2b_1 \int_{q}^{\infty} z\phi(z)dz + \int_{q}^{\infty} z^2\phi(z)dz - \mathrm{E}^2(A_{\text{pretest}})$

$= b_1^2(\Phi(r) + 1 - \Phi(q)) + 2b_1(-\phi(r) + \phi(q)) - r\phi(r) + \Phi(r) + q\phi(q) + 1 - \Phi(q) - \mathrm{E}^2(A_{\text{pretest}})$.
Now, $\tilde{\beta}_{1\text{pretest}} = v_1^{1/2}A_{\text{pretest}}$, then $\mathrm{E}(\tilde{\beta}_{1\text{pretest}}) = v_1^{1/2}\mathrm{E}(A_{\text{pretest}})$ and $\mathrm{Var}(\tilde{\beta}_{1\text{pretest}}) = v_1\mathrm{Var}(A_{\text{pretest}})$. Replacing these values in (3.14) yields the result.
The important thing to note in these derivations is that the moment of the pretest estimator depend on the data through v_1=$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}$.
This means that only the quantity $Sxx = \Sigma_{i=1}^{n}(x_i - \overline{x})^2$ is useful. This explains why PMSEs are not sensitive to the data, therefore, it does not matter whether one uses real or simulated data. For multivariate regression, the important quantity is $(X'X)^{-1}$. The properties of PMSEs do not change too much from one data set to another. We will therefore mostly use simulated data in the illustration.

3.6.1.2 Post-information theory approach

To choose between M_0 and M_1, let consider model selection procedure of the form

$$IC_k = -2\log L_k(\hat{\theta}_k) + h_n p_k, \tag{3.15}$$

where $L_k(\hat{\theta}_k) = L_k$ is the likelihood value for model M_k, $p_1 = 2$, $p_0 = 1$.

Lemma 3.6.1 *Under (3.15),*

$$\frac{L_1}{L_0} = e^{\frac{1}{2}(Z_1+b_1)^2}, \tag{3.16}$$

$$IC_1 - IC_0 = -(Z_1+b_1)^2 + h_n. \tag{3.17}$$

Proof. $IC_1 - IC_0 = -2(\log L_1 - \log L_0) + (p_1 - p_0)h_n = -2\log\{\frac{L_1}{L_0}\} + h_n.$

$\log L_1 = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma_{i=1}^n(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$

$\log L_0 = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma_{i=1}^n(y_i - \hat{\beta}_0)^2.$

$\log L_1 - \log L_0 = \frac{1}{2\sigma^2}[\Sigma_{i=1}^n\{(y_i - \overline{y})^2 - (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\}]$

$= \frac{1}{2\sigma^2}\Sigma_{i=1}^n[2(y_i - \overline{y}) - \hat{\beta}_1^2 x_i^2] = \frac{1}{\sigma^2}\Sigma_{i=1}^n(y_i - \overline{y}) - \frac{1}{2\sigma^2}\Sigma_{i=1}^n\hat{\beta}_1^2 x_i^2$

$= -\frac{1}{2\sigma^2}\hat{\beta}_1^2\Sigma_{i=1}^n x_i^2 = -\frac{1}{2\sigma^2}\hat{\beta}_1^2\frac{\sigma^2}{v_1} = -\frac{1}{2}\frac{\hat{\beta}_1^2}{v_1} = -\frac{1}{2}(Z_1+b_1)^2$, yielding the result.

Corollary 3.6.1 *Under the model (3.5) and (3.15), the mean, bias, variance and MSE of PMSE are given by*

$$\begin{aligned}
&\mathrm{E}(\tilde{\triangle}_{IC}) = \beta_0 + x_+ v_1^{1/2}[b_1(\Phi(r_n) + 1 - \Phi(q_n)) - \phi(r_n) + \phi(q_n)],\\
&\mathrm{Bias}(\tilde{\triangle}_{IC}) = x_+ v_1^{1/2}[b_1(\Phi(r_n) - \Phi(q_n)) - \phi(r_n) + \phi(q_n)],\\
&\mathrm{Var}(\tilde{\triangle}_{IC}) = \tfrac{\sigma^2}{n} + x_+^2 v_1[(b_1^2+1)(\Phi(r_n) + 1 - \Phi(q_n)) + 2b_1(-\phi(r_n)\\
&+\phi(q_n)) - r_n\phi(r_n) + q_n\phi(q_n) - \mathrm{E}^2(\tilde{\triangle}_{IC})],\\
&\mathrm{MSE}(\tilde{\triangle}_{IC}) = \mathrm{Bias}^2(\tilde{\triangle}_{IC}) + \mathrm{Var}(\tilde{\triangle}_{IC}),
\end{aligned} \tag{3.18}$$

where $r_n = -h_n^{1/2} - b_1$ *and* $q_n = h_n^{1/2} - b_1$.

Proof. The model M_1 is chosen if $IC_1 - IC_0 < 0$, using Lemma 3.6.1, the partition of the sample space is

$\mathcal{X}_0(IC, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : |Z_1 + b_1| = \frac{|\hat{\beta}_1|}{v_1^{1/2}} < h_n^{1/2}\}$,
$\mathcal{X}_1(IC, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : |Z_1 + b_1| = \frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq h_n^{1/2}\}$.
This means that the analysis above for the case pre-test is valid by replacing $z_{1-\frac{\alpha}{2}}$ by $h_n^{1/2}$. The result then follows from Proposition 3.6.1.
This is an important connection between hypothesis testing as model selection method and information criterion of the form (3.15). For $z_{1-\frac{\alpha}{2}} = h_n^{1/2}$ (that is $h_n = z^2_{1-\frac{\alpha}{2}}$), the two forms of model selection methods are equivalent for

$$\alpha^* = 2[1 - \Phi(h_n^{1/2})]. \tag{3.19}$$

Special cases of (3.15) are AIC ($h_n = 2$), BIC ($h_n = \log n$), HQ ($h_n = \log\log n$). E.g., for pre-test to be equivalent to AIC, $\alpha^* = 0.16$. For information criteria of the form Equation (3.15),

$$h_n = \begin{cases} z^2_{1-\frac{\alpha}{2}} & \text{for Hypothesis Testing} \\ 2 & \text{for AIC} \\ \log(n) & \text{for BIC} \\ \log(\log(n)) & \text{for HQ.} \end{cases}$$

3.6.1.3 Post-Mallows Cp estimators

$Cp_k = \frac{RSS_k}{S_K^2} - n + 2p_k$, where $S_k^2 = \frac{RSS_k}{n-2}$, $K = 1$, corresponding to M_1 and RSS are the residuals sum of squares.
$Cp_0 = \frac{RSS_0}{S_1^2} - n + 2$ and $Cp_1 = \frac{RSS_1}{S_1^2} - n + 2$.
Model M_k is chosen if $Cp_1 - Cp_0 < 0$.
$Cp_1 - Cp_0 = (n-2)\frac{RSS_1 - RSS_0}{RSS_1} + 2$, because of normality, $\frac{(n-2)S_1^2}{\sigma^2} \sim \chi^2_{n-2}$ and are independent of $RSS_1 - RSS_0$, this gives that $RSS_1 - RSS_0$ and RSS_1 are independent. From likelihood computations in pre-test,
$RSS_1 - RSS_0 = -\frac{\hat{\beta}_1^2\sigma^2}{v_1} = -(b_1 + Z_1)^2\sigma^2$, therefore $Cp_1 - Cp_0 = (n-2)\frac{(b_1+Z_1)^2}{\chi^2_{n-2}} + 2 = (n-2)\frac{\chi^2_1(b_1^2)}{\chi^2_{n-2}} + 2 = -F(1, n-2, b_1^2) + 2$, a non-central F distribution with 1 and n-2 degrees of freedom with non-central parameter b_1^2. The partition of the sample space is then
$\mathcal{X}_0(Cp, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : F(1, n-2, b_1^2) \leq 2\}$,
$\mathcal{X}_1(Cp, \mathcal{M}) = \{X = (X_1, \ldots, X_n) : F(1, n-2, b_1^2) > 2\}$.

The behaviour of post-Cp model selection can be derived as above (pre-test) where A_{pretest} is replaced. That is

$$\tilde{\beta}_{1Cp} = v_1^{1/2}(Z_1 + b_1)I_1(F(1, n-2, b_1^2) > 2) = v_1^{1/2}A_{Cp},$$

where $A_{Cp} = (Z_1 + b_1)I_1(F(1, n-2, b_1^2) > 2)$.

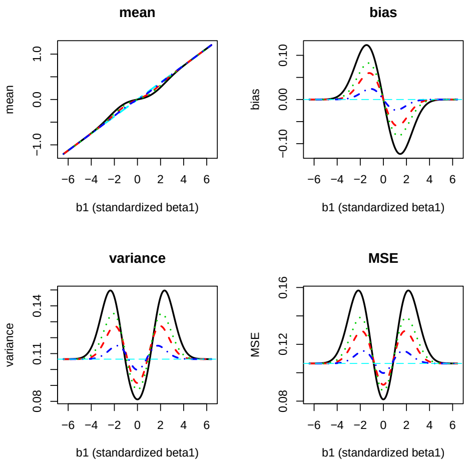

Figure 3.4: Mean, bias, variance and MSE of PMSEs as function of b_1, pre-test (solid line), post-BIC (dotted line), post-AIC (dashed line), post-HQ (broken line).

3.6.1.4 Optimal selection criteria and optimal significance level α

The properties of PMSEs are given in Figure (3.4). For all model selection criteria, when $|b_1|$ is large enough, $|b_1| > 4$, the variance, bias and MSE of PMSEs are close to the values corresponding to those of the model M_1. The performances of PMSE are similar to that of the larger model (properties of MLE full model, the same variance, same MSE and unbiased). This is because model M_1 is likely to be selected for $|b_1|$ is large enough.

Maximum variance and maximum MSE are smaller for post-HQ, followed by post-AIC, post-BIC and pre-test for $|b_1|$ large, and the reverse is observed for $|b_1|$ small. The bias is always smaller for post-HQ, followed by post-AIC, post-BIC and pre-test. As the penalty h_n decreases, the bias also decreases. This means that if a choice is based on the bias, one should choose the criterion with the smaller penalty. However, this is only true for this simple example. In general, it is not always possible to choose among criteria based on their bias. Post-AIC and Cp have the same performance and are better than other criteria for $|b_1|$ large and worse for $|b_1|$ small. Pre-test estimators just have the reverse effect. The risk of post-BIC estimator is between that of post-AIC risk and pre-test risk. The worse performance of post AIC near 0 is due to overfitting properties of AIC. The variance of the prediction is 1.13. From Figure (3.5), the MSE of all the criteria cross, meaning that there is no single criterion that dominates all the others in terms of MSE.

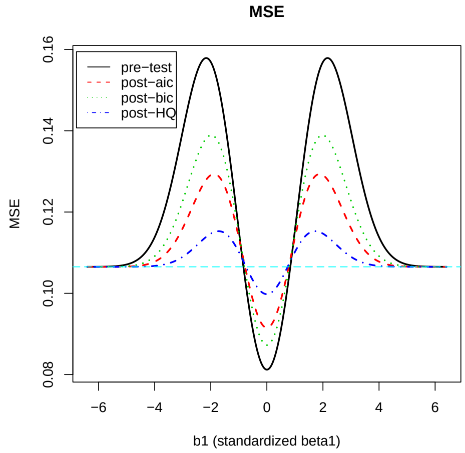

Figure 3.5: MSE of PMSEs as function of b_1.

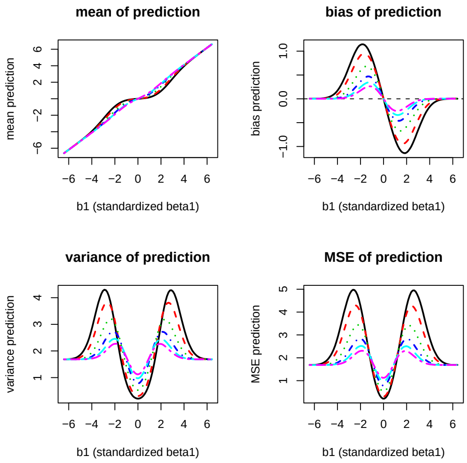

Figure 3.6: Mean, bias, variance and MSE of pre-test as function of b_1 for different values of $\alpha \in [0.01$ (solid line), 0.02 (dashed line), 0.05 (dotted line), 0.10 (broken line), 0.15, 0.20].

As can be seen in Figure (3.6), as the α increases, the bias gets smaller uniformly. For the variance and MSE, this is only valid when the parameter is too large, namely when the null hypothesis is clearly rejected. For b_1 small, the variance and MSE increase as α gets larger. This means that, if interest is focused on unbiased estimation, one should choose α reasonably large. But this shows that, in terms of risk, there is no optimal value of α, as illustrated in Figure (3.6).

In general, there is no optimal choice of the penalty h^* from the set $\{h \in \mathbb{R} : h > 0\}$ to get a criterion that dominates all the others in terms of MSE (all the risks cross). For this example, when the penalty is small, the full model is likely to be selected, therefore the risk function is smaller for $|b_1|$ larger and larger for $|b_1|$ smaller. The reverse happens when the penalty is larger. This observation explains why the risks will cross, making the choice of a particular criterion as best difficult. That there is no optimal level of significance, this fact is not surprising since pre-testing is equivalent to using an information criterion of the form (3.15).

3.6.2 Multiple linear regression

Consider the following multiple regression model

$$Y = X_1\beta_1 + \ldots + X_p\beta_p + \varepsilon = X\beta + \varepsilon, \tag{3.20}$$

where Y is $n \times 1$, $X = (X_1, \ldots, X_p)$ is $n \times p$, $\beta = (\beta_1, \ldots, \beta_p)'$ is $p \times 1$, and $\varepsilon \sim N_n(0, \sigma^2 I_n)$. Suppose that interest is focused on the estimation of the mean $\triangle = \mathrm{E}(Y) = X\beta$. This means that it is enough to have an estimate of the β.

Suppose that one proceeds with hypothesis testing, and then based on the outcome, estimation is performed. Consider the following testing problem

$$H_0 : R\beta = r \quad \text{against} \quad H_1 : R\beta \neq r, \tag{3.21}$$

where R is $m \times p$ known matrix with rank m and r is $m \times 1$ vector. This corresponds to a choice between two models: unrestricted model (model M_1) and a restricted model (model M_0). Let $\hat{\beta}^0$ and $\hat{\beta}$ be restricted and unrestricted estimators of β and let $A = X'X; \delta = R\beta - r$, Then under $H_0 : \delta = 0$.
The OLS estimators of β under model M_1 and model M_0 are given by
$\hat{\beta} = (X'X)^{-1}X'Y$,
$\hat{\beta^0} = \hat{\beta} - A^{-1}R'(RA^{-1}R')^{-1}(R\beta - r)$.
The OLS estimator of σ^2 under M_1 is given by $\hat{\sigma}^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{n-p}$.
Let $F = \frac{(R\beta-r)'(RA^{-1}R')^{-1}(R\beta-r)}{m\hat{\sigma}^2}$ and $\lambda = \frac{\delta'(RA^{-1}R')^{-1}\delta}{2\sigma^2}$.
Then under $H_0, F \sim F(m, n-p)$ and under $H_1, F \sim F(m, n-p, \lambda)$ non-central

F with non-centrality parameter λ.
The PMSE is $\tilde{\triangle}$ given by

$$\tilde{\triangle}_{\text{pretest}} = \begin{cases} X\hat{\beta}^0 & F < F^{1-\alpha}(m, n-p) \\ X\hat{\beta} & F \geq F^{1-\alpha}(m, n-p), \end{cases} \tag{3.22}$$

where $F^{1-\alpha}(m, n-p)$ is the $100(1-\alpha)^{th}$ percentile of F distribution with m and $n-p$ degree of freedom. Since the behaviour of $\tilde{\triangle}$ is given by that of $\tilde{\triangle}$

$$\tilde{\beta} = \begin{cases} \hat{\beta}^0 & F < F^{1-\alpha}(m, n-p) \\ \hat{\beta} & F \geq F^{1-\alpha}(m, n-p). \end{cases} \tag{3.23}$$

We will restrict analysis to the properties of $\tilde{\beta}$. These properties are derived in Judge and Bock (1978). Let W be a positive weighted matrix, a weighted quadratic loss for an estimator $\hat{\triangle}$ is

$$L(\triangle, \hat{\triangle}) = (\hat{\triangle} - \triangle)'W(\hat{\triangle} - \triangle)$$

and the risk of $\hat{\triangle}$ is Risk($\hat{\triangle}$)=$E[L(\triangle, \hat{\triangle})]$.
We define the following quantities:
$f_\lambda(s) = \text{Prob}(F(m+s, n-p, \lambda) < \frac{F^{1-\alpha}(m,n-p)m}{m+s})$,
$\zeta = \frac{\delta'(RA^{-1}R')^{-1}RA^{-1}WA^{-1}R'(RA^{-1}R')^{-1}\delta}{2\sigma^2}$,
$U = RA^{-1}WA^{-1}R'(RA^{-1}R')^{-1}$.
The properties of the estimators under M_0, M_1 and the PMSEs are given by

$\text{Bias}(\hat{\beta}^0) = -A^{-1}R'(RA^{-1}R')^{-1}\delta$,

$\text{Bias}(\tilde{\beta}) = -\text{Prob}(F(m+2, n-p, \lambda) < \frac{F^{1-\alpha}(m,n-p)m}{m+2})A^{-1}R'(RA^{-1}R')^{-1}\delta$,

$\text{Bias}(\hat{\beta}) = 0$,

$\text{Var}(\hat{\beta}^0) = \sigma^2[A^{-1} - A^{-1}R'(RA^{-1}R')^{-1}RA^{-1}]$,

$\text{Var}(\hat{\beta}) = \sigma^2 A^{-1}$,

$\text{Var}(\tilde{\beta}) = \sigma^2 A^{-1} + \sigma^2 f_\lambda(2)A^{-1} - [2f_\lambda(2) - f_\lambda(4)$
$+ f_\lambda^2(2)][A^{-1}R'(RA^{-1}R')^{-1}\delta][\delta'(RA^{-1}R')^{-1}RA^{-1}]$,

$\text{Risk}(\hat{\beta}^0) = \sigma^2 tr(A^{-1}W) - \sigma^2 tr(U) + 2\sigma^2\zeta$,

$\text{Risk}(\hat{\beta}) = \sigma^2 tr(A^{-1}W)$,

$\text{Risk}(\tilde{\beta}) = \sigma^2 tr(A^{-1}W) - \sigma^2 tr(U)[f_\lambda(2) + 2(f_\lambda(4) - 2f_\lambda(2))\frac{\zeta}{tr(U)}]$.

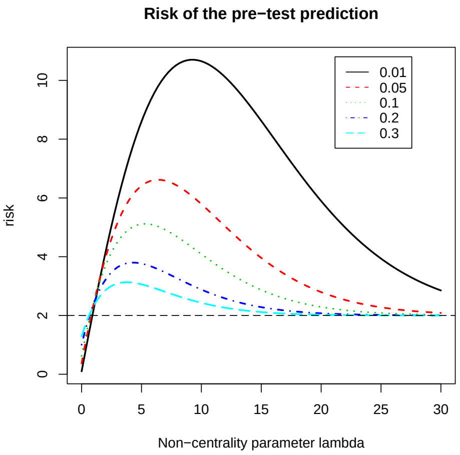

Figure 3.7: Risk bounds of the pre-test estimator as a function of the non-centrality parameter λ for different level of significance α.

For simplicity, the risks of these estimators are computed only for orthogonal designs, that is for $A = X'X = I_p = W$. The risk of $\tilde{\beta}$ is then the same as that of $\tilde{\triangle}$ since $E[(\tilde{\triangle} - \triangle)'(\tilde{\triangle} - \triangle)] = E[(X\tilde{\beta} - X\beta)'(X\tilde{\beta} - X\beta)]$
$= E[(\tilde{\beta} - X\beta)'X'X(\tilde{\beta} - \beta)] = E[(\tilde{\beta} - X\beta)W(\tilde{\beta} - \beta)] = \text{Risk}(\tilde{\beta})$.
That is the simple risk of $\tilde{\triangle}$ is the weighted risk of $\tilde{\beta}$.

Figure (3.7) illustrates the risk for multiple regression for an orthogonal design, $n = 30$ and $m = p = 3$ (intercept and two covariates). For large values of the non-centrality parameter λ corresponding to the rejection of the null hypothesis, the risk gets smaller as the significance level increases (the maximum risk also decreases), as illustrated in Table (3.1). For very large values of λ, the risks are close to that of full model (MLE estimator), the horizontal line in Figure (3.7). However, for smaller values of λ, corresponding to the non-rejection of the null hypothesis, as the significance level increases, the risk gets larger. Therefore, there is no optimal level of α that guarantees that its risk will uniformly dominate

α	0.01	0.05	0.10	0.20	0.30
λ^*	9.25	6.39	5.29	4.30	3.77
$risk^*$	10.70	6.62	5.12	3.80	3.13

Table 3.1: Significance level α and the corresponding maximum risk and λ.

the other risks (all the risks cross).

3.6.3 Testing for a given variance

Consider a modification of the multiple linear regression (3.20) by allowing a possibility for known or unknown variance. Consider two models M_0 corresponding to model (3.20) with known variance, that is $\sigma^2 = \sigma_0^2$ and M_1 to a model (3.20). If $\sigma^2 = \sigma_0^2$, then M_0 is selected, otherwise model M_1 is selected. That is testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$.
We have that T=$\frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-p}$. One can base the decision on the test statistic T and reject H_0 if T is large and does not reject H_0 if T is smaller. This can be done by finding two critical points $t_1 = \chi^2_{\alpha/2}(n-p)$ and $t_2 = \chi^2_{1-\alpha/2}(n-p)$ corresponding to $(100\alpha/2)^{th}$ and $100(1-\alpha/2)^{th}$ percentiles of the chi-squared distribution with $n-p$ degrees of freedom.
The behaviour of any quantity of interest after this test is governed by that of the PMSE $\tilde{\sigma}^2$, whose properties and risks are studied in Judge and Yancy (1986) and expressed below

$$\tilde{\sigma}^2 = \begin{cases} \sigma_0^2 & t_1 < \frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} < t_2 \\ \hat{\sigma}^2 & \text{otherwise}, \end{cases} \tag{3.24}$$

where $\hat{\sigma}^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{n-p}$.
The risk of $\tilde{\sigma}^2$ under square error loss is

Risk($\tilde{\sigma}^2$) $= \frac{2\sigma^4}{n-p} + \sigma^4[(1/\delta^2 - 2/\delta)P(t_1/\delta < \chi^2_{n-p} < t_2/\delta)$
$+ 2P(t_1/\delta < \chi^2_{n-p+2} < t_2/\delta) - \frac{n-p+2}{n-p}P(t_1/\delta < \chi^2_{n-p+4} < t_2/\delta)]$, with $\delta = \frac{\sigma^2}{\sigma_0^2}$.

Risk($\hat{\sigma}^2$) $= \frac{2\sigma^4}{n-p}$, the risk of MLE.
To facilitate the comparison between risk of MLE and pre-test risk, we divide the risk by that of MLE, so the MLE risk is normalized to 1. From Figure (3.8) with $\sigma_0 = 1$, $n = 30$, $p = 2$ and $\alpha = 0.05$, we see that the pre-test estimator is biased and has higher variance and MSE unless δ is larger (rejection of null hypothesis), MLE risk and pre-test risk are close for δ close to 1, H_0 not rejected and MLE is better than pre-test in terms of variance and MSE.

3.6.4 Testing for a given mean

Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, σ known and interest is focused on choosing a model $M_0 : \mu = \mu_0$ and $M_1 : \mu \neq \mu_0$.
Suppose one uses hypothesis testing $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. We can assume $\mu_0 = 0$ without loss the generality since in the case $\mu_0 \neq 0$, transform

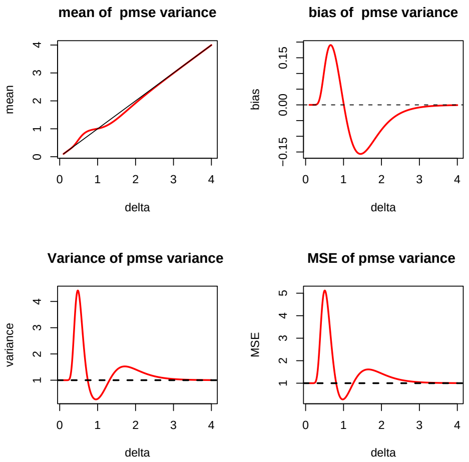

Figure 3.8: Mean, bias, variance and MSE functions of the pre-test estimator as a function δ for $\alpha = 0.05$.

$\lambda = \mu - \mu_0$ and test for $\lambda = 0$.
The decision is then the following: choose M_0 if $\frac{|\hat{\mu}|}{v^{1/2}} < z_{1-\frac{\alpha}{2}}$ and M_1 otherwise, where $\hat{\mu} = \overline{X}$ and $v = \sigma^2/n$. The PMSE for μ is then given by

$$\tilde{\mu} = \begin{cases} 0 & \frac{|\hat{\mu}|}{v^{1/2}} < z_{1-\frac{\alpha}{2}} \\ \hat{\mu} & \frac{|\hat{\mu}|}{v^{1/2}} \geq z_{1-\frac{\alpha}{2}}. \end{cases} \tag{3.25}$$

Let $Z = \frac{\hat{\mu}-\mu}{v^{1/2}}$, $\delta = \frac{\mu}{v^{1/2}}$, $\hat{\delta} = \frac{\hat{\mu}}{v^{1/2}}$, and $\hat{\delta} = Z + \delta$ and $Z \sim N(0,1)$. Equation (3.25) can then be written as

$$\tilde{\mu} = v^{1/2}(Z+\delta)I(|Z+\delta| \geq z_{1-\frac{\alpha}{2}}) = v^{1/2}A_m, \tag{3.26}$$

where $A_m = (Z+\delta)I(|Z+\delta| \geq z_{1-\frac{\alpha}{2}})$.
The properties of $\tilde{\mu}$ can then be derived as Equation (3.11).
For a test of proportion, let $X_i \sim$ Bernouilli(π), $i = 1, \ldots, n$. Choosing between 2 models $M_0 : \pi = \pi_0$ and $M_1 : \pi \neq \pi_0$, using the large sample approximation, $Z = \frac{\hat{\pi}-\mu_0}{v^{1/2}} \longrightarrow N(0,1)$, with $v = \frac{\pi_0(1-\pi_0)}{n}$, properties of PMSEs for μ are given as above.

Figure (3.9) ($n = 20$ and $\sigma = 1$) compares pre-test estimators for various values of α. For any data set for testing for normality, the behaviour is the same as illustrated in the case of simple linear regression. The horizontal line corresponds to the properties of unbiased MLE $\overline{X} = \hat{\mu}$. The pre-test for proportion yields the same behaviour.

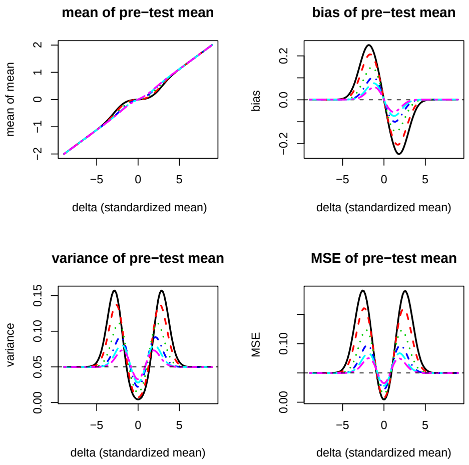

Figure 3.9: Mean, bias, variance and MSE of pre-test as function of true standardised mean δ for different values of $\alpha \in$ [0.01 (solid line), 0.02 (dashed line), 0.05 (dotted line), 0.10 (broken line), 0.15, 0.20].

3.7 Partition of the sample space

Consider the partition of the sample space given in Equation (3.4). Let $\tilde{\mathcal{X}}(X|S,\mathcal{M})$ be the (random) selected subset. Let E an arbitrary even. From the law of total probability,

$$P_\theta(E) = \Sigma_{k=1}^{K} P_\theta(E|\tilde{\mathcal{X}}(X|S,\mathcal{M}) = \mathcal{X}_k) P_\theta(\tilde{\mathcal{X}}(X|S,\mathcal{M}) = \mathcal{X}_k), \tag{3.27}$$

where $P_\theta(\tilde{\mathcal{X}}(X|S,\mathcal{M}) = \mathcal{X}_k)$ is the probability of subset $\mathcal{X}_k$, i.e. the probability of selecting M_k. $P_\theta(E|\tilde{\mathcal{X}}(X|S,\mathcal{M}) = \mathcal{X}_k)$ is the probability of the event E given that model M_k is selected. To simplify the notation, we denote by $\tilde{\mathcal{X}}_k(X|S,\mathcal{M})$ the event ”$\tilde{\mathcal{X}}(X|S,\mathcal{M}) = \mathcal{X}_k$”, that is the event that model M_k is selected. Later we will also use the simpler notation $\tilde{\mathcal{X}}_k$.

3.7.1 Model selection probabilities

We will refer to $P_\theta(\tilde{\mathcal{X}}_k(X|S,\mathcal{M}))$ as the *model selection probability* for model M_k, given as

$$P_\theta(\tilde{\mathcal{X}}_k(X|S,\mathcal{M})) = P_\theta(\tilde{M}(X|S,\mathcal{M}) = M_k) = E_\theta(I_k(X|S,\mathcal{M})). \tag{3.28}$$

The subscript is needed to indicate that this is a function of the parameter θ. For the simple linear model and pre-test selection procedure, the probability of

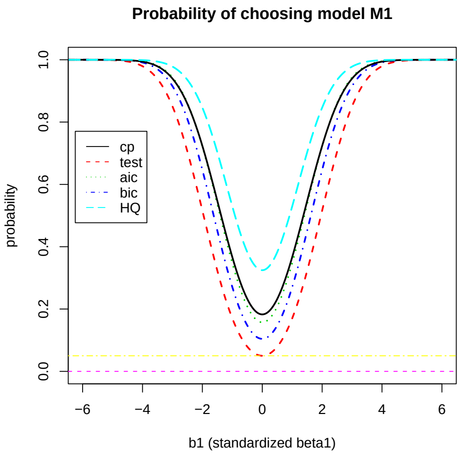

Figure 3.10: Probability of selecting model M_1 for different selection procedures as function of standardized slope b_1, $\alpha = 0.05$.

selecting M_1 is given by
$P_\theta(\tilde{M}(X|\text{pre-test}, \mathcal{M}) = M_1) = E_\theta(I_1(X|\text{pre-test}, \mathcal{M}))$
=P(rejecting H_0)=$P(Z > q$ or $Z < r)$
=$\Phi(r) + 1 - \Phi(q)$.
Note that $P_\theta(\tilde{M}(X|\text{pre-test}, \mathcal{M}) = M_1)$ is simply the power of the test.

For an information criterion of the form Equation (3.15), the probability of selecting M_1 is $P_\theta(\tilde{M}(X|IC_n, \mathcal{M}) = M_1) = E_\theta(I_1(X|IC_n, \mathcal{M}))$
$=P(Z > q_n$ or $Z < r_n)$=$\Phi(r_n) + 1 - \Phi(q_n)$.

From Figure (3.10) with $\alpha = 0.05$, as can be expected, for $|b_1|$ large, the probability is almost 1. One is likely to choose model M_1. For any b_1, this probability is smaller for pre-test, followed by post-BIC, post-AIC and post-Cp. For $b_1 = 0$, this probability $\Phi(-z_{1-\alpha/2}) + 1 - \Phi(z_{1-\alpha/2})$ for pre-test and $\Phi(-h_n^{1/2}) + 1 - \Phi(h_n^{1/2})$ for information criteria of the form (3.15). One would have expected these probabilities to be 0. However, due to overfitting properties of such model selection criteria for finite sample size, this is not the case. As is already known, AIC and Cp “overfit” more than hypothesis testing and BIC. When the significance level α gets small, for $|b_1| = 0$, the probability is near 0, indicating the underfitting properties of hypothesis testing. That is BIC and pre-test for α smaller tend to select models with fewer parameters.

The partition of the sample space that leads to Equation (3.27) is not only relevant to model selection problems, but also to any decision problem, as has been explained in the decision part.

Let k^* be the index corresponding to the naive model. Then Equation (3.27) can be written as

$$P_\theta(E) = \underbrace{P_\theta(E|\tilde{\mathcal{X}}_{k^*})P_\theta(\tilde{\mathcal{X}}_{k^*})}_{(1)} + \underbrace{\Sigma_{k=1,k\neq k^*}^{K} P_\theta(E|\tilde{\mathcal{X}}_k)P_\theta(\tilde{\mathcal{X}}_k)}_{(2)}. \tag{3.29}$$

By using the naive procedure, one assumes (implicitely) that $P_\theta(\tilde{\mathcal{X}}_{k^*}) = 1$, therefore that (2) is zero; one then uses only the (unconditional) distribution of M_{k^*}. Some particular cases correspond to the correct distribution of PMSEs, the computation of p-values, the coverage probability and the consistency of PMSE, each being an event.

3.7.2 Distribution of PMSEs

From (3.27), the (unconditional) distribution of $\tilde{\triangle}(X|S,\mathcal{M})$ is given by

$$P_\theta(\tilde{\triangle}(X|S,\mathcal{M}) \leq s) = \Sigma_{k=1}^{K} P_\theta(\tilde{\triangle}(X|S,\mathcal{M}) \leq s|\tilde{\mathcal{X}}_k)P_\theta(\tilde{\mathcal{X}}_k), \tag{3.30}$$

i.e. a mixture of conditional distributions $P_\theta(\tilde{\triangle}(X|S,\mathcal{M}) \leq s|\tilde{\mathcal{X}}_k)$. The moments of this estimator can be computed, provided that they exist.

3.7.2.1 Distribution of likelihood ratio

Consider an information criterion of the form of Equation (3.15), for simplicity with fixed h (for e.g., $h = 2$ for AIC), and let M_0 be the true model.
$IC_k - IC_0 = -2\log[\frac{L_k(\hat{\theta}_k)}{L_0(\hat{\theta}_0)}] + h(p_k - p_0)$. It is well known that $2\log[\frac{L_k(\hat{\theta}_k)}{L_0(\hat{\theta}_0)}]$ tends in distribution to χ^2 with $(p_k - p_0)$ degrees of freedom, as n goes to ∞. However, if a selection criterion is applied to get $\tilde{\theta} = \hat{\theta}_{\hat{k}}$ with the corresponding model $M_{\hat{k}}$, then the distribution of $2\log[\frac{L_{\hat{k}}(\hat{\theta}_{\hat{k}})}{L_0(\hat{\theta}_0)}]$ is not necesarily chi-squared and is not easy to compute. Therefore the distribution of $IC_{\hat{k}} - IC_0$ is not easily obtained; it is likely to be a mixture.

3.7.2.2 The simple linear regression example

Let $Z_0 = \frac{\hat{\beta}_0 - \beta_0}{v_0^{1/2}}$, then $Z_0 \sim N(0,1)$, then $\hat{\beta}_0 = v_0^{1/2}(Z_0 + b_0)$, where $b_0 = \frac{\beta_0}{v_0^{1/2}}$ (standardized intercept), $v_0 = \frac{\sigma^2}{n}$.
From Equation (3.7), the pre-test estimators,

$$\tilde{\triangle}_{\text{pretest}} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1)I_1(|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}}), \tag{3.31}$$

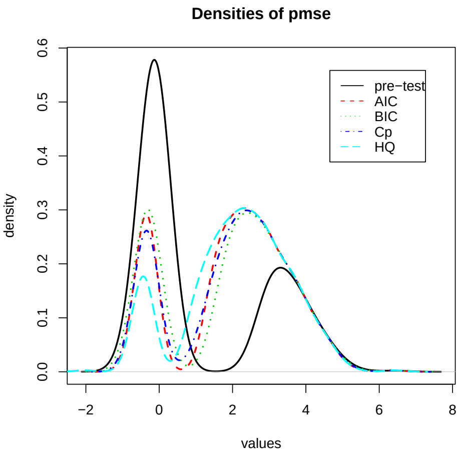

Figure 3.11: Densities for PMSEs for $b_1 = 0.2$.

where Z_0 and Z_1 are independent normal since $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent. For model selection criterion of the form of Equation (3.15), PMSEs (e.g. AIC, BIC) are given by

$$\tilde{\triangle}_{IC} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1) I_1(|Z_1 + b_1| \geq h_n^{1/2}) \tag{3.32}$$

and for Mallows Cp, PMSE is given by

$$\tilde{\triangle}_{Cp} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1) I_1(F(1, n-2, b_1^2) > 2). \tag{3.33}$$

Figure (3.11) illustrates the non-normal and mixture nature of PMSEs for $b_1 = 0.1$. Figure (3.12) displays different finite sample distributions for the post-AIC estimation. Other PMSEs have similar behaviour. As $|b_1|$ gets larger, the distribution becomes uni-modal. This is because, it is likely that model 1 will be chosen, because the probability of selecting model M_1 approaches one for large values of $|b_1|$. However, since one does not know the true value of the parameter, from Equation (3.30), PMSEs are in general not normal.

3.7.3 Coverage probabilities

Let

$$\begin{aligned}\text{CI} = [\tilde{\triangle}(x|S, \mathcal{M}) - z_{1-\alpha/2}\text{Var}^{1/2}[\tilde{\triangle}(x|S, \mathcal{M})], \tilde{\triangle}(x|S, \mathcal{M}) +\\ z_{1-\alpha/2}\text{Var}^{1/2}[\tilde{\triangle}(x|S, \mathcal{M})]]\end{aligned}$$

be a naive confidence interval for $\triangle$ under normality assumption, and $1 - \alpha$ the nominal level for each model been considered as true model, that is, $P_k(\theta \in CI) =$

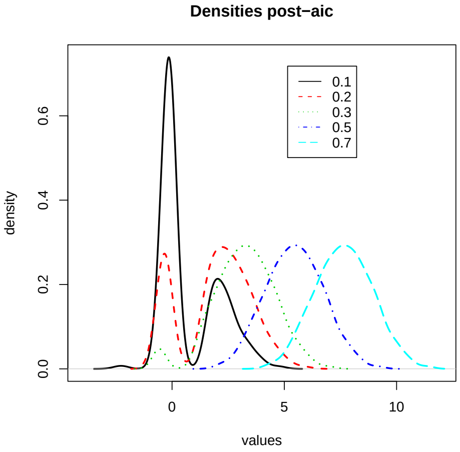

Figure 3.12: Densities for post-AIC estimators for different values of b_1.

$1-\alpha$, under the use of model M_k. The event to consider here is "$\triangle \in CI$": from Equation (3.30), the coverage probability of such interval is given by

$$P_\theta(\triangle \in CI) = \Sigma_{k=1}^{K} P_\theta(\triangle \in CI|\tilde{\mathcal{X}}_k)P_\theta(\tilde{\mathcal{X}}_k), \tag{3.34}$$

where $P_\theta(\triangle \in CI|\tilde{\mathcal{X}}_k)$ is the conditional coverage probability. The nominal (naive) coverage for model M_k is $P_k(\triangle \in CI) = 1-\alpha$.

3.7.4 P-value and goodness of fit tests after model selection

Consider an hypothesis of the form

$$H_0 : X \sim g(x|\theta), \tag{3.35}$$

and a sample (data) x. The question is whether data x are compatible with H_0. To investigate this compatibility, one chooses a statistic $\triangle(X)$. Suppose that and large values of $\triangle(X)$ indicate that data are less compatible with H_0. A common measure of compatibility is the p-value defined by p-value=$P_\theta(\triangle(X) \geq \triangle(x))$. The null hypothesis is rejected for smaller values, for e.g. 0.05.

We are concerned with the choice of the statistic $\triangle(X)$. In general $\triangle(X)$ depends on the estimator $\hat{\theta}$ of θ. Suppose that one considers a set of models $\mathcal{M} = (M_1, \ldots, M_K)$, uses a model selection criterion, for e.g. AIC, to select a model and then estimate θ under that model. The statistic is then $\tilde{\triangle}(X|S,\mathcal{M}) = h(\tilde{\theta}(X|S,\mathcal{M}))$, where $\tilde{\theta}(X|S,\mathcal{M})$ is the PMSE of θ. For e.g., a set of model can be under H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta_0^c$ (complement of Θ_0), and respective estimators are $\hat{\theta}_0$ and $\hat{\theta}_1$, and one model is selected. Consider the event "$\tilde{\triangle}(X|S, \mathcal{M}) > \tilde{\triangle}(x|S, \mathcal{M}) = \triangle_{k^*}(x)$", $\triangle_{k^*}(x)$ being the observed statistic for the selected model. The probability of this event is given by

$$P_\theta(\tilde{\triangle}(X|S, \mathcal{M}) > \triangle_{k^*}(x)) = \Sigma_{k=1}^{K} P_\theta(\tilde{\triangle}(X|S, \mathcal{M}) > \triangle_{k^*}(x)|\tilde{\mathcal{X}_k}) P_\theta(\tilde{\mathcal{X}_k}) \quad (3.36)$$

where the probability is computed with respect to H_0. Equation (3.36) gives the valid p-value after model selection. A p-value of 1 will indicate that model g fits the data well. If one does not compute p-value as in Equation (3.36), the resulting decision could be wrong. This is not uncommon in applied work where one first uses a selection criterion (e.g. AIC) to select a model and then tests whether this model fits the data.

3.7.5 Consistency and PMSEs

The consistency of PMSEs is given by the event "$|\tilde{\triangle}(X|S, \mathcal{M}) - \triangle| > \epsilon$", $\epsilon > 0$. The probability of this event is given by

$$P_\theta(|\tilde{\triangle}(X|S, \mathcal{M}) - \triangle| > \epsilon) = \Sigma_{k=1}^{K} P_\theta(|\tilde{\triangle}(X|S, \mathcal{M}) - \triangle| > \epsilon|\tilde{\mathcal{X}_k}) P_\theta(\tilde{\mathcal{X}_k}). \quad (3.37)$$

The estimator $\tilde{\triangle}(X|S, \mathcal{M})$ is consistent if

$$\lim_{n\to\infty} P_\theta(|\tilde{\triangle}(X|S, \mathcal{M}) - \triangle| > \epsilon) \longrightarrow 0, \quad \forall \epsilon > 0 \quad (3.38)$$

and uniformly consistent if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} P_\theta(|\tilde{\triangle}(X|S, \mathcal{M}) - \triangle| > \epsilon) \longrightarrow 0, \quad \forall \epsilon > 0. \quad (3.39)$$

The point is that it is hard to check for consistency and uniform consistency of $\tilde{\triangle}$, as can be seen in Equation (6.7). Even, if for each model considered as true, estimators are consistent, there is no such guarantee that this holds for the PMSE.

3.8 Conditional analysis

A conditional analysis involves conditioning the analysis on the selected subset. For instance, suppose that the selected subset is $\mathcal{X}_{k^*}$. The estimator to study is the conditional estimator $\tilde{\triangle}|\mathcal{X}_{k^*}$. Such conditional analysis is studied by Miller (2002) and model selection uncertainty is defined with respect to this conditional estimator. From this point of view, for any event E that can appear in any subset $\mathcal{X}_k$, its probability is given by

$$P_\theta(E) = \Sigma_{k=1}^{K} P_\theta(E|\tilde{\mathcal{X}}_k) I_k(X|S, \mathcal{M}). \tag{3.40}$$

After the data have been observed, one computes $P_\theta(E|\tilde{\mathcal{X}}_{k^*})$, the conditional probability given subset $\tilde{\mathcal{X}}_{k^*}$. One may argue that after data have been observed, the relevant subset to consider is $\mathcal{X}_{k^*}$, but it is important to take into account the probability of selecting this particular subset, which is unlikely to be one. Equation (3.27) explains why although this approach is better than the naive approach, it is incomplete and does not cover all aspects of the problem. In fact, Equation (3.29) can be written as

$$P_\theta(E) = \underbrace{P_\theta(E|\tilde{\mathcal{X}}_{k^*})}_{(1)} \underbrace{P_\theta(\tilde{\mathcal{X}}_{k^*})}_{(2)} + \underbrace{\Sigma_{k=1,k\neq k^*}^{K} P_\theta(E|\tilde{\mathcal{X}}_k) P_\theta(\tilde{\mathcal{X}}_k)}_{(3)}. \tag{3.41}$$

Using a conditional analysis involves using only (1) and ignoring (2) and (3).

3.9 The use of consistent model selection criteria

3.9.1 Describing consistent criteria

Consider Equation (3.29) and suppose that there exists a subset such that the probability of selecting this subset is 1. Let k_0 denoting the index of that subset and assume that the true model is one of the competing models. If a consistent criterion is used to select this true model, we have then the following

$$P_\theta(E) = \underbrace{P_\theta(E|\tilde{\mathcal{X}}_{k_0}) P_\theta(\tilde{\mathcal{X}}_{k_0})}_{(1)} + \underbrace{\Sigma_{k=1,k\neq k_0}^{K} P_\theta(E|\tilde{\mathcal{X}}_k) P_\theta(\tilde{\mathcal{X}}_k)}_{(2)}. \tag{3.42}$$

Clearly, if $P_\theta(\tilde{\mathcal{X}}_k) = 1$, then (2) is equal to 0.
Let A and B be two events such that $P(B) = 1$ and $P(A|B) > 0$, then $P(A|B) = P(A)$. Then applying this to Equation (3.42), $P_\theta(\tilde{\mathcal{X}}_{k_0}) = 1$, $P_\theta(E|\tilde{\mathcal{X}}_{k_0}) P_\theta(\tilde{\mathcal{X}}_{k_0}) = P_\theta(E)$, assuming that $P_\theta(E|\tilde{\mathcal{X}}_{k_0}) > 0$. This means that inference can be based on the selected model M_0 since e.g., the distribution of PMSE and naive estimator will be the same.
This is in fact true if $P_\theta(\tilde{\mathcal{X}}_{k_0})$ does not depend on the true parameter (in general not true for every θ). For each subset $\mathcal{X}_k$, one can define a parameter space Θ_k for which the probability of landing on this subset is 1, namely

$$\Theta_k = \{\theta \in \Theta : P_\theta(\tilde{\mathcal{X}}_k) = P_\theta(\tilde{M}(X|S, \mathcal{M}) = M_k) = 1\}. \tag{3.43}$$

That the probability of selecting the true model is 1 for all values of the parameter space is in general difficult to achieve. E.g., suppose that the selection criterion is the pre-test with H_0 the null hypothesis and H_1 the alternative hypothesis. Suppose that the true model is the one under H_1. One would like the power of the test to be close to 1 for parameter under H_1 and close to 0 for parameter under H_0. One cannot expect the power of the test be 1 under H_0 and H_1. In practice, Θ_k can't be known (it can also be an empty set) and is unlikely to be Θ.

For consistent selection procedure, for every θ, $P_\theta(\tilde{\mathcal{X}}_{k_0}) \rightarrow 1$ as $n \longrightarrow \infty$. That is we have, for every θ, $P_{(\theta,n)}(\tilde{\mathcal{X}}_{k_0}) \rightarrow 1$ as $n \longrightarrow \infty$ (pointwise convergence).

From Equation (3.42), one can then use the asymptotic distribution of the selected (naive) model, M^*, namely

$$n^{1/2}(\tilde{\triangle}(X|S,\mathcal{M}) - \triangle) \longrightarrow_d G^*_\theta, \tag{3.44}$$

where G^*_θ is the asymptotic distribution the normalised naive estimator $n^{1/2}(\triangle^* - \triangle)$.

3.9.2 Asymptotic efficiency

A convenient approach to compare performance of estimators for large sample sizes is to study the normalised risks of these estimators. Let $X \sim f(x|\theta)$ and $\triangle = h(\theta)$, the quantity of interest (function of parameter θ). Let $\hat{\triangle}_1$ and $\hat{\triangle}_2$ be two estimators of $\triangle$ with respective risks R_{1n} and R_{2n}. One could compare their normalised risks $n^r R_{1n}$ and $n^r R_{2n}$, $r > 0$, by finding $\lim_{n\to\infty} n^r R_{in}$.

Such approaches are used in Lehmann (1983), Lehmann and Casella (1998, 2001). Let $\tilde{\triangle}(X|S,\mathcal{M})$ be a PMSE corresponding to a choice between the two estimators $\hat{\triangle}_1$ and $\hat{\triangle}_2$. Let $R_n(\tilde{\triangle}(X|S,\mathcal{M}),\triangle) = \mathrm{E}(n^d(\tilde{\triangle}(X|S,\mathcal{M}) - \triangle)^2$ the normalised risk of $\tilde{\triangle}(X|S,\mathcal{M})$, we choose $d = 1/2$. Suppose that

$$\lim_{n\to\infty} R_n(\tilde{\triangle}(X|S,\mathcal{M}),\triangle) = g(\theta), \quad \forall\theta \in \Theta.$$

The point is that there may exist a sequence of parameters θ_n, therefore $\triangle_n$ for which

$$\lim_{n\to\infty} R_n(\tilde{\triangle}(X|S,\mathcal{M}),\triangle_n) = \infty.$$

It follows that

$$\lim_{n\to\infty} \sup_{\Gamma} R_n(\tilde{\theta}(X|S,\mathcal{M}),\triangle) = \infty.$$

In this case, the convergence will not be uniform.

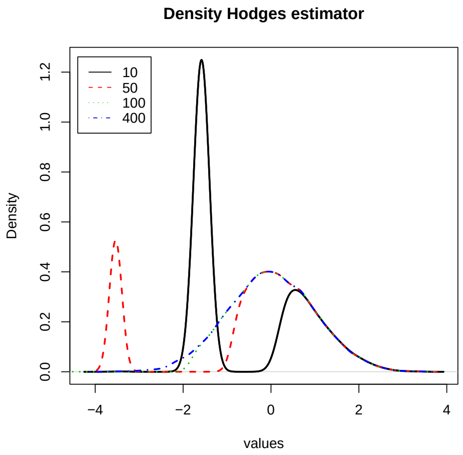

Figure 3.13: Densities for Hodges' estimator, $n^{1/2}(\tilde{\theta} - \theta)$ for $\theta = 0.5$ for various values of the sample size.

3.9.3 The Hodges' estimator example

Let $\hat{\theta}$ be the mean of a sample size n from $N(\theta, 1)$ distribution. For this example, $\triangle = \theta$. Define another estimator $\tilde{\theta}$ by

$$\tilde{\theta} = \begin{cases} \hat{\theta} & \text{if} \quad |\hat{\theta}| > n^{-1/4} \\ 0 & \text{otherwise.} \end{cases} \tag{3.45}$$

Note that Lehmann (1983), Lehmann(1999), Lehmann and Casella (1998, 2001) consider instead

$$\tilde{\theta} = \begin{cases} \hat{\theta} & \text{if} \quad |\hat{\theta}| > n^{-1/4} \\ a\hat{\theta} & \text{otherwise,} \end{cases}$$

where with $|a| < 1$. They were concerned to illustrate that Hodges' estimators were superefficient. These are estimators that are asymptotically normal, but whose asymptotic variance is not greater than the inverse of the amount of Fisher information at some points.
We use the simpler estimator (3.45), since the main point to be illustrated does not change.

The idea of this estimator is that if $\hat{\theta}$ is close to 0, then it is set to exactly 0, otherwise it does not change. The cutoff value $n^{-1/4}$ is chosen in such a way that its limit behaviour is the same as that of $\hat{\theta}$ for $\theta \neq 0$, but one tries to improve at $\theta = 0$. The estimator $\tilde{\theta}$ can be viewed as a PMSE. Figure (3.13) gives

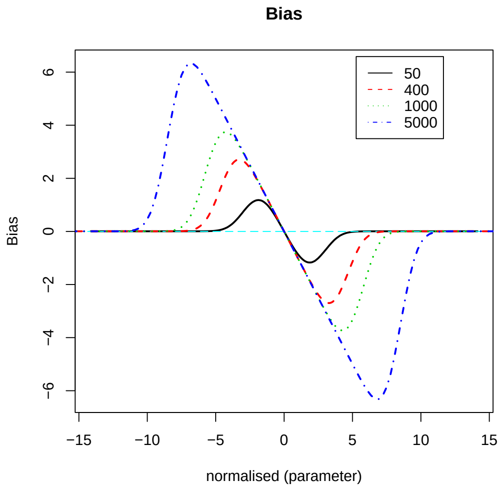

Figure 3.14: Sample size effects on bias for Hodges' estimator as a function of normalised θ, $n \in \{50$ (solid line), 400 (dashed line), 1000 (dotted line), 5000 (broken line)$\}$.

the density of $\tilde{\theta}$ for different values of n. It can be seen that its distribution is a mixture for small sample size, but as the sample increases, this distribution tends to be unimodal. It can be shown that $\lim_{n\to\infty} P(\tilde{\theta} = \hat{\theta}) = 1$ for $\theta \neq 0$, and $\lim_{n\to\infty} P(\tilde{\theta} = 0) = 1$ for $\theta = 0$, for example Lehmann (1999), Van der Vaart (2000).

Moreover, we have that

$$n^{1/2}(\tilde{\theta} - \theta) \longrightarrow_d N(0, g(\theta)), \tag{3.46}$$

where $g(\theta) = 1$ when $\theta \neq 0$ and $g(\theta) = 0$ for $\theta = 0$. At a first glance, $\tilde{\theta}$ seems to be an improvement over $\hat{\theta}$. Let look at its normalised risk. From Lehmann (1983), Lehmann and Casella (1998, 2001), $\lim_{n\to\infty} R_n(\tilde{\theta}, \theta) = 1$ for $\theta \neq 0$ and $\lim_{n\to\infty} R_n(\tilde{\theta}, 0) = 0$. Consider $\theta_n = n^{-1/4}$, then $\lim_{n\to\infty} R_n(\tilde{\theta}, \theta_n) = \infty$, therefore $\lim_{n\to\infty} \sup_\Theta R_n(\tilde{\theta}, \theta_n) = \infty$.

Figures (3.14) and (3.15) illustrate that the maximum normalised bias, hence MSE of $\tilde{\theta}$ increases without bound as the sample size increases. This is the reason why consistent selection criteria do not solve the model selection uncertainty problem.

Unlike the case for fixed θ, Figure (3.16) shows that for $\theta_n = n^{-1/4}$, no matter how large the sample size, the density of PMSE Hodges' estimator is bimodal. That is, for fixed true paramater θ, as illustrated in (3.13), the density is unimodal

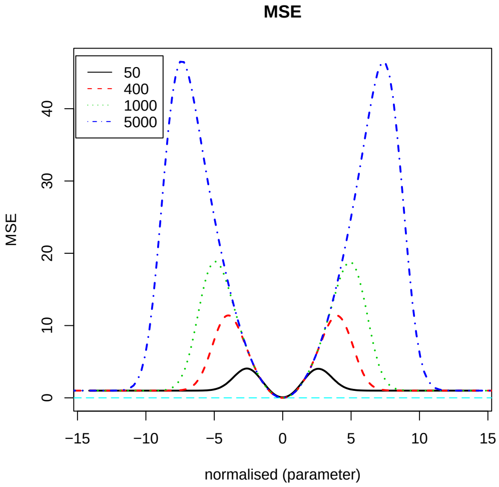

Figure 3.15: Sample size effects on MSE for Hodges' estimator as a function of normalised θ, $n \in \{50$ (solid line), 400 (dashed line), 1000 (dotted line), 5000 (broken line)$\}$.

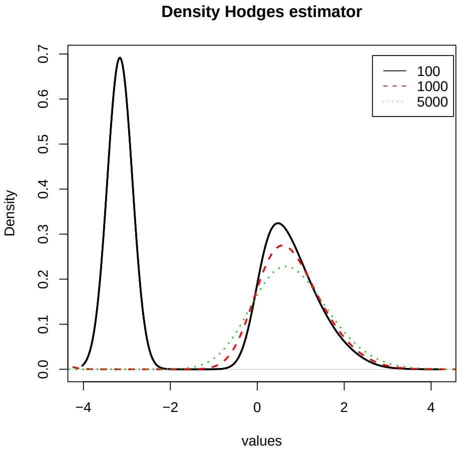

Figure 3.16: Densities for Hodges' estimator, $n^{1/2}(\tilde{\theta} - \theta_n)$ for $\theta_n = n^{-1/4}$ for various values of the sample size.

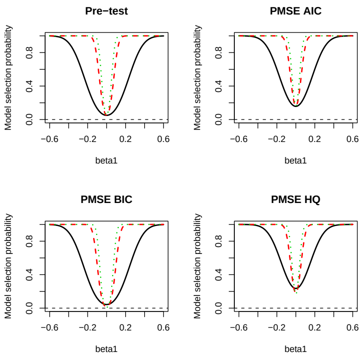

Figure 3.17: Sample size effects on model selection probabilities as a function of β_1. $n \in \{60$ (solid line), 500 (dashed line), 1000 (dotted line)$\}$.

as the sample size increases, but this is different when the true parameter depends on the sample size, e.g. $\theta_n = n^{-1/4}$. Figure (3.16) shows how the non-uniformity problem manifests for the density. Also, if θ_n is chosen in such a way that $\theta_n \to 0$, $n^{1/2}\theta_n \to \infty$ and $n^{1/4}\theta_n \to \infty$, then $n^{1/2}(\tilde{\theta} - \theta_n) \to -\infty$ as $n \to \infty$. This suggests that it is not enough to study the behaviour of $\tilde{\theta}$ pointwise (for every θ).

3.9.4 Linear regression

3.9.4.1 Model selection probabilities

Equations (3.17) and (3.18) suggest that one can see that the probability of selecting the full model increases with the sample size. However, this is not always the case. For example, in the case of multiple linear regression as illustrated in (3.18), this probability increases for small sample size, but tends to remain unchanged for large sample size. When these probabilities are scaled, from Figure (3.19), model selection probabilities decrease as the simple size increases for consistent criteria (BIC and HQ), that is, for all consistent criterion where the penalty h_n satisfy: h_n tends to infinity and h_n/n tends to 0 as n tends to infinity. These selection probabilities remain unchanged whatever the sample size for model selection criteria with fixed penalty (AIC, Pretest).

Consider the probability of selecting the true model at $b_1 = 0$, $P_n(0)$,

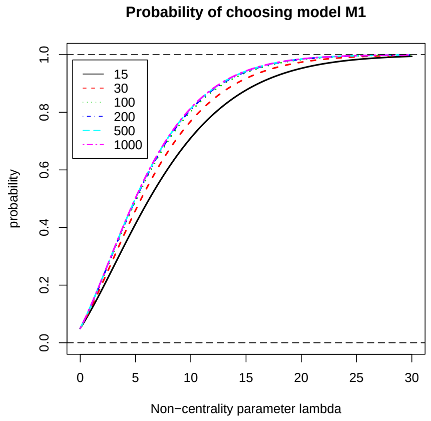

Figure 3.18: Multivariate regression: sample size effects on model selection probability for pre-test estimators as a function of δ with $\alpha = 0.05$.

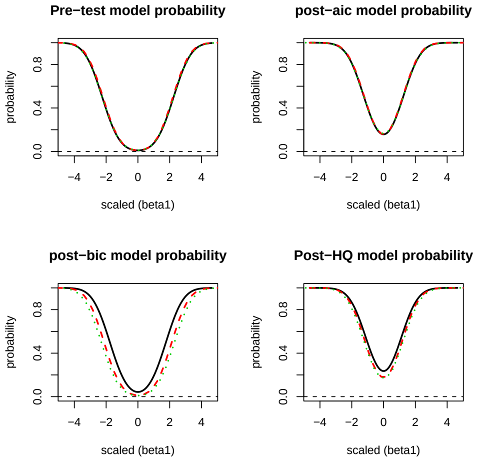

Figure 3.19: Sample size effects on model selection probabilities as a function of scaled β_1, $n = 60$ (solid line), $n = 500$ (dashed line), $n = 1000$ (dotted line).

$$P_n(0) = \begin{cases} \Phi(-z_{1-\alpha/2}) + 1 - \Phi(z_{1-\alpha/2}) & \text{for Hypothesis Testing} \\ \Phi(-2^{1/2}) + 1 - \Phi(2^{1/2}) & \text{for AIC} \\ \Phi(-h^{1/2}) + 1 - \Phi(h^{1/2}) & \text{for any constant penalty h} \\ \Phi(-h_n^{1/2}) + 1 - \Phi(h_n^{1/2}) & \text{for any penalty } h_n \text{ depending on } n \\ \log(n) & \text{for BIC} \\ \log(\log(n)) & \text{for HQ.} \end{cases}$$

Under $b_1 = 0$, the probability of selecting model the full model is $\Phi(-z_{1-\alpha/2}) + 1 - \Phi(z_{1-\alpha/2})$ for pre-test. For smaller values of α, for example 0.01, 0.005, this probability is nearly 0, that is, the probability of selecting the true model is nearly 1. This means that hypothesis testing can be considered as consistent for very smaller value of α. But the probability of selecting the full model is not 0 for reasonnable level of significance level, for e.g. 0.05.

For all criteria with fixed penalty (e.g., AIC), it is clear that the probability $P_n(0)$ does not converge to 0 (AIC is not a consistent selection criterion).

Consider criteria with penalty h_n, depending on n (e.g. BIC, HQ), then $P_n(0) = \Phi(-h_n^{1/2}) + 1 - \Phi(h_n^{1/2}) = 2(1 - \Phi(h_n^{1/2}))$. $P_n(0)$ gets smaller as n gets larger for a penalty that increases as n gets larger (e.g. BIC, HQ). For example, for BIC and HQ, $P_n(0) \longrightarrow 0$ as $n \longrightarrow \infty$. The reason is that these criteria are consistent. For consistent model selection criteria, the penalty h_n tends to infinity and h_n/n tends to 0 as n tends to infinity (Shao, 1997). These features are illustrated in Figure (3.17).

3.9.4.2 Moments

For scaled parameters, from Figure (3.20), one can see that the bias (also the maximum bias) increases with the sample size. The phenomenon is more pronounced with BIC and HQ. Variance (Figure 3.21) and Mean square error (Figure 3.22) increase with sample size when the absolute value of the scaled parameter is larger and decrease when it is smaller for BIC and HQ.
However, for model selection criteria with fixed penalty (Pre-test, AIC), the maximum bias, variance and MSE are bounded, see Figures 3.20, 3.21 and 3.22.

3.9.4.3 Densities

Of particular interest is the behaviour of consistent criteria like BIC. For any fixed value of β_1, there exists a sample size for which the density is unimodal, Figure (3.24) shows that for some values of β_1, e.g. $\beta_1 = n^{-2/5}$, there is no sample size for which the density is unimodal. This is due to non-uniformity.

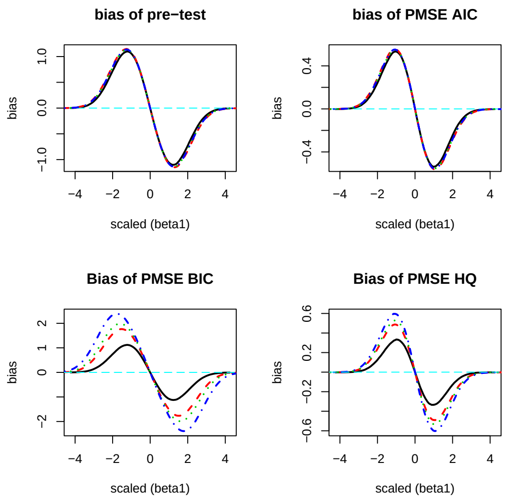

Figure 3.20: Sample size effects on bias for PMSEs as a function of scaled β_1 , $n = 50$ (solid line), $n = 400$ (dashed line), $n = 1000$ (dotted line), $n = 5000$ (broken line).

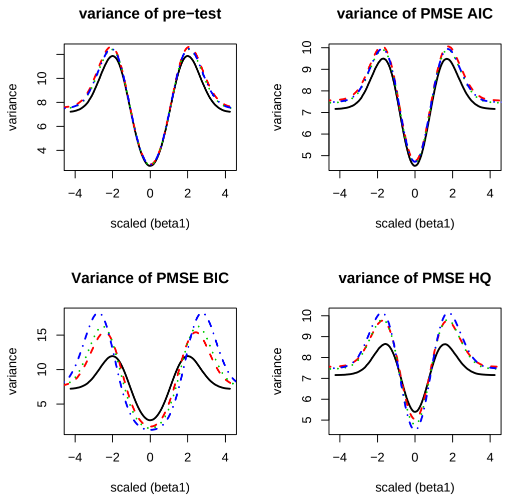

Figure 3.21: Sample size effects on variance for PMSEs as a function of scaled β_1, $n = 50$ (solid line), $n = 400$ (dashed line), $n = 1000$ (dotted line), $n = 5000$ (broken line).

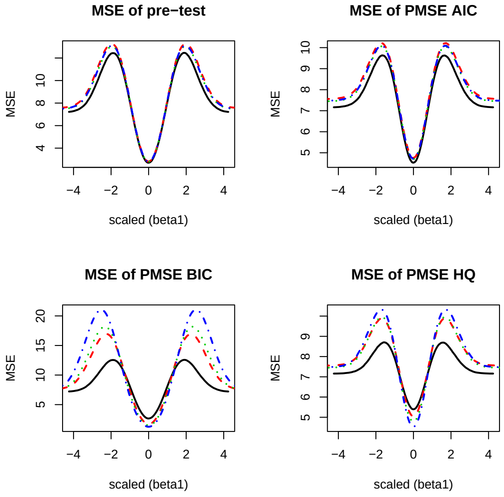

Figure 3.22: Sample size effects on MSE for PMSEs as a function of scaled β_1, $n = 50$ (solid line), $n = 400$ (dashed line), $n = 1000$ (dotted line), $n = 5000$ (broken line).

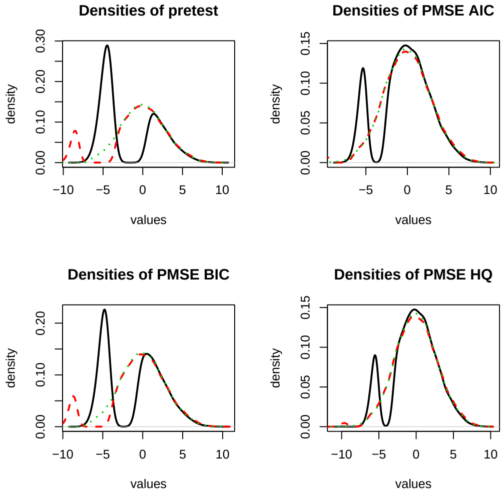

Figure 3.23: Densities for PMSEs for $\beta_1 = 0.2$, $\alpha = 0.01$ as a function of sample size: $n = 100$ (solid line), $n = 300$ (dashed line), $n = 500$ (dotted line).

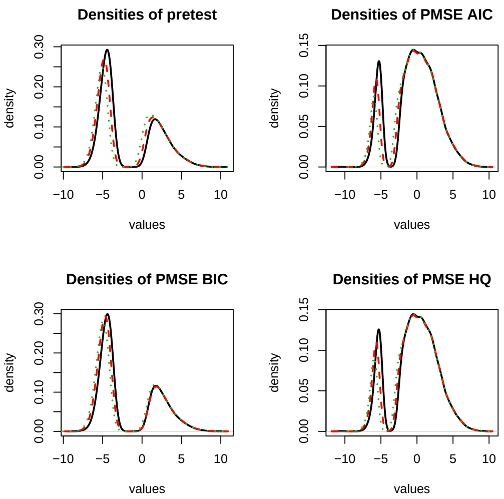

Figure 3.24: Scaled densities for PMSEs for $\beta_{1_n} = n^{-2/5}$ as a function of sample size: $n = 1000$ (solid line), $n = 2000$ (dashed line), $n = 3000$ (dotted line).

As can be seen in Figure (3.23), as $|\beta_1|$ decreases, the size required for unimodality of PMSEs increases. This means that asymptotic unimodality depends on the true parameter and the sample size. From Table (3.2), the minimum sample size, $n^*_{\beta 1}$ to achieve unimodality depends on the true parameter β_1. For $\beta_1 = 0.08$, this minimum is 2000 for post-AIC and post-Cp, and 3500 for pre-test and post-BIC. The reason is that for smaller values of $|\beta_1|$, the probability that post-BIC and pre-test select the M_1 is 0 whereas, it is not 0 for post-AIC and Cp. As one can see, this convergence is then not uniform for β_1, it is only pointwise (for fixed β_1), whether the selection criterion is consistent or not. In general, PMSEs are not unimodal uniformly for large n.

β_1	0.6	0.2	0.08
$n^*_{\beta 1}$	50	500	2000 or 3500

Table 3.2: Minimun sample size to achieve unimodality as function of β_1, $\alpha = 0.01$.

Chapter 4

Model Selection and Frequentist Model Averaging

4.1 Introduction

The problems that arise when one has more than a single model at one's disposal have now been considered from two points of view. In Chapter 2 we examined model averaging estimators, in which a weighted average of the models is used to estimate the quantity of interest. In Chapter 3 we focused on model selection, in which a single model is selected to estimate the quantity of interest. These two approaches are of course different. However, mathematically, any post-model-selection estimator is simply a special case of model averaging, the case in which all except one of the weights are set equal to zero. Furthermore these 0-1 weights, considered unconditionally, are, as in the model averaging case, random variables. *Before* we see the data it is uncertain which model will be selected, i.e. which of the binary 0-1 random variables (the weights) will equal 1. This point of view is exploited to demonstrate the fact that the two methodologies are closely related and mathematically comparable.

It has been suggested in the literature that model averaging outperforms PMS estimation. However, our analysis shows that, as long as the selection procedure is not taken into account in classical model averaging estimators, these are unlikely to outperform PMSEs. We propose a method to take selection into account in classical model averaging by making use of the key quantity in PMS estimation, namely the model selection probabilities. In effect we are proposing alternatives to the Akaike weights, which we call *adjusted Akaike weights* (AAW), and alternatives to likelihood weights, called *adjusted likelihood weights* (ALW). The adjusting factor is simply the model selection probability (Nguefack and

Zucchini, 2005).

The new averaging method is applied to the estimation of the mean of a multivariate normal distribution. It is shown that, under certain conditions, it is a minimax estimator, and that it can outperform Stein estimation. The method is also illustrated using a simple linear regression model, and an example on the analysis of proportions. In both examples it is clearly better than PMSEs.

4.2 Similarities and differences

We have defined model uncertainty as the fact that the true model is not known and model selection uncertainty as the fact that the model to be used should be selected from a set of candidate models. Let $\mathcal{M} = (M_1, \ldots, M_K)$ be a finite set of models, X the data and each model is parametric with parameter θ. Let $\triangle$ be the quantity of interest (a function of θ). Let θ_k represent the parameter under M_k and $\hat{\theta}_k$ be its estimator. Let $\triangle_k$ represent the quantity of interest under M_k and $\hat{\triangle}_k$ be its estimator. The difference between the two is that the latter includes a selection procedure, through the quantity $I_k(X|S, \mathcal{M})$ for each model M_k. In general, likelihoods are used as weights. Since it can only be used for models with the same dimension, different penalties have been tried to penalize the likelihood. We will explain that, as long as weighted model scheme does not take into account model selection procedure, the resulting weighted estimator is not garanteed to outperform post-model-selection estimators.

4.3 Combining model averaging and model selection

The PMSE is defined as

$$\tilde{\triangle}(X|S, \mathcal{M}) = \sum_{k=1}^{K} I_k(X|S, \mathcal{M})\hat{\triangle}_k, \tag{4.1}$$

where $I_k(X|S, \mathcal{M}) = 1$ if M_k is selected by S and 0 otherwise. This PMSE is then also a special model weighting (0-1 degenerate random weights); it depends on the selection procedure S and the entire set of models $\mathcal{M}$. An important ingredient of Equation (4.1) is the selection procedure S. Suppose that one is selecting between two models using hypothesis testing as the selection procedure. A fundamental quantity is the power of test. We can expect the properties of the estimator in (4.1) to be different if another selection procedure is used. A selection procedure of the form (3.15), or more generally a parsimonious selection procedure taking into account underfitting and overfitting, will result in different types of estimators (4.1). This suggests that taking into account the selection procedure could improve the estimators of the form (4.1). For instance, in the case of hypothesis testing, one may use the information concerning the power of the test. Since the likelihood of each model contains important information about that model, one could use this to weight competing models. In this case, the more complex model, in terms of number of parameters, will have high weights. The selection procedure S partitions the sample space and there is probability

$$P_\theta(M_k|S) = \mathrm{E}_\theta(I_k(X|S, \mathcal{M})) \tag{4.2}$$

that the subset $\mathcal{X}_k$ (model M_k) is selected. The selected subset is a random subset. A common feature for classical model averaging is that the selection procedure S is not taken into account. However it is necessary to include the selection procedure S in the model averaging estimator. We suggest to include the probability of selecting each model M_k into the weights, in particular in Akaike weights and simple likelihood weights. The point is how to get these model probabilities. As can be seen in (4.2), these are computed as an expectation and depend on the parameter θ. If a closed form exists, one can find an estimator of θ, and then obtain an estimator of these probabilities. In case there is no close form, Miller (2002) suggests a Monte Carlo method based on projection to get these probabilities. We denote by $p(M_k|S)$, an estimator or a Monte Carlo estimator of model selection probability for model M_k. The naive bootstrap estimates, where estimate selection probability of model M_k is estimated by the proportion of resample in which M_k is selected, does not work (Hjort and Claeskens, 2003). We propose to introduce the selection procedure into classical model averaging using Akaike weights and simple likelihood weights.

4.3.1 Adjusted Akaike weights

We define *Adjusted Akaike weights* as

$$W_{aa_k} = \frac{p(M_k|S)\exp(-s_k/2)L_k}{\Sigma_{i=1}^{K} p(M_i|S)\exp(-s_i/2)L_i)} = \frac{p(M_k|S)\exp(-\frac{AIC_k}{2})}{\Sigma_{i=1}^{K} p(M_i|S)\exp(-\frac{AIC_i}{2})}. \tag{4.3}$$

If S is of the form $I_k = -2\log L_k + s_k$, (for example, the AIC or BIC), then $p(M_k|S)$ already takes account of the penalty term s_k, and so the likelihood would be penalized twice, by $\exp(-s_k/2)$ and by $p(M_k|S)$. This "double penalty" would apply to **any** selection procedure S that itself penalizes the complexity of the model, e.g. the number of parameters. Therefore we recommend the use of adjusted Akaike weights only for models with the same dimension.

4.3.2 Adjusted likelihood weights

We define *Adjusted likelihood weights* as

$$W_{al_k} = \frac{p(M_k|S)L_k}{\Sigma_{i=1}^K p(M_i|S)L_i}. \tag{4.4}$$

The likelihood, which determines the "degree of fit" of the model, enters the weights. The selection probability $p(M_k|S)$ adjusts the weights for the selection procedure. Note that both of these components are required. To use only $p(M_k|S)$ would not fully account for the fit of the model. Using only the likelihood does not take into account the way in which the model is selected.

Relation between Akaike weights and adjusted likelihood weights

$$W_{al_k} = \frac{p(M_k|S)\exp(s_k/2)\exp(-\frac{AIC_k}{2})}{\Sigma_{i=1}^K p(M_i|S)\exp(s_i/2)\exp(-\frac{AIC_i}{2})}.$$

If models have the same dimension and probabilities of selection then the adjusted likelihood weights reduce to Akaike weights.

For finite samples, one can use the variance formulae proposed in Buckland et al. (1997); the first when estimates are perfectly correlated and the second when they are independent:

$$\begin{aligned} \mathrm{Var}(\hat{\triangle}_{\mathrm{MA}}(S)) &= \left\{\Sigma_{k=1}^K W_{al_k}\sqrt{\mathrm{Var}(\hat{\triangle}_k) + [\hat{\triangle}_k - \hat{\triangle}_{\mathrm{MA}}(S)]^2}\right\}^2 \quad \text{perfect correlation,} \\ \mathrm{Var}(\hat{\triangle}_{\mathrm{MA}}(S)) &= \Sigma_{k=1}^K W_{al_k}^2\{\mathrm{Var}(\hat{\triangle}_k) + [\hat{\triangle}_k - \hat{\triangle}_{\mathrm{MA}}(S)]^2\} \quad \text{independence.} \end{aligned} \tag{4.5}$$

where $\hat{\triangle}_{\mathrm{MA}}(S)$ is the weighted estimator and W_{al_k} the weight for model M_k.

For large samples, one can use the limiting risk properties and limiting distributions of general model weights as given in Hjort and Claeskens (2003). In this way, even if a model is more complex, in terms of the number of parameters, a "bad" model will be penalised by any reasonable selection procedure through the probability $p(M_k|S)$. If the model really fits the data, it will receive a higher weight. Here we let the selection procedure determine in how far a model is penalised.

4.4 Estimating a multivariate mean

4.4.1 Variance known

Suppose $X = (X_1, ..., X_p)' \sim N_p(\theta, \sigma^2 I_p)$, with unknown mean $\theta = (\theta_1, ..., \theta_p)'$, σ known. The quantity of interest is $\triangle = \theta$.

It is well known that for $p \leq 2$, one can just find the maximun likelihood estimator $\hat{\theta} = X$ and the risk is $R(\hat{\theta}) = \text{MSE}(\hat{\theta}) = p$. However, for $p \geq 3$, the maximum likelihood estimator is inadmissible and the problem is to find alternative estimators that yield small risk comparing to the MLE. Using the proposed weights, we propose an alternative approach for estimating θ. It will also assume that model selection probabilities are computed independently of the data, for example using the Monte Carlo technique of Miller (2002). Let $\Gamma(x|S, \mathcal{M}) = \Sigma_{k=1}^{K} p(M_k|S) L_k$.

Theorem 4.4.1 *Assume the following*

1. $X \sim N_p(\theta, \sigma^2 I_p), \sigma$ *known,*
2. $\hat{\theta}_k = X + \bigtriangledown \log L_k; \forall M_k$,*where* $\bigtriangledown \log L_k = (\partial \log L_k / \partial x_1, ..., \partial \log L_k / \partial x_p)$,
3. $\Gamma(x|S, \mathcal{M})$ *is almost differentiable (a.d) for which* $\bigtriangledown\Gamma$ *is also a.d ,*
   $\Gamma(x + z|S, \mathcal{M}) - \Gamma(x|S, \mathcal{M}) = \int\limits_0^1 z' \bigtriangledown \Gamma(x + tz|S, \mathcal{M})\, dt, \forall z \in Dom(x)$,
4. $\Gamma(x|S, \mathcal{M})$ *is superharmonic; that is* $\Sigma_{i=1}^{p} \partial^2 \Gamma(x|S, \mathcal{M}) / \partial x_i^2 \leq 0, \forall M_k$,
5. *Moment condition*
   (a) $\text{E}_\theta |\frac{\partial^2 \Gamma(X|S,\mathcal{M})/\partial x_i^2}{\Gamma(x|S,\mathcal{M})}| < \infty$,

   (b) $\text{E}_\theta \| \bigtriangledown \log \Gamma(X|S, \mathcal{M}) \|^2 < \infty$.

then $\hat{\theta}_{\text{MA}}(S) = \Sigma_{k=1}^{K} W_{al_k} \hat{\theta}_k$*is a minimax estimator for* θ *and its risk is*

$$\text{R}(\hat{\theta}_{\text{MA}}(S)) = \text{E}_\theta \| \hat{\theta}_{\text{MA}}(S)) - \theta \|^2 = p - 4\text{E}_\theta [- \frac{\bigtriangledown^2 \Gamma^{1/2}(x|S, \mathcal{M})}{\Gamma^{1/2}(x|S, \mathcal{M})}].$$

Proof. We start with the following lemma which is straightforward but important.

Lemma 4.4.1 *Under assumption (2),*

$$\hat{\theta}_{\text{MA}}(S) = X + \bigtriangledown \log \Gamma(X|S, \mathcal{M}). \tag{4.6}$$

Proof of Lemma 4.4.1. From assumption (2), each estimator is of the form $\hat{\theta}_k = X + \bigtriangledown \log L_k$.
Let denote $\bigtriangledown f = (\partial f / \partial x_1, \ldots, \partial f / \partial x_p)'$ and $\bigtriangledown Log f = \frac{\bigtriangledown f}{f}$, $W_{al_k} = \frac{p(M_k|S) L_k}{\Gamma(X|S,\mathcal{M})}$,

therefore $\hat{\theta}_{\text{MA}}(S) = \Sigma_{k=1}^{K} W_{al_k} \hat{\theta}_k$

$= \Sigma_{k=1}^{K} \left\{ \frac{p(M_k|S) L_k}{\Gamma(X|S,\mathcal{M})} \right\} \hat{\theta}_k$

$= \frac{1}{\Gamma(X|S,\mathcal{M})}\Sigma_{k=1}^{K}p(M_k|S)L_k[X + \bigtriangledown \log L_k]$

$= X + \frac{1}{\Gamma(X|S,\mathcal{M})}\Sigma_{k=1}^{K}p(M_k|S)L_k[\frac{\bigtriangledown L_k}{L_k}]$

$= X + \frac{1}{\Gamma(X|S,\mathcal{M})}\Sigma_{k=1}^{K}p(M_k|S) \bigtriangledown L_k$
$= X + \frac{1}{\Gamma(X|S,\mathcal{M})} \bigtriangledown (\Sigma_{k=1}^{K}p(M_k|S)L_k)$

$= X + \frac{1}{\Gamma(X|S,\mathcal{M})} \bigtriangledown \Gamma(X|S,\mathcal{M})$
$= X + \bigtriangledown Log\Gamma(X|S,\mathcal{M})$.
Using Lemma Lemma 4.4.1 and applying Stein's results (Stein, 1981, Corollary 1, p.1139), the result follows.

4.4.1.1 Improvement over James-Stein estimator

Efron and Morris (1971, 1972) propose to modify the James-Stein estimator given by $\hat{\theta}_0 = (1 - \frac{p-2}{\|X\|^2})_+X$. This modification was based on requiring that no coordinate of $\hat{\theta}_0$ be changed by more than a predetermined quantity d. This resulted in an improvement of $\hat{\theta}_0$ when the empirical distribution of $\mid \theta_i \mid$ is long tailed. We now also consider a modification of Efron and Morris based on order statistic.

Let $Y_i =\mid X_i \mid$ and the order statistics defined by $Y_{(1)} < ... < Y_{(p)}$. Let j be a large fraction of p. Suppose also that the coordinates of $h(X|S,\mathcal{M}) = \bigtriangledown \log\Gamma(X|S,\mathcal{M})$ are defined as

$$h_i(X|S,\mathcal{M}) = \begin{cases} c[\Sigma_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}X_i & \text{if } Y_i \leq Y_{(j)} \\ c[\Sigma_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}Y_{(j)}sgnX_i & \text{otherwise,} \end{cases} \tag{4.7}$$

where c is a constant and the optimum choice of c is determined to be $(j-2)$ and the risk of $\hat{\theta}_{\text{MA}}(S)^{(j)}$ is

$$\text{E}_\theta\|\hat{\theta}_{\text{MA}}(S)^{(j)} - \theta\|^2 = p - (j-2)^2\text{E}_\theta[\Sigma_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}.$$

We assume that p is large and $z = \frac{j}{p}$, so that j is closed to p (Stein, 1981, p.1146). We also assume that $\hat{\theta}_{\text{MA}_i}(S)^{(j)}$ are independently normally distributed with variance δ^2, let $w = \Phi^{-1}(0.5(1+z))$.
The estimated improvement risk for $\hat{\theta}_{MA}(S)^{(j)}$ and $\hat{\theta}_0$ over the MLE $\hat{\theta}$ are $\text{Imp}(\hat{\theta}_{\text{MA}}(S)^{(j)}, \hat{\theta})$, $\text{Imp}(\hat{\theta}_0, \hat{\theta})$ defined by

$$\text{Imp}(\hat{\theta}_{\text{MA}}(S)^{(j)}, \hat{\theta}) = \text{Risk}(\hat{\theta}) - \text{Risk}(\hat{\theta}_{\text{MA}}(S)^{(j)}) = (j-2)^2[\Sigma_{l=1}^{p}(min(X_l^2, Y_{(j)}^2))]^{-1},$$

$$\text{Imp}(\hat{\theta}_0, \hat{\theta}) = \text{Risk}(\hat{\theta}) - \text{Risk}(\hat{\theta}_0) = (p-2)^2[\Sigma_{l=1}^{p}X_l^2]^{-1}.$$

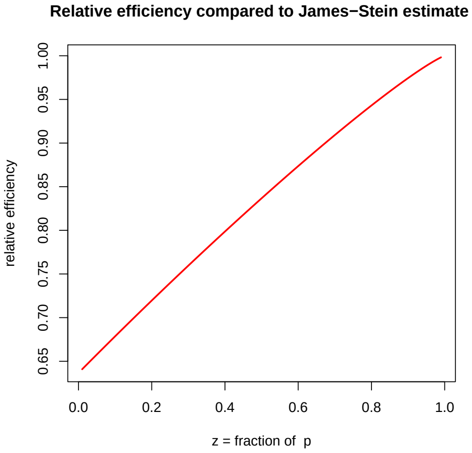

Figure 4.1: Relative efficiency of $\hat{\theta}_{\mathrm{MA}}(S)^{(j)}$ compared to James-Stein estimate $\hat{\theta}_0$ as a function of the proportion of the dimension of the parameter θ.

The relative efficiency of $\hat{\theta}_{\mathrm{MA}}(S)^{(j)}$ compared to the James- Stein estimate $\hat{\theta}_0$ is defined by

$$\mathrm{eff}(z) = \frac{\mathrm{Imp}(\hat{\theta}_0, \hat{\theta})}{\mathrm{Imp}(\hat{\theta}_{\mathrm{MA}}(S)^{(j)}, \hat{\theta})} = \frac{z^2}{(1-z)w^2 - 2w\phi(w) + z}, \tag{4.8}$$

where ϕ and Φ are the density and distribution functions of standard normal. From Equation (4.8), it follows that $\mathrm{eff}(z) < 1, \forall z \in (0,1)$,
$\Rightarrow \mathrm{Risk}(\hat{\theta}_{\mathrm{MA}}(S)^{(j)}) < \mathrm{Risk}(\hat{\theta}_0)$. This means that the modified version of the weighted estimator given in (4.7) is better than Stein estimator, for any proportion z of the data. An illustration is given in Figure (4.1). One can see that this relative efficiency is an increasing function of z. This means that as the proportion of data increases, the relative efficiency also increases.

4.4.1.2 Confidence sets for the mean

Here we illustrate how to obtain an approximate confidence sets for the true parameter θ. We start with the following theorem.

Theorem 4.4.2 *Suppose that* $h(X|S,\mathcal{M}) = \bigtriangledown \log \Gamma(X|S,\mathcal{M})$ *is twice continuously differentiable and such that*

$$\begin{aligned} &\mathrm{E}_\theta[\|h(X|S,\mathcal{M})\|^2 + \Sigma_{j=1}^{p}\Sigma_{i=1}^{p}(\partial h_i(X|S,\mathcal{M})/\partial X_j)^2 \\ &\quad + \Sigma_{j=1}^{p}\Sigma_{i=1}^{p}(\partial^2 h_i(X|S,\mathcal{M})/\partial X_i \partial X_j)^2] < \infty, \end{aligned}$$

then

$$\begin{aligned}&\mathrm{E}_\theta[\|\hat{\theta}_{\mathrm{MA}}(S)-\theta\|^2-(p+\|h(X|S,\mathcal{M})\|^2+2\bigtriangledown' h(X|S,\mathcal{M})]\\ &=2p+4\mathrm{E}_\theta[\|h(X|S,\mathcal{M})\|^2+2\bigtriangledown' h(X|S,\mathcal{M}))+tr^2(\bigtriangledown h'(X|S,\mathcal{M})].\end{aligned}$$

Proof. Follows directly from Theorem 3, Stein (1981), p.1149.

Let $V=2p+4[\|h(X|S,\mathcal{M})\|^2+2\bigtriangledown' h(X|S,\mathcal{M})+tr^2(\bigtriangledown h'(X|S,\mathcal{M})]$ and
$U=p+\|h(X|S,\mathcal{M})\|^2+2\bigtriangledown' h(X|S,\mathcal{M})]$
For p large, a confidence set with $1-\alpha$ approximate probability of covering θ can be

$$CS(X|S,\mathcal{M})=\{\theta:\|\hat{\theta}_{MA}(S)-\theta\|^2<U(X|S,\mathcal{M})+Z_{1-\alpha}\sqrt{V}(X|S,\mathcal{M})\},$$

where $Z_{1-\alpha}$ is the $1-\alpha$ quantile of standard normal.

4.4.2 Variance unknown

Consider the most realistic case σ^2 unknown. Let $\hat{\sigma}^2$ be an independent estimate of σ^2, such that $\hat{\sigma}^2\sim\sigma^2\chi^2_m$, chi-square with m degree of freedom. Consider the following estimator of θ

$$\theta^*(X|S,\mathcal{M})=\mathbf{X}+\frac{\hat{\sigma}^2}{m+2}\bigtriangledown\log\Gamma(X|S,\mathcal{M}).$$

The risk of $\theta^*(X|S,\mathcal{M})$ is $\mathrm{E}_{\theta,\sigma}\|\theta^*(X|S,\mathcal{M})-\theta\|^2$ and is given by

$$\mathrm{E}_{\theta,\sigma}\{p\frac{\hat{\sigma}}{m}+\frac{\hat{\sigma}^2}{(m+2)^2}[\|\bigtriangledown\log\Gamma(X|S,\mathcal{M})\|^2+2\bigtriangledown.\bigtriangledown log\Gamma(X|S,\mathcal{M})]\}.$$

It can be observed that we lose the proportion $\frac{2}{(m+2)}$ of the reduction of the risk that would have been achieved if σ^2 were known.
The proof is based on Stein (1981).

4.5 Illustrative examples

4.5.1 A simple linear regression

Consider a simple linear regression model discussed in Chapter 3.

$$Y_i=\beta_0+\beta_1x_i+\epsilon_i, i=1,\dots,n, \tag{4.9}$$

where the $\epsilon_i\sim N(0,\sigma^2)$, σ known (for simplicity), but similar results are obtained for the case where σ is unknown. Let x_+ be a future value of the covariate. The

aim is to estimate the mean $\triangle = E(Y|x_+)$.
Consider two models
$M_0 : \triangle = \beta_0$ and
$M_1 : \triangle = \beta_0 + \beta_1 x_+$.
One method of selecting between the 2 models is testing:
$H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 \neq 0$.
This means that the model selection method here is the pre-test. The PMSE of $\triangle$ is then given by

$$\tilde{\triangle} = \hat{\beta}_0 I_0(\frac{|\hat{\beta}_1|}{v_1^{1/2}} < z_{1-\frac{\alpha}{2}}) + (\hat{\beta}_0 + \hat{\beta}_1 x_+) I_1(\frac{|\hat{\beta}_1|}{v_1^{1/2}} \geq z_{1-\frac{\alpha}{2}}), \tag{4.10}$$

where I_0 and I_1 are indicator functions under M_0 and M_1, respectively, with $I_0 + I_1 = 1$, v_1=Var$(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma_{i=1}^n (x_i - \overline{x})^2}$, $\overline{x} = \frac{1}{n}\Sigma_{i=1}^n x_i$ and $z_{1-\frac{\alpha}{2}}$ is the quantile of standard normal. Now any PMSE can be written in the form of Equation (4.10). Here we will compare the moments of PMSEs and that of AIC model weights and the adjusted estimator defined above. Let $Z_0 = \frac{\hat{\beta}_0 - \beta_0}{v_0^{1/2}}$, then $Z_0 \sim N(0,1)$, then $\hat{\beta}_0 = v_0^{1/2}(Z_0 + b_0)$, where $b_0 = \frac{\beta_0}{v_0^{1/2}}$ (standardized intercept), $v_0 = \frac{\sigma^2}{n}$.
Let $Z_1 = \frac{\hat{\beta}_1 - \beta_1}{v_1^{1/2}}$, then $Z_1 \sim N(0,1)$, $\hat{\beta}_1 = v_1^{1/2}(Z_1 + b_1)$, where $b_1 = \frac{\beta_1}{v_1^{1/2}}$ (standardized slope).
From Chapter 3, the pre-test estimators and more general PMSEs are given by

$$\tilde{\triangle}_{\text{pretest}} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1) I_1(|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}}), \tag{4.11}$$

where Z_0 and Z_1 are independent normal since $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent.
For model selection criterion of the form

$$IC_k = -2 \log L_k + h_n p_k. \tag{4.12}$$

The PMSEs (e.g. AIC, BIC) are given by

$$\tilde{\triangle}_{IC} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1) I_1(|Z_1 + b_1| \geq h_n^{1/2}), \tag{4.13}$$

and for Mallows Cp, PMSE is given by

$$\tilde{\triangle}_{Cp} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1) I_1(F(1, n-2, b_1^2) > 2). \tag{4.14}$$

The properties of PMSEs were given in Chapter 3. For any weight scheme w_0 and w_1 for model M_0 and model M_1, $w_1 + w_2 = 1$, the model averaging estimator is

$$\hat{\triangle}_{\text{MA}} = w_0 \hat{\triangle}_0 + w_1 \hat{\triangle}_1 = w_0 \hat{\beta}_0 + w_1(\hat{\beta}_0 + x_+ \hat{\beta}_1) = \hat{\triangle}_0 + x_+ w_1 \hat{\beta}_1. \tag{4.15}$$

Replacing estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by standard normal yields

$$\hat{\triangle}_{\text{MA}} = v_0^{1/2}(Z_0 + b_0) + x_+ v_1^{1/2}(Z_1 + b_1)w_1. \tag{4.16}$$

We shall then use only w_1 for any model averaging scheme we will consider. We have derived in Chapter 3 the following

$$\frac{L_1}{L_0} = e^{\frac{1}{2}(Z_1+b_1)^2}, \tag{4.17}$$

where L_k is the likelihood function under model M_k. We have closed form model selection probabilities. For the simple linear model, the probability of selecting M_1 is given by

$p(M_1|\text{pre-test}, \mathcal{M}) = \Phi(\hat{r}) + 1 - \Phi(\hat{q})$, estimated power of the test.

For information criteria of the form (3.15), the probability of selecting M_1 is

$p(M_1|IC_n, \mathcal{M}) = \Phi(\hat{r}_n) + 1 - \Phi(\hat{q}_n)$, and
$p(M_1|Cp, \mathcal{M}) = 1 - P(F(1, n-2, \hat{b}_1^2) < 2)$ for the Cp criterion.

4.5.1.1 Likelihood weights

The likelihood weights are defined by $W_{sl_1} = \frac{L_1}{L_0+L_1} = \frac{\frac{L_1}{L_0}}{1+\frac{L_1}{L_0}}$. From (4.17) the simple likelihood weights are given by

$$W_{sl_1} = \frac{e^{\frac{1}{2}(Z_1+b_1)^2}}{1 + e^{\frac{1}{2}(Z_1+b_1)^2}}. \tag{4.18}$$

4.5.1.2 Akaike weights

We have $AIC_1 - AIC_0 = 2(\log L_0 - \log L_1) + 2$ and $W_{a_1} = \frac{e^{(AIC_0 - AIC_1)}}{1+e^{(AIC_0-AIC_1)}}$. Akaike weights are then given by

$$W_{a_1} = \frac{e^{\frac{1}{2}(Z_1+b_1)^2-1}}{1 + e^{\frac{1}{2}(Z_1+b_1)^2-1}}. \tag{4.19}$$

We do not illustrate for HQ since the same remarks apply as for the case of BIC.

Figure (4.2) displays post-AIC, corrected by Akaike and likelihood weights. One can see that model averaging using these classical weights methods does not improve on post-HQ. The same is valid for other information criteria, AIC, BIC, Cp and Pre-test. Figure (4.3) shows that no weight (simple likelihood or Akaike weight) dominates PMSE in terms of risk and variance functions. The same is

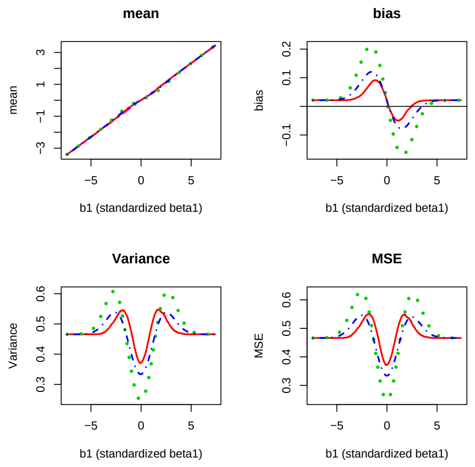

Figure 4.2: Properties of post-HQ (solid line), corrected by Akaike weight (dotted line), likelihood weight (dashed line) as a function of b_1.

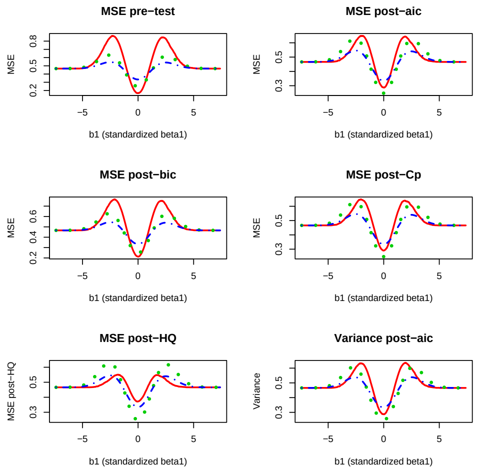

Figure 4.3: MSE and variance of PMSEs (solid line), corrected by Akaike weight (dotted line), likelihood weight (dashed line) as a function of b_1.

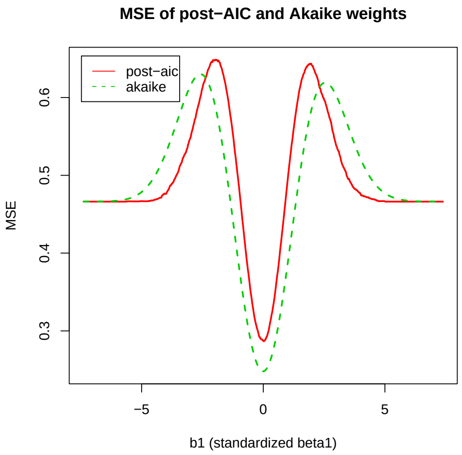

Figure 4.4: MSE of post-AIC (solid line), corrected by Akaike weight (dashed line) as a function of b_1.

valid for the variance.

Figure (4.4) illustrates the typical behaviour of Akaike weights and post-AIC. It can be seen that Akaike weigthing does not perform better than post-AIC.

4.5.1.3 Adjusted Akaike weights

The adjusted Akaike weight is given by

$$W_{aa_k} = \frac{\delta_1(M_1|S)e^{\frac{1}{2}(Z_1+b_1)^2-1}}{1+\delta_1(M_1|S)e^{\frac{1}{2}(Z_1+b_1)^2-1}}, \tag{4.20}$$

where $\delta_1(M_1|S) = \frac{p(M_1|S)}{p(M_0|S)} = \frac{p(M_1|S)}{1-p(M_1|S)}$.

Figure (4.5) shows that, in terms of risks, adjusted Akaike weights are better than PMSEs. However, in terms of bias, adjusted Akaike weights do not improve on PMSEs. The reason is that the models are not of the same dimension. The bias is then due to the double penalty.

4.5.1.4 Adjusted likelihood weights

For any selection procedure S, with selection probability $p(M_k|S)$ for model M_k, the adjusted likelihood is

$$W_{al_1} = \frac{p(M_1|S)L_1}{p(M_0|S)L_0 + p(M_1|S)L_1} = \frac{\frac{p(M_1|S)L_1}{p(M_0|S)L_0}}{1+\frac{p(M_1|S)L_1}{p(M_0|S)L_0}}. \tag{4.21}$$

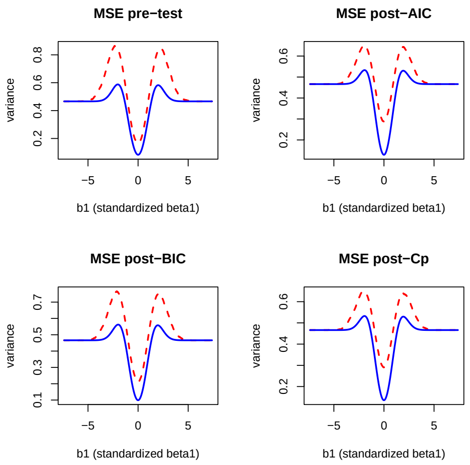

Figure 4.5: MSE of PMSEs (solid line), corrected by adjusted Akaike weight (dashed line) as a function of b_1.

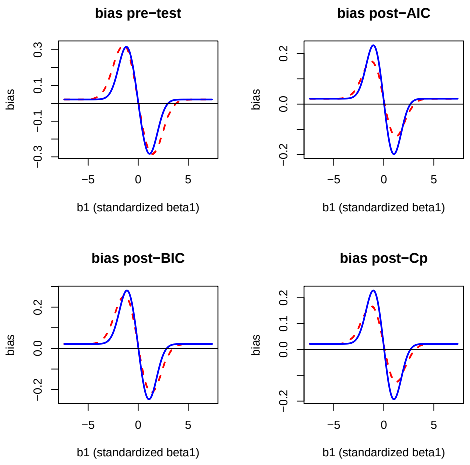

Figure 4.6: Bias of PMSEs (solid line), corrected by adjusted Akaike weight (dashed line) as a function of b_1.

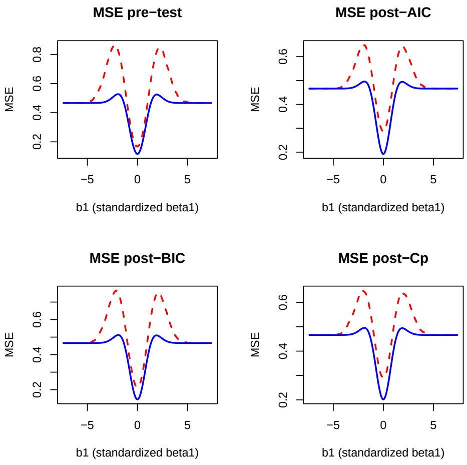

Figure 4.7: MSE of PMSEs (solid line), corrected by adjusted likelihood weight (dashed line) as a function of b_1.

The adjusted likelihood weights can be re-written as

$$W_{al_k} = \frac{\delta_1(M_1|S)e^{\frac{1}{2}(Z_1+b_1)^2}}{1+\delta_1(M_1|S)e^{\frac{1}{2}(Z_1+b_1)^2}}. \tag{4.22}$$

Figures (4.7), (4.8) and (4.9) show that in terms of risk, variance and bias, adjusted likelihood weights estimators improve on PMSEs. The same applies when one uses HQ as selection criterion.

Figure (4.10) displays the bias due to using only the probability of selecting each model as weights. One can see that, in general, this has more bias than PMSEs, due to not taking into account the likelihood of each model.

4.5.2 Estimation of proportions

Let $X_1, \ldots, X_n$ be n independent Bernouilli trials, that is $X_i \sim Be(\theta)$, $\quad Y = \sum_{i=1}^n X_i$ is the number of successes. Y is a binomial(n, θ), θ unknown. We will base inference on Y, since the likelihood function of the X_i's is $\theta^Y(1-\theta)^{n-Y}$ and involves the sufficient statistic Y.
$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y = 0, 1, \ldots, n$, is the probability mass function (PMF) of Y. Our quantity of interest is the unknown $\triangle = \theta$.

4.5.2.1 A two-model selection problem

(a) Consider the choice between the 2 models: $M_1 : \theta = \theta_1$ and $M_2 : \theta = \theta_2$. The true model may or may not belong to these 2 models.

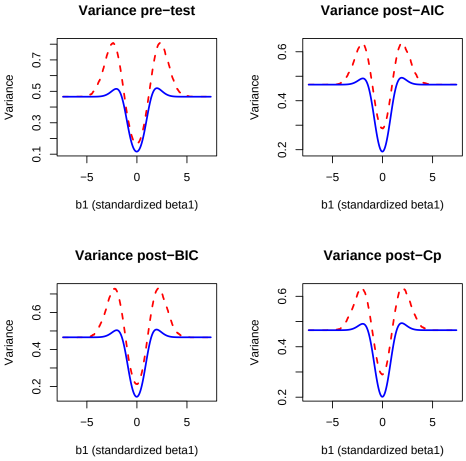

Figure 4.8: Variance of PMSEs (solid line), corrected by adjusted likelihood weight (dashed line) as a function of b_1.

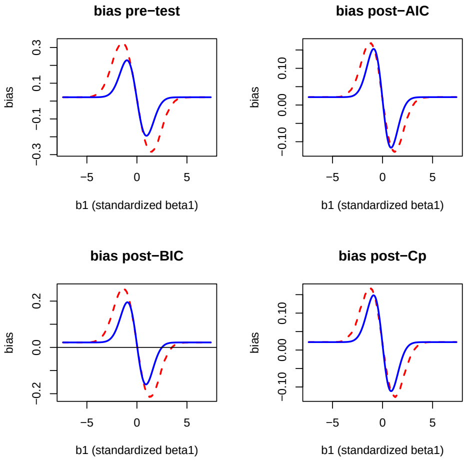

Figure 4.9: Bias of PMSEs (solid line), corrected by adjusted likelihood weight (dashed line) as a function of b_1.

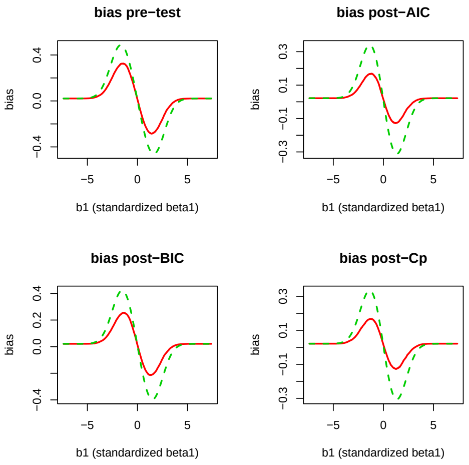

Figure 4.10: Bias of PMSEs (solid line), corrected only by the probability of selecting each model (dashed line) as a function of b_1.

Suppose that the selection procedure chooses the model with smaller AIC. In this case, this reduces to choosing the model with higher likelihood, since there is no parameter to be estimated for each model.
M_1 will be chosen if $f(y|\theta_1) > f(y|\theta_2)$ or equivalently if $R = \log(f(y|\theta_1)) - \log(f(y|\theta_2)) > 0$.
$R = \log(\binom{n}{y}\theta_1^y(1-\theta_1)^{n-y}) - \log(\binom{n}{y}\theta_2^y(1-\theta_2)^{n-y})$

$= \log\binom{n}{y} + y\log\theta_1 + (n-y)\log(1-\theta_1) - \log(\binom{n}{y}) - y\log\theta_2 - (n-y)\log(1-\theta_2)$

$= y\log\frac{\theta_1}{\theta_2} + (n-y)\log[\frac{1-\theta_1}{1-\theta_2}] = y[\log\frac{\theta_1}{\theta_2} - Log[\frac{1-\theta_1}{1-\theta_2}]] + n\log[\frac{1-\theta_1}{1-\theta_2}]$

$R > 0 \Longleftrightarrow y > \frac{-n\log[\frac{1-\theta_1}{1-\theta_2}]}{\log[\frac{\theta_1(1-\theta_2)}{\theta_2(1-\theta_1)}]} = a_n(\theta_1,\theta_2).$

Let $P_\theta(M_1|AIC,\mathcal{M}$ and $P_\theta(M_2|AIC,\mathcal{M} = 1 - P_\theta(M_1|AIC,\mathcal{M})$ be the probabilities of choosing models 1 and 2, respectively.

$$P_\theta(M_1|AIC,\mathcal{M}) = P_\theta(Y > a_n(\theta_1,\theta_2)) = 1 - P_\theta(Y \le a_n(\theta_1,\theta_2)) = 1 - F_{B(n,\theta)}(a_n(\theta_1,\theta_2)),$$

where $F_{B(n,\theta)}$ is the cumulative distribution function of binomial(n,θ).
The estimated probabilities are given by $p(M_1|AIC) = 1 - F_{B(n,\hat\theta)}(a_n(\theta_1,\theta_2))$, where $\hat\theta = y/n$ and $p(M_1|AIC) = 1 - p(M_1|AIC)$.
The PMSE $\tilde\theta = \theta_1$ if $y > a_n(\theta_1,\theta_2)$ and θ_2 otherwise.
The properties of $\tilde\theta$ are given by

$\mathrm{E}_\theta(\tilde{\theta}) = \sum_{y>a_n(\theta_1,\theta_2)} \theta_1 f(y|\theta) + \sum_{y\le a_n(\theta_1,\theta_2)} \theta_2 f(y|\theta)$

$= \theta_1 \sum_{y>a_n(\theta_1,\theta_2)} f(y|\theta) + \theta_2 \sum_{y\le a_n(\theta_1,\theta_2)} f(y|\theta) = \theta_1 p_1 + \theta_2 p_2.$

$\mathrm{Var}_\theta(\tilde{\theta}) = \sum_{y>a_n(\theta_1,\theta_2)} (\theta_1 - \mathrm{E}_\theta(\tilde{\theta}))^2 f(y|\theta) + \sum_{y\le a_n(\theta_1,\theta_2)} (\theta_2 - \mathrm{E}_\theta(\tilde{\theta}))^2 f(y|\theta)$

$= \mathrm{E}_\theta(\tilde{\theta})^2 - \mathrm{E}^2_\theta(\tilde{\theta}) = \theta_1^2 p_1 + \theta_2^2 p_2 - (\theta_1 p_1 + \theta_2 p_2)^2.$

$\mathrm{Bias}_\theta(\tilde{\theta}) = \mathrm{E}_\theta(\tilde{\theta}) - \theta.$

$\mathrm{MSE}_\theta(\tilde{\theta}) = \mathrm{Var}_\theta(\tilde{\theta}) + \mathrm{Bias}^2_\theta(\tilde{\theta}).$

The Akaike weights are defined by

$W_{a_1} = \frac{f(y|\theta_1)}{f(y|\theta_1)+f(y|\theta_2)},\ W_{aka_2} = \frac{f(y|\theta_2)}{f(y|\theta_1)+f(y|\theta_2)}.$

The adjusted likelihood weights are defined by

$W_{al_1} = \frac{p(M_1|AIC)f(y|\theta_1)}{p(M_1|AIC)f(y|\theta_1)+p(M_2|AIC)f(y|\theta_2)},\ W_{al_2} = \frac{p(M_2|AIC)f(y|\theta_2)}{p(M_1|AIC)f(y|\theta_1)+p(M_2|AIC)f(y|\theta_2)}\ .$

The weighted estimators are

$\hat{\theta}_{\mathrm{a}} = \theta_1 W_{a_1} + \theta_2 W_{a_2}\ .$

$\hat{\theta}_{\mathrm{al}} = \theta_1 W_{al_1} + \theta_2 W_{al_2}\ .$

$\mathrm{MSE}_\theta(\hat{\theta}_{\mathrm{a}}) = \Sigma_{y=0}^{n} (\hat{\theta}_{\mathrm{a}} - \theta)^2 f(y|\theta)\ .$

$\mathrm{MSE}_\theta(\hat{\theta}_{\mathrm{al}}) = \Sigma_{y=0}^{n} (\hat{\theta}_{\mathrm{al}} - \theta)^2 f(y|\theta)\ .$

Illustrating pictures correspond to $n = 41$, $\theta_1 = 0.6$ and $\theta_2 = 0.4$.
Figure (4.11) concerns model selection probabilities for $\theta_1 = 0.6$ and $\theta_2 = 0.4$ for the range of parameter space.

Figure (4.12) compares PMSE to that using AIC weights and adjusted weights using true model selection probabilities. It can be seen that adjusted likelihood is always better than PMSE and Akaike weights estimators. However, for some values of the true parameter, the risk of Akaike weight tends to be slightly bigger than that of PMSEs..

(b) Consider now a choice between the following two models:
$M_1 : Y \sim \text{binomial}(\theta_1, n)$ and $M_2 : Y \sim \text{binomial}(\theta, n)$.
AIC is used to select a model, $\hat{\theta}_2 = y/n$, for illustration, we choose $\theta_1 = 0.5$.
$AIC_{M_1} = -2\log(f(y|\theta_1)), \quad AIC_{M_2} = -2\log(f(y|\hat{\theta}_2)) + 2.$

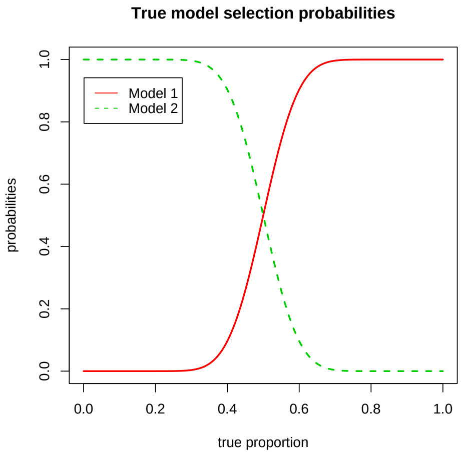

Figure 4.11: Model selection probabilities as a function θ, $\theta_1 = 0.6$ and $\theta_2 = 0.4$.

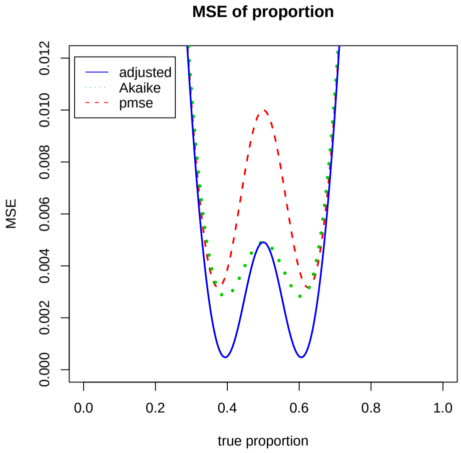

Figure 4.12: Risk of two simple proportions comparing PMSEs, Akaike weights estimators and adjusted estimators as a function of θ.

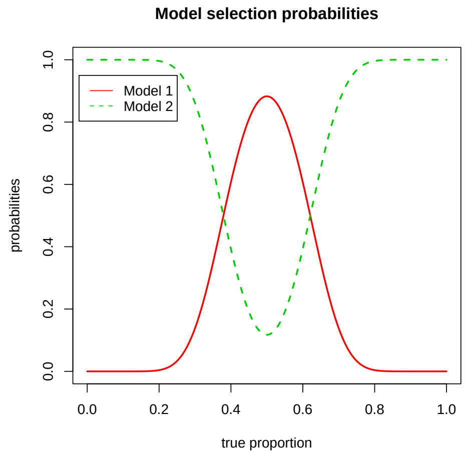

Figure 4.13: Model selection probabilities as a function θ.

Model 1 is chosen if
$AIC_{M_1} > AIC_{M_2}$, $P(M_1|AIC, \mathcal{M}) = P_\theta(AIC_{M_1} > AIC_{M_2})$,
$P(M_2|AIC, \mathcal{M}) = P_\theta(AIC_{M_1} \leq AIC_{M_2})$.
$p(M_1|AIC)$ and $p(M_2|AIC)$ are obtained by replacing θ by $\hat{\theta}_2 = y/n$.
The PMSE $\tilde{\theta} = \theta_1$ if $AIC_{M_1} > AIC_{M_2}$ and $\hat{\theta}_2$ otherwise.

$$\text{MSE}_\theta(\tilde{\theta}) = \sum_{AIC_{M_1} > AIC_{M_2}} \theta_1 f(y|\theta) + \sum_{AIC_{M_1} \leq AIC_{M_2}} \hat{\theta}_2 f(y|\theta).$$

The Akaike weights are defined by
$W_{a_1} = \frac{f(y|\theta_1)}{f(y|\theta_1)+f(y|\hat{\theta}_2)}$, $W_{a_2} = \frac{f(y|\hat{\theta}_2)}{f(y|\theta_1)+f(y|\hat{\theta}_2)}$

and the adjusted weights is defined by
$W_{al_1} = \frac{p(M_1|AIC)f(y|\theta_1)}{p(M_1|AIC)f(y|\theta_1)+p(M_2|AIC)f(y|\hat{\theta}_2)}$, $W_{al_2} = \frac{p(M_2|AIC)f(y|\hat{\theta}_2)}{p(M_1|AIC)f(y|\theta_1)+p(M_2|AIC)f(y|\hat{\theta}_2)}$.

Figure (4.13) displays model selection probabilities and Figure (4.14) displays risks performance of estimators. It can be seen that Akaike weighting does not perform better than PMSEs when the true parameter is between $(0, 0.3)$ and between $(0.7, 1)$. However, the adjusted weights perform better than both.

4.5.2.2 Multi-model choice

Consider also a choice between the following models: $M_k : Y \sim \text{binomial}(\theta_k, n)$ for arbitrary K models. Each parameter θ_k is known.
For a choice using AIC criterion, since there is no unknown parameter, this is the same as selecting the model with higher likelihood. Model M_{max} is chosen if

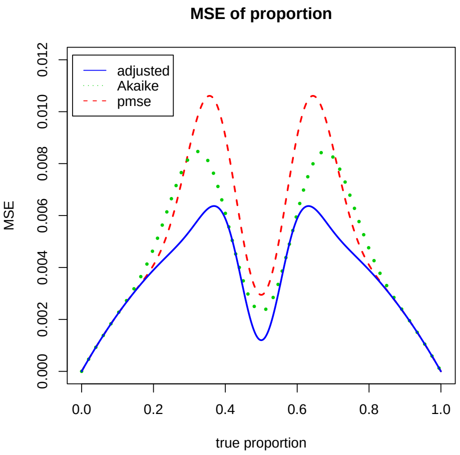

Figure 4.14: Risk of two proportions comparing PMSEs, Akaike weights estimators and adjusted estimators as a function of θ.

$L_{max} \geq L_k, \forall k \in \{1, \ldots, K\}$.
PMSE $\tilde{\theta} = \theta_k$ if M_k is selected.
$\tilde{\theta} = \sum_{k=1}^{K} I_k(f(y|\theta_k) = L_{max})\theta_k$, $I_k = 1$ if M_k is chosen and 0 otherwise. Model selection probability for model M_k is given by: $P_\theta(M_k|AIC, \mathcal{M}) = P_\theta(f(y|\theta_k) = L_{max})$.
The estimated model selection probabilities $p(M_k|AIC)$ are given by replacing θ by the estimated $\hat{\theta} = y/n$. The Akaike weights are defined by $W_{a_k} = \frac{f(y|\theta_k)}{\Sigma_{i=1}^{K} f(y|\theta_i)}$, and the adjusted weights by $W_{al_k} = \frac{p(M_k|AIC)f(y|\theta_k)}{\Sigma_{i=1}^{K} p(M_i|AIC)f(y|\theta_i)}$.

Numerical computations of the properties for these estimators are for $n = 41$, $K = 30$, models are between 0.1 and 0.9 and are given in Figure (4.15). One can see that Akaike weights are not better than PMSEs for certain regions of the parameter space, but the adjusted likelihood weights are better than both.

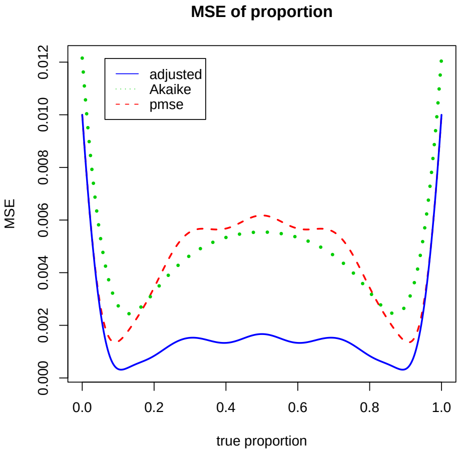

Figure 4.15: Risk of 30 models comparing PMSEs, Akaike weights estimators and adjusted estimators as a function of θ.

Chapter 5

Bayesian Model Selection and Model Averaging

5.1 Introduction

This chapter considers model selection uncertainty in the Bayesian context. We first explain that choosing a model selection method depends on whether interest is focused on identifying the true model, or on choosing a model for inference purposes. We are concerned with the latter. It is explained that, as long as one is concerned with posterior evaluation (Bayes risks, posterior variance, etc.) of an estimate, i.e. conditional on the data, model selection uncertainty is not an issue; the data are held fixed. In this case, model selection has no effect on subsequent inference.

However, if interest is focused on frequentist performance of estimators (e.g. frequentist risk) then the problem of model selection uncertainty exists and can be really severe. This is analogous to the situation discussed in Chapter 3. The properties of *Bayesian post-model-selection estimators* (BPMSEs) are difficult to derive. For example, one can compute confidence regions, but it is not clear how to compute their true coverage probabilities. Secondly, again in the framework of frequentist performance, it is explained that BMA estimators are unlikely to dominate BPMSEs. The reason is that the former do not take account of the selection procedure. An alternative approach, adjusted Bayesian model averaging (ABMA) is proposed which takes into account the selection procedure. The approach, which is based on *prior model selection probabilities* is illustrated using a simple example involving the estimation of proportions.

5.2 Bayesian model selection

Consider the data x and the set of K models $\mathcal{M} = (M_1, \ldots, M_K)$, containing the true model M_t, where each model M_k consists of a family of distributions $P(x|\theta_k, M_k)$, θ_k a possible vector of parameters. One assigns a prior probability $P(\theta_k|M_k)$ to the parameter of each model and a prior probability $P(M_k)$ that model M_k is the true model. $\tilde{M}(X|S, \mathcal{M})$ the selected model (depends on the data X viewed as random), $\triangle$ the quantity of interest, Γ the parameter space for $\triangle$, $\hat{\triangle}_k$ the Bayes estimate of $\triangle$ for each model M_k.

5.2.1 Utility approach and analyst's goal

The aim of this section is to stress how various model selection criteria can been derived from different choices of the utility function (depending on the goal of the analyst). Let $u(m, \triangle)$ be the utility of negative loss of action m given the unknown quantity of interest $\triangle$. The optimal action m^* is that maximising

$$\overline{u}(m|x) = \int_{\triangle} u(m, \triangle) P(\triangle|x) d\triangle, \tag{5.1}$$

where $P(\triangle|x)$ represents the actual beliefs about $\triangle$ having observed x. One can see that (5.1) depends on the choice of the utility function and the computation of the posterior of $\triangle$. More on utility approach for model selection can be found in Chipman, George and McCulloch (2001), Bernado and Rueda (2002). In our framework, the action space is referred to model space so that (5.1) for the optimal model choice m^* becomes

$$\overline{u}(m^*|x) = \sup_{m \in \mathcal{N}} \int_{\triangle} u(m, \triangle) P(\triangle|x) d\triangle\,. \tag{5.2}$$

In the following, our choice of the posterior of $\triangle$ will be based on the total law of probability over model given

$$P(\triangle|x) = \Sigma_{k=1}^{K} P(\triangle|x, M_k) P(M_k|x). \tag{5.3}$$

The aim of the analyst could be either identifying the true model (M_t) or choosing a model for inference.

5.2.1.1 Identifying the true model

Here we assume that the action to be taken is $\triangle$=M_t, *choosing the true model.* In this case a natural choice of utility function is $u(m_k, \triangle) = 1$ if $\triangle = M_k$ and 0 otherwise. A natural choice for the posterior of $\triangle$ for each model can be

$P(\triangle|x, M_k) = 1$ if $\triangle = M_k$ and 0 otherwise. From (5.3), $P(\triangle|x) = P(\triangle|x, M_k)$ if $\triangle = M_k$ and 0 otherwise. The expected utility for the decision m_k given x is

$$\overline{u}(m_k|x) = \int_{\triangle} u(m_k, \triangle)P(\triangle|x)d\triangle = P(M_k|x).$$

This means that the optimal decision is to choose the model with the highest posterior probability. From the Bayes factor framework, the Bayes factor for model M_i versus model M_j is defined to be

$$B_{ij} = \frac{P(M_i|x)}{P(M_j|x)}\frac{P(M_j)}{P(M_i)}.$$

Model M_i is chosen if $B_{ij} > 1$. For equal model prior probability, model M_i is chosen if $P(M_i|x) > P(M_j|x)$.

Hypothesis testing

Consider again that the parameter of interest is the choice of the true model and consider 2 models M_1 and M_2. Define the utility function of the form: $u(m_k, \triangle) = -u_{kl}$ where $u_{11} = u_{22} = 0$ and $u_{12}, u_{21} > 0$. Using $P(\triangle|x, M_k) = 1$ if $\triangle = M_k$ and 0 otherwise. The expected utility of m_k is given by

$$\overline{u}(m_k|x) = -u_{k1}P(M_1|x) - u_{k2}P(M_2|x),$$

then model M_1 is preferred if

$$\frac{P(M_1|x)}{P(M_2|x)} > \frac{u_{12}}{u_{21}}.$$

If $u_{12} = u_{21}$, we recognise the 0-1 utility case with the choice of model with highest posterior.

5.2.1.2 Choosing a model for inference

Quadratic loss for prediction or estimation
Let define the utility by $u(m_k, \hat{\triangle}_k, \triangle) = -(\hat{\triangle}_k - \triangle)^2$, $\triangle$ may be a future observation or any unknown quantity.
Conditioning on model M_k, the optimal choice is the value of $\hat{\triangle}$ minimising

$$\int_{\triangle} (\hat{\triangle} - \triangle)^2 P(\triangle|x, M_k)d\,\triangle\,. \tag{5.4}$$

The optimal choice is $\hat{\triangle}_k = \int_{\triangle} \triangle P(\triangle|x, M_k)d\triangle = \mathrm{E}(\triangle|x, M_k)$.

5.2.1.3 Other loss functions

For the logarithm form $u(m_k, \triangle) = \log P(\triangle|x, M_k)$, the expected utility of choosing model M_k is given by

$$\overline{u}(m_k|x) = \textstyle\int_\Gamma \log P(\triangle|x, M_k)P(\triangle|x)d\triangle$$

$$= \textstyle\int_\triangle \log P(\triangle|x, M_k)(\Sigma_{l=1}^K P(\triangle|x, M_l)P(M_l|x))d\triangle, \text{under}(5.3)$$

$$= \Sigma_{l=1}^K P(M_l|x)\{\textstyle\int_\Gamma \log P(\triangle|x, M_k)P(\triangle|x, M_l)d\triangle\}$$

$$= \Sigma_{l=1}^K P(M_l|x)\{\textstyle\int_\Gamma (\log P(\triangle|x, M_l) - \log[\frac{P(\triangle|x,M_l)}{P(\triangle|x,M_k)}])P(\triangle|x, M_l)d\triangle\}$$

$$= \underbrace{\Sigma_{l=1}^K P(M_l|x)\{\int_\Gamma \log[\frac{P(\triangle|x, M_l)}{P(\triangle|x, M_k)}]P(\triangle|x, M_l)d\triangle\} +}_{(1)}$$

$$\underbrace{\Sigma_{l=1}^K P(M_l|x)\{\int_\Gamma (\log P(\triangle|x, M_l)P(\triangle|x, M_l)d\triangle\}}_{(2)}.$$

The second term (2) does not depend on M_k, therefore the selected model is given by minimising over M_k

$$\Sigma_{l=1}^K \{P(M_l|x)\int_\triangle \log[\frac{P(\triangle|x, M_l)}{P(\triangle|x, M_k)}]P(\triangle|x, M_l)d\triangle\}. \tag{5.5}$$

For a utility function of the form

$$u(m_k, \triangle) = 2P(\triangle|x, M_k) - \int_\triangle P^2(\triangle|x, M_k)d\triangle,$$

similar computations yield that the selected model is given by minimising over M_k

$$\Sigma_{l=1}^K P(M_l|X)\Big\{\int_\triangle \{[2P(\triangle|x, M_l) - \int_\triangle P^2(\triangle|x, M_l)d\triangle] - [2P(\triangle|x, M_k) - \int_\triangle P^2(\triangle|x, M_k)d\triangle]\}P(\triangle|x, M_l)d\triangle\Big\}.$$

It follows that, if one needs to select a model followed by inference, other loss functions are more appropriate. In general, a change in the utility function corresponds to a different optimal decision. For instance, Bernado and Rueda (2002) propose some continuous loss functions for that purpose. Barbieri and Berger (2004) note also that choosing the model with higher posterior is true under more general conditions, for example in linear models, this is often true for orthogonal matrices designs. Barbieri and Berger (2004) show that for normal linear model, the optimal predictive model is the median probability model, under some strong conditions. Their findings are also explained using a geometric representation.

5.2.2 BMA as model selection criterion

Proposition 5.2.1 *Under the square error loss and the weighted posterior probability of Equation (5.3), the selected model is the one whose estimate is closest to BMA estimate.*

Proof. Conditioning on all models, that is under Equation (5.3), the optimal choice is the value $\hat{\triangle}$ minimising

$$\int_{\triangle} (\hat{\triangle}_k - \triangle)^2 P(\triangle|x) d\,\triangle\,. \tag{5.6}$$

This is equivalent to minimise

$$\begin{aligned}
&\textstyle\int_\Gamma (\hat{\triangle}_k - \triangle)^2 \{\Sigma_{l=1}^K P(\triangle|x, M_l) P(M_l|x) d\triangle\} \\
&= \Sigma_{l=1}^K P(M_l|x) \{\textstyle\int_\Gamma (\hat{\triangle}_k - \hat{\triangle}_l + \hat{\triangle}_l - \triangle)^2 P(\triangle|x, M_l) d\triangle\} \\
&= \underbrace{\Sigma_{l=1}^K (\hat{\triangle}_k - \hat{\triangle}_l)^2 P(M_l|x)}_{(1)} + \underbrace{\Sigma_{l=1}^K P(M_l|x) \mathrm{Var}(\triangle|x, M_l)}_{(2)}.
\end{aligned}$$

The second term (2) does not depend on model M_k and denoting

$$\hat{\triangle}_{\text{bma}} = \Sigma_{l=1}^K \hat{\triangle}_l P(M_l|x), \tag{5.7}$$

the first term (1) can be rearranged as

$$\begin{aligned}
&\Sigma_{l=1}^K (\hat{\triangle}_k - \hat{\triangle}_l)^2 P(M_l|x) = \Sigma_{l=1}^K ((\hat{\triangle}_{\text{bma}} - \hat{\triangle}_l) + (\hat{\triangle}_k - \hat{\triangle}_{\text{bma}}))^2 P(M_l|x) \\
&= \underbrace{\Sigma_{l=1}^K (\hat{\triangle}_{\text{bma}} - \hat{\triangle}_l)^2 P(M_l|x)}_{(1')} + \underbrace{\Sigma_{l=1}^K (\hat{\triangle}_k - \hat{\triangle}_{\text{bma}})^2 P(M_l|x)}_{(2')} \\
&+ \underbrace{2\Sigma_{l=1}^K (\hat{\triangle}_{\text{bma}} - \hat{\triangle}_l)(\hat{\triangle}_k - \hat{\triangle}_{\text{bma}}) P(M_l|x)}_{(3')}. \\
&(2') = (\hat{\triangle}_k - \hat{\triangle}_{\text{bma}})^2 \Sigma_{l=1}^K P(M_l|x) = (\hat{\triangle}_k - \hat{\triangle}_{\text{bma}})^2. \\
&(3') = 2(\hat{\triangle}_k - \hat{\triangle}_{\text{bma}}) \Sigma_{l=1}^K (\hat{\triangle}_{\text{bma}} - \hat{\triangle}_l) P(M_l|x). \\
&= 2(\hat{\triangle}_k - \hat{\triangle}_{\text{bma}})(\hat{\triangle}_{\text{bma}} - \Sigma_{l=1}^K \hat{\triangle}_l P(M_l|x)) = 2(\hat{\triangle}_k - \hat{\triangle}_{\text{bma}})(\hat{\triangle}_{\text{bma}} - \hat{\triangle}_{\text{bma}}) = 0.
\end{aligned}$$

(1’) does not depend on M_k, therefore the only term that depends on M_k is (2’)=$(\hat{\triangle}_k - \hat{\triangle}_{\text{bma}})^2$.
One can see that the preferred model M_k is the one whose estimate $\hat{\triangle}_k$ is closest to the BMA estimate $\hat{\triangle}_{\text{bma}}$.

Corollary 5.2.1 *For two models, the selected model is the one with the highest posterior probability.*

Proof. From Proposition 5.2.1, let find the distance between each model and BMA model. $(\hat{\triangle}_1 - \hat{\triangle}_{\text{bma}})^2 = (\hat{\triangle}_1 - P(M_1|x)\hat{\triangle}_1 - P(M_2|x)\hat{\triangle}_2)^2$
$= (\hat{\triangle}_1(1-P(M_1|x)-P(M_2|x)\hat{\triangle}_2)^2 = (P(M_2|x)\hat{\triangle}_1-P(M_2|x)\hat{\triangle}_2)^2 = P^2(M_2|x)(\hat{\triangle}_1-\hat{\triangle}_2)^2$.
Similarly, $(\hat{\triangle}_2 - \hat{\triangle}_{\text{bma}})^2 = P^2(M_1|x)(\hat{\triangle}_2 - \hat{\triangle}_1)^2$.
Model M_1 is selected if $P^2(M_2|x)(\hat{\triangle}_1 - \hat{\triangle}_2)^2 < P^2(M_1|x)(\hat{\triangle}_2 - \hat{\triangle}_1)^2$.
That is, if $P(M_1|x) > P(M_2|x)$.

5.2.3 Robustness for prior specification

Besides the implementation issue of BMA, priors specification is also of concern. For parametric priors, efforts have been made to develop priors that are robust, even if the results are not really satisfactory. The methods include: natural conjugate, non-informative priors, flat-tailed priors, hierarchical priors, maximum entropy prior, ML-II priors, reference prior (Bernado and Smith, 1994). Details on parameter prior robustness can be found in Robert (2001) and Berger (1985). For the model space prior, a popular and simple choice is $P(M_k) = \frac{1}{K}$. However, as noted in Chipman et al. (2001), when many models are very similar, and only few are distinct, these priors are not robust and may bias the posterior away from the good models. Many model space priors for variable selection can be found in George and McCulloch (1993), Madigan and Raftery (1994), Hoeting et al. (1999), Chipman and al. (2001). Spiegelhalter et al. (1993) and Lauritzen (1996) analyse the benefits of incorporating informative prior models and show improvement in predictive performance. Kass and Raftery (1995) and George (1999) note that posterior model probabilities can quite be sensitive to the prior distribution.

5.2.4 M-open framework

For the M-open framework, the true model M_t does not belong to the set of competing models. The probabilities $P(M_k)$ and $P(M_k|x)$ have no meaningful interpretation, so that the classical Bayesian model selection and BMA are not valid. Suppose that one is interested in the choice of a model for inference about a quantity $\triangle$. For the M-closed perspective, the optimal action $m*$ is that maximising:

$$\overline{u}(m_k, \hat{\triangle}_k) = \int_{\triangle} u(m_k, \triangle, x)P(\triangle|x)d\triangle, \tag{5.8}$$

where $P(\triangle|x) = \Sigma_{k=1}^{K} P(\triangle|x, M_k)P(M_k|x)$, for each model M_k, $k = 1, \ldots, K$. However, $P(\triangle|x)$ is not available and one needs to find the expected utility function in (5.8). Consider the utility function $u(m_k, \triangle, x)$ and a special case where the quantity of interest is a future observation, that is $\triangle = z$. The idea of cross-validation is to leave out one observation, and use the set $x_{-i} = x - x_i$. With this set, one computes $\hat{\triangle}_k^{(i)}$ for each model and the selection rule in Equation (5.8) becomes the optimal action $m*$ is that maximises

$$\overline{u}(m_k, \hat{\triangle}_k) = \frac{1}{n}\Sigma_{i=1}^{n} u(m_k, x_i, \hat{\triangle}_k^{(i)}|x_{-i}). \tag{5.9}$$

For quadratic loss, this maximises over M_k

$$\overline{u}(m_k, \hat{\triangle}_k) = \frac{1}{n}\Sigma_{i=1}^{n}(\hat{\triangle}_k^{(i)} - x_i)^2. \tag{5.10}$$

For logarithm score function, this maximises over M_k

$$\overline{u}(m_k, \hat{\triangle}_k) = \frac{1}{n}\Sigma_{i=1}^{n} logP(x_i|M_k, x_{-i}). \tag{5.11}$$

Other loss functions can be considered. More on cross-validation methods is given in Bernado and Smith (1994), Berger and Pericchi (1996), Key, Pericchi and Smith (1999) and Marriott, Spencer and Pettitt (2001).

5.3 Applied Bayesian inference and Bayesian model selection inference

Here, we compare Bayesian methods in which only informal selection criteria are used to those with both informal and formal selection criteria. Conditioning on observed data, if the same model is selected for both approaches, their properties are identical. If one needs to evaluate the long run performances of the resulting estimator, these properties are different. Both approaches therefore suffer of model selection uncertainty from a frequentist point of view.

5.3.1 Bayesian approach to statistical data analysis

Bayesian data analysis can be summarised as follows:

1. Quantity of interest $\triangle$.

2. Data $x = (x_1, \ldots, x_n)$.

3. use x for exploratory data analysis.

4. From (3), specify a distribution family for the data $M = f(x|\theta)$, $(\triangle = h(\theta))$: there is model uncertainty, since the true model is unknown.

5. Specify a prior distribution for $\theta : \pi(\theta)$.

6. Compute the posterior distribution for $\theta : \pi(\theta|x)$.

7. Define a loss function.

8. Find the optimal decision rule. E.g. for square error loss, $\mathrm{E}(\theta|x)$, $\mathrm{Var}(\theta|x)$ or any quantity, i.e. the posterior properties for $\triangle$.

Here, the analysis is conditioned on the observed data. But, if one needs the frequentist properties, the data should be viewed as random. The step (3) is an informal model selection procedure. Therefore, the use of exploratory data analysis introduces model selection uncertainty.

5.3.2 Bayesian model selection approach to statistical data analysis

1. Quantity of interest $\triangle$.

2. Data $x = (x_1, \dots, x_n)$.

3. Use x for exploratory data analysis.

4. From (3), specify $\mathcal{M} = (M_1, \dots, M_K)$, alternative plausible (parametric θ) models, $\triangle = h(\theta)$.

5. Use $\underline{\text{any}}$ model selection criteria and data x to select a model (model uncertainty) $\hat{M}(x) = M_{\hat{k}(x)} \in \mathcal{M}$, $\hat{k}(x) \in \{1, \dots, K\}$.

6. Specify a prior distribution for $\theta : \pi(\theta)$ from the selected model.

7. Compute the posterior distribution for $\theta : \pi(\theta|x)$ from the selected model.

8. Define a loss function.

9. Find the optimal decision rule. E.g. for square error loss, $\mathrm{E}(\theta|x)$, $\mathrm{Var}(\theta|x)$ or any quantity, e.g. posterior properties for $\triangle$.

The analysis is conditioned on the observed data ($\underline{\text{conditional inference}}$). There is no model selection uncertainty, only model uncertainty, since the data x (viewed as fixed) are used for all steps (including steps 3 and 4). However, if one needs the frequentist properties, the data should be viewed as random because steps 3 and 4 introduce model selection uncertainty and $\hat{M}(X) = M_{\hat{k}(X)} \notin \mathcal{M}$, $\hat{k}(X) \notin \{1, \ldots, K\}$. The difficulties are now similar those of frequentist model selection. In some cases, frequentist performance is not of interest, in which case, Bayesian methods are not concerned with model selection uncertainty. The remaining uncertainty includes the choice of the statistical model, the prior and the loss function.

5.4 Model selection uncertainty

5.4.1 Bayesian post-model-selection estimator

We refer to *Bayesian post-model-selection estimator* (BPMSE), the Bayes estimator after a model selection procedure has been applied. Here, we consider the squared error loss, but the main idea remains unchanged for any other loss function. Given the selection procedure, BPMSE can been written as

$$\tilde{\triangle}(X|S, \mathcal{M}) = \sum_{k=1}^{K} I_k(X|S, \mathcal{M}) \mathrm{E}(\triangle|X, M_k), \tag{5.12}$$

where $I_k(X|S, \mathcal{M}) = 1$ if model M_k is selected and 0 otherwise. In the following, we define the posterior quantity and derive Bayesian-post-model selection in a coherent way. For simplicity, $\triangle_k$ for each model M_k will be replaced only by $\triangle$ in the integrals.

5.4.2 Long-run performance of Bayes estimators

Here the goal of the analysis is to select a model for inference using any selection procedure. One is interested in evaluating the long run performance (frequentist performance) of the selected model. In general, Bayes estimators have good frequentist properties (see, Carlin and Louis, 1996; Bayarri and Berger, 2004). The Bayesian approach can also produce interval estimation with good performance, for example coverage probabilities. It is also known that if a Bayes estimator associated with a prior is unique, then it is admissible (see e.g., Robert 2001). There are also conditions under which Bayes estimator are minimax. The point is to see whether these good frequentist approaches still hold for Bayes estimators after model selection.

$$\tilde{\triangle}(X|S, \mathcal{M}) \notin \{\mathrm{E}(\triangle|X, M_1), \ldots, \mathrm{E}(\triangle|X, M_K)\}.$$

We are interested in studying the frequentist properties of $\tilde{\triangle}(X|S,\mathcal{M})$. The difficulty we will describe here is similar to those encountered in frequentist PMSEs. This is due to the partition of the sample space $\mathcal{X}$ by the selection procedure. This makes it difficult to derive the coverage probability of confidence intervals.

5.4.2.1 The frequentist risk

The frequentist risk of BPMSEs is defined as

$$\mathrm{R}(\triangle,\tilde{\triangle}(X|S,\mathcal{M}))=\mathrm{E}_t[\mathrm{L}(\triangle,\tilde{\triangle}(X|S,\mathcal{M}))], \tag{5.13}$$

where L is a loss function. One can now see that this risk is difficult to compute. It follows that it is difficult to prove admissibility and minimaxity properties of BPMSEs, since their associated priors are not known.

5.4.2.2 Coverage probabilities

When the data have been observed, one can construct a confidence region. Suppose that after observing the data, model M_{k^*} is selected. For large samples, Berger (1985) considers the normal approximation

$$\triangle|x\sim\mathcal{N}_p(\mathrm{E}(\triangle|x,M_{k^*}),\mathrm{Var}(\triangle|x,M_{k^*})) \tag{5.14}$$

and then derives an approximate region at the $1-\alpha$ level given by

$$C_\alpha(x)=\{\triangle;(\triangle-\mathrm{E}'(\triangle|x,M_{k^*})\mathrm{Var}^{-1}(\triangle|x,M_{k^*})(\triangle-\mathrm{E}(\triangle|x,M_{k^*}))\leq d_\alpha^2\},$$

where d_α^2 is the α-quantile of χ_p^2.
A stochastic version (assuming normality) is given by

$$C_\alpha(X)=\{\triangle;(\triangle-\tilde{\triangle}'(X|S,\mathcal{M})Var^{-1}(\tilde{\triangle}'(X|S,\mathcal{M}))(\triangle-\tilde{\triangle}(X|S,\mathcal{M}))\leq d_\alpha^2\}.$$

The coverage probability of the stochastic form is given by

$$P_\triangle(\triangle\in C_\alpha(X))=\mathrm{E}_\triangle I_{C_\alpha(X)}(\triangle),$$

which is now difficult, as it involves computing the variance and expectation of BPMSE.

5.4.2.3 Consistency

Another frequentist property of Bayes estimators is consistency. It is known (e.g., Bayarri and Berger, 2004) that, under appropriate regularity conditions, Bayes estimators are consistent. A question is whether BPMSEs are consistent, but this is difficult to prove because one does not know the priors associated with BPMSEs.

5.4.3 Conditional performance of Bayes estimates

We now consider the conditional performance, the properties of the Bayes estimates conditioned on the observed data. Consider a realisation of the data, $X = x$, then

$$\tilde{\triangle}(x|S,\mathcal{M}) \in \{\mathrm{E}(\triangle|x, M_1), \ldots, \mathrm{E}(\triangle|x, M_K)\} \quad \text{and}$$
$$\mathrm{Var}(\tilde{\triangle}(x|S,\mathcal{M})) \in \{\mathrm{Var}(\triangle|M_1, x), \ldots, \mathrm{Var}(\triangle|M_K, x)\}.$$

That is, $\tilde{\triangle}(x|S,\mathcal{M}) = \mathrm{E}(\triangle|X, M_{k^*})$ and $\mathrm{Var}(\tilde{\triangle}(x|S,\mathcal{M})) = \mathrm{Var}(\triangle|x, M_{k^*})$ where M_{k^*} is the selected model. For example the posterior risk of $\tilde{\triangle}(x|S,\mathcal{M})$ is given by that of the Bayesian estimate under M_{k^*}. That is,

$$\rho(\tilde{\triangle}(x|S,\mathcal{M})) = \rho(\mathrm{E}(\triangle|X, M_{k^*})) = \mathrm{E}_{k^*}[\mathrm{L}(\triangle, \mathrm{E}(\triangle|X, M_{k^*}))].$$

One can construct confidence region (but does not know the coverage). This means that, if only posterior analysis is of interest, Bayesian model selection does not suffer from model uncertainty problem.

5.5 Adjusted Bayesian model averaging

In this framework, we are concerned with the long run performance of BPMSES, not on posterior evaluation, since in the posterior evaluation, the model selection uncertainty problem does not exist. Under model selection uncertainty, from (5.12), a fundamental ingredient is the selection procedure S. This selection procedure should depend on the objective of the analyst and should be taken into account in modelling uncertainty at two levels: prior and posterior to the data analysis.

5.5.1 Prior representation of model selection uncertainty

The initial representation of model uncertainty is captured by parameter prior uncertainty and the model space prior, the selection procedure is used to update model prior. Formally, consider the possible models $M_1, \ldots, M_K$; assign a prior probability $P(\theta_k|M_k)$ to the parameter of each model and a prior probability $P(M_k)$ to each model with the data X viewed as random. Let $M_k(S)$ be event *model M_k is selected*, M_k is considered to be the event *model M_k is true.* We refer to the probability of this event as *prior model selection probability* of model M_k and denoted by $P(M_k(S))$. This is to update prior model $P(M_k)$ using the selection procedure S. $P(M_k)$ may be informative or not, but $P(M_k(S))$ is an informative prior. Making use of the fact that one of the models is true, $P(M_k(S))$ can been computed as

$$P(M_k(S)) = \Sigma_{j=1}^{K} P(M_k(S)|M_j)P(M_j), \tag{5.15}$$

where $P(M_k(S)|M_j)$ is the *prior model selection probabilities* of model M_k given that M_j is the true model. $P(M_k(S)|M_k)$ is the probability that M_k is actually selected given that it is really the true model. The true state of the nature is that a given model is true. The decision here is to select a model. Given that model M_j is true, $\Sigma_{k=1}^{K} P(M_k(S)|M_j) = 1$. These probabilities can be computed as

$$P(M_k(S)|M_j) = \mathrm{E}_j^{\theta}[\mathrm{E}_j^{X}(I_k(X))]. \tag{5.16}$$

The expectation is taken with respect to the true model M_j, provided that these expectations exist. Note that these probabilities do not depend on the observed data.

Table (5.1) gives the true state of the world (nature) and the decision (the selected model). The $P_{jk} = P(M_k(S)|M_j)$, the probability that M_k is selected, given that M_j is the true model. Suppose that M_j is the true model, one would like P_{jj} to be higher, ideally 1 (the correct decision). If model M_j is not selected with probability one, $\alpha_j = 1 - P_{jj} = 1 - \Sigma_{k=1,k\neq j}^{K} P_{jk}$ is called the probability of *Type I error for model M_j*. That is, if M_j is the true model and the selection procedure S incorrectly does not select it, then the selection procedure has made a Type I Error.

On the other hand, if M_k is the true model, but the selection procedure selects M_j, then this selection procedure has made a *Type II error*, with probability P_{kj}, $j \neq k$. The reliability of the selection criterion is given by the closeness of P_{jj} to 1.

Nature and Decision	$M_1(\mathcal{S})$	$M_2(\mathcal{S})$	...	$M_j(\mathcal{S})$	...	$M_K(\mathcal{S})$
M_1	$\mathbf{P_{11}}$	P_{12}	-	P_{1j}	-	P_{1K}
M_2	P_{21}	$\mathbf{P_{22}}$	-	P_{2j}	-	P_{2K}
...	-	-	-	-	-	-
M_j	P_{j1}	P_{j2}	-	$\mathbf{P_{jj}}$	-	P_{jK}
...	-	-	-	-	-	-
M_K	P_{K1}	P_{K2}	-	P_{Kj}	-	$\mathbf{P_{KK}}$

Table 5.1: True and selected models.

5.5.2 Posterior representation of model selection uncertainty

When the data have been observed, the *posterior model selection probability* for each model M_k is given by

$$P(M_k(S)|x) = \frac{P(x|M_k(S))P(M_k(S))}{\Sigma_{j=1}^{K} P(x|M_j(S))P(M_j(S))}, \tag{5.17}$$

where

$$P(x|M_j(S)) = L(x|M_k(S)) = \int_{\Theta} P(x|\theta_k, M_k(S))P(\theta_k|M_k(S))d\theta_k \tag{5.18}$$

is the marginal likelihood of $M_k(S)$. For $P(\theta_k|M_k(S))$ discrete, (5.18) is a summation. $P(M_k(S)|x)$ is the conditional probability that M_k was the selected model. Computations are conditioned on each model, since one will never know the selection for random data. This is similar to the fact that the true model will never be known, and each of the models can be viewed as a possible true model.

5.5.2.1 Posterior distribution

Now, after the data x is observed, and given the selection procedure S, from the law of total probability, the posterior distribution of $\triangle$ is then given by

$$P(\triangle|x, S) = \Sigma_{k=1}^{K} P(\triangle|x, M_k(S))P(M_k(S)|x). \tag{5.19}$$

$P(\triangle|x, S)$ is an average of the posterior of each model $M_k(S)$, $P(\triangle|x, M_k(S))$, weighted by posterior model selection probability.

5.5.2.2 Posterior mean and variance

Proposition 5.5.1 *Under (5.19), the posterior mean and variance are given by*

$$\begin{aligned}&\hat{\triangle} = \mathrm{E}(\triangle|x, S) = \Sigma_{k=1}^{K} \mathrm{E}(\triangle|x, M_k(S))P(M_k(S)|x),\\ &\mathrm{Var}(\triangle|x, S) = \Sigma_{k=1}^{K} P(M_k(S)|x)\{\mathrm{Var}(\triangle|x, M_k(S)) + (\mathrm{E}(\triangle|x, M_k(S)) - \mathrm{E}(\triangle|x, S))^2\},\end{aligned} \tag{5.20}$$

where $\mathrm{E}(\triangle|x, M_k(S))$ *and* $\mathrm{Var}(\triangle|x, M_k(S))$ *are respectively the posterior mean and the posterion variance of* $\triangle$ *for model* M_k *if* M_k *was the selected model.*

Proof. Under (5.19), the posterior mean is
$\mathrm{E}(\triangle|x, S) = \int_\Lambda \triangle P(\triangle|x, S)d\triangle = \int_\Lambda \triangle\{\Sigma_{k=1}^{K} P(\triangle|x, M_k(S))P(M_k(S)|x)\}d\,\triangle$

$= \Sigma_{k=1}^{K}[P(M_k(S)|x)\{\int_\Lambda \triangle P(\triangle|x, M_k(S))d\triangle\}] = \Sigma_{k=1}^{K} P(M_k(S)|x)\mathrm{E}(\triangle|x, M_k(S)).$

The posterior variance under (5.19) is

$$\mathrm{Var}(\triangle|x,S) = \int_{\triangle}(\triangle - \mathrm{E}(\triangle))^2 P(\triangle|x)d\triangle$$

$$= \int_{\Gamma}(\triangle - \hat{\triangle})^2\{\Sigma_{k=1}^{K} P(\triangle|x, M_k(S))P(M_k(S)|x)\}d\triangle$$

$$= \Sigma_{k=1}^{K}[P(M_k(S)|x)\underbrace{\{\int_{\Gamma}(\triangle - \hat{\triangle})^2 P(\triangle|x, M_k(S))d\triangle\}}_{R_k(\hat{\triangle})}].$$

$$\mathrm{R}_k(\hat{\triangle}|S) = \mathrm{E}_k(\triangle - \hat{\triangle})^2 = \mathrm{E}_k(\triangle - \mathrm{E}(\triangle|x, M_k(S)) + \mathrm{E}(\triangle|x, M_k(S)) - \hat{\triangle})^2$$

$$= \mathrm{E}_k(\triangle - \mathrm{E}(\triangle|x, M_k(S)))^2 + \mathrm{E}(\mathrm{E}(\triangle|x, M_k(S)) - \hat{\triangle})^2$$
$$-2\mathrm{E}[(\triangle - \mathrm{E}(\triangle|x, M_k(S)))((\mathrm{E}(\triangle|x, M_k(S)) - \hat{\triangle})]$$

$$= \mathrm{Var}(\triangle|x, M_k(S)) + (\mathrm{E}(\triangle|x, M_k(S)) - \hat{\triangle})^2 + 2(\mathrm{E}(\triangle|x, M_k(S)) -$$
$$\hat{\triangle})\underbrace{(\mathrm{E}^{\triangle_k|x)}(\triangle - \mathrm{E}(\triangle|x, M_k(S))))}_{=0}$$

$$= \mathrm{Var}(\triangle|x, M_k(S)) + (\mathrm{E}(\triangle|x, M_k(S)) - \hat{\triangle})^2.$$

$\mathrm{R}_k(\hat{\triangle}|S)$ is the posterior expectation loss for model M_k for taking the decision rule $\hat{\triangle}$ rather than $\mathrm{E}(\triangle|x, M_k(S))$.

The method can be then summarised as follows:

1. $P(M_k)$ represents the prior model uncertainty,

2. $P(M_k(S))$ updates prior model uncertainty by taking into account the selection procedure,

3. $P(M_k(S)|x)$ is the overall posterior representation of the model selection uncertainty.

Note that if the unconditional model selection probability is equal to model prior, then the proposed weights are the same as BMA weights, namely the probability that each model is true given the data, $P(M_k|x)$. For the proposed weights, one needs to compute the marginal likelihood and these model selection probabilities. This is not an easy task. However, methods exist in the literature for doing such computations. These include Markov chain Monte Carlo methods, noniterative Monte Carlo methods and asymptotic methods.

5.5.2.3 A basic property

From the nonnegativity of Kullback-Leiber information divergence, it follows that $\forall j = 1, \ldots, K$:

$$\mathrm{E}[log\{\Sigma_{k=1}^{K} P(\triangle|x, M_k(S))P(M_k(S)|x)\}] \geq \mathrm{E}[logP(\triangle_j|x, M_j(S))], \tag{5.21}$$

where the expectation is taken with respect to the posterior distribution in (5.19). This logarithm score rule was suggested by Good (1952). This means that under the use of a selection criterion and the posterior distribution given in (5.19), FBMA provides better predictive ability (under logarithm score rule) than any single selected model.

For computational purposes, $P(M_k(S)|x)$ can be written as

$$P(M_k(S)|x) = \frac{P(M_k(S))B_{kj}(x|S)}{\Sigma_{i=1}^{K} P_i(M_i(S))B_{ij}(x|S)}, \tag{5.22}$$

where $B_{ij}(x|S)$ is the Bayes factor, summarising the relative support for model M_i versus model M_j using posterior model selection probabilities. Using Laplace approximation of the marginal likekihood, the weights in (5.22) become

$$P(M_k(S)|x) = \frac{P_k(M_k(S))exp(-\frac{BIC_k(S)}{2})}{\Sigma_{i=1}^{K} P_i(M_i(S))exp(-\frac{BIC_i(S)}{2})}, \tag{5.23}$$

where $BIC_k(S)$ is Bayesian information criterion for model $M_k(S)$.

5.6 Estimating a multivariate mean

Let $X = (X_1, ..., X_p)'$ be a p-dimensional random vector, $p \geq 3$. Suppose $X \sim N_p(\theta, I_p)$, with unknown mean $\theta = (\theta_1, ..., \theta_p)'$ and θ is a random with prior probability τ. The marginal distribution of X is given by

$$L(X) = \frac{1}{(2\pi)^{p/2}} \int_{\Theta} e^{-\|x-\theta\|^2/2} \tau(\theta) d\theta. \tag{5.24}$$

The Bayes estimate $\hat{\theta}$ of θ with respect to τ, obtained by minimising over ρ the quantity $\mathrm{E}\|\theta - \rho(X)\|^2$ is given by (Stein, 1981):

$$\hat{\theta} = X + \bigtriangledown \log L, \tag{5.25}$$

where $\bigtriangledown \log L = \partial \log L / \partial x$. Now for a set of models, suppose that the marginal of each model, L_k, is of the form (5.24). This means that each Bayes estimator will be of the form (5.25), that is $\hat{\theta}_k = X + \bigtriangledown \log L_k$. Let $P(M_k|S)$ be model

selection probability of each model M_k.
Let $\Gamma(X|S,\mathcal{M}) = \Sigma_{k=1}^{K} P(M_k|S)L_k$ and $\tau(\theta) = \Sigma_{k=1}^{K} P(M_k|S)\tau_k(\theta)$,
where L_k and τ_k are respectively marginal likelihood and prior for θ in each model.

Theorem 5.6.1 *Assume the following*

1. $X \sim N_p(\theta, I_p)$,

2. $L_k = \frac{1}{(2\pi)^{p/2}} \int_\Theta e^{-\|x-\theta\|^2/2} \tau_k(\theta) d\theta$,

3. $\tau(\theta)$ *is superharmonic, that is* $\Sigma_{i=1}^{p} \partial^2\tau/\partial\theta_i^2 \le 0$,

then $\tilde{\theta}_{\text{abma}}(S) = \Sigma_{k=1}^{K} P(M_k(S)|x)\hat{\theta}_k$ *is a minimax estimator for* θ *and its risk is*

$$\mathrm{R}(\tilde{\theta}_{\text{abma}}(S)) = p + \frac{\bigtriangledown^2 \Gamma(X|S,\mathcal{M})}{\Gamma(X|S,\mathcal{M})} - \| \bigtriangledown \log \Gamma(X|S,\mathcal{M})\|^2.$$

Proof. Let start with the following lemma which is straightforward but important.

Lemma 5.6.1 *Under assumption (2),*

$$\tilde{\theta}_{\text{abma}}(S) = X + \bigtriangledown \log \Gamma(X|S,\mathcal{M}) \tag{5.26}$$

and $\Gamma(X|S,\mathcal{M})$ *is the marginal of* X *under mixture prior* τ.

Proof of the Lemma 5.6.1. From assumption (2), and Equation (5.25) each Bayesian estimator is of the form $\hat{\theta}_k = X + \bigtriangledown \log L_k$.

We have the notation $\bigtriangledown f = (\partial f/\partial X_1, \ldots, \partial f/\partial X_p)'$ and $\bigtriangledown \log f = \frac{\bigtriangledown f}{f}$.

$P(M_k(S)|X) = \frac{P(M_k|S)L_k}{\Sigma_{i=1}^{K} P(M_i|S)L_i} = \frac{P(M_k|S)L_k}{\Gamma(X|S,\mathcal{M})}$.

then $\tilde{\theta}_{\text{abma}}(S) = \Sigma_{k=1}^{K} P(M_k(S)|X)\hat{\theta}_k$

$= \Sigma_{k=1}^{K} \left\{ \frac{P(M_k|S)L_k}{\Gamma(X|S,\mathcal{M})} \right\} \hat{\theta}_k$

$= \frac{1}{\Gamma(X|S,\mathcal{M})} \Sigma_{k=1}^{K} P(M_k|S) L_k (X + \bigtriangledown log L_k)$

$= X + \frac{1}{\Gamma(X|S,\mathcal{M})} \Sigma_{k=1}^{K} P(M_k|S) L_k (\frac{\bigtriangledown L_k}{L_k})$

$= X + \frac{1}{\Gamma(X|S,\mathcal{M})} \Sigma_{k=1}^{K} P(M_k|S) \bigtriangledown L_k = X + \frac{1}{\Gamma(X|S,\mathcal{M})} \bigtriangledown (\Sigma_{k=1}^{K} P(M_k|S) L_k)$

$= X + \frac{1}{\Gamma(X|S,\mathcal{M})} \bigtriangledown \Gamma(X|S,\mathcal{M}) = X + \bigtriangledown \log \Gamma(X|S,\mathcal{M})$.

Under $\tau(\theta)$, the marginal L of X is derived as:
$L = \frac{1}{(2\pi)^{p/2}} \int_\Theta e^{-\|x-\theta\|^2/2}[\Sigma_{k=1}^K P(M_k|S)\tau_k(\theta)]d\theta$

$= \Sigma_{k=1}^K P(M_k|S)\{\frac{1}{(2\pi)^{p/2}} \int_\Theta e^{-\|x-\theta\|^2/2}\tau_k(\theta)d\theta\} = \Sigma_{k=1}^K P(M_k|S)L_k = \Gamma(X|S,\mathcal{M})$.
Now we are ready to use Stein's results. The posterior risk of $\tilde{\theta}_{\text{abma}}(S)$ is given by $R(\tilde{\theta}_{\text{abma}}(S)) = p + \frac{\bigtriangledown^2\Gamma(X|S,\mathcal{M})}{\Gamma(X|S,\mathcal{M})} - \| \bigtriangledown \log \Gamma(X|S,\mathcal{M})\|^2$.
Now, from Stein (1981), p.1141, $\tau(\theta)$ is superharmonic. This implies that $\Gamma(X|S)$ is also superharmonic. Using Lemma 5.6.1 and Stein's (1981) results, $\tilde{\theta}_{\text{abma}}(S)$ is a minimax estimator for θ.
The theorem remains valid if one assumes that each $\tau_k(\theta)$ is superharmonic. The theorem is also valid for classical BMA by just replacing $P(M_k|S)$ by $P(M_k)$, the model prior probability and the weights $P(M_k(S)|X)$ by posterior model probabilities $P(M_k|X)$.

5.7 Application to one-way ANOVA

A typical example for estimation of the mean of a multivariate normal distribution is a one-way ANOVA.
$Y_{ij} = \theta_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0,1)$, iid, $i = 1,\dots,p$, $p \geq 3$, $j = 1,\dots,n$; $\theta_i \sim N(0,\varrho)$ and are iid, $\theta = (\theta_1,\dots,\theta_p)'$.
A Bayes estimator of θ_i is $\hat{\theta}_i = (\frac{\varrho^2}{\varrho^2+1})Y_i.$, where $Y_i. = \frac{1}{n}\Sigma_{j=1}^n Y_{ij}$.
Now, suppose that the linear form is correct and the only uncertainty is the choice of prior probability over the parameter θ. One needs to select among different models, where priors over θ_i differ.
That is, M_k is just characterised by $\theta_i \sim N(0,\varrho_k)$, $k = 1,\dots,K$. The Bayes estimator for each model is $\hat{\theta}_{ki} = (\frac{\varrho_k^2}{\varrho_k^2+1})Y_i$. Suppose that a selection procedure S is given, for example choosing a model with the highest posterior. The naive procedure consists of selecting a model M_0 and then obtaining the Bayes estimator $\hat{\theta}_{0i} = (\frac{\varrho_0^2}{\varrho_0^2+1})Y_i$.
Instead of applying this naive procedure, one can use the proposed method to weight each Bayes estimator.

5.8 Estimating a proportion

In Bayesian analysis, we use the following notation: quantity of interest $\triangle = \theta$, with prior $\pi(\theta)$, given data x, posterior of θ is $\pi(\theta|x)$, and sample space χ for any

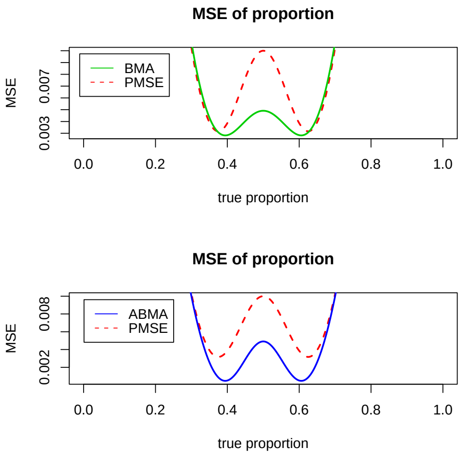

Figure 5.1: Risk of two proportions comparing BPMSE, BMA and ABMA estimators as a function of θ.

decision rule $\delta(x)$, given θ, the distribution of X is $f(x|\theta)$. The frequentist risk of $\delta(x)$ is

$$\mathrm{R}(\theta, \delta) = \mathrm{MSE}_\theta(\delta) = \int_{\mathcal{X}} (\delta(x) - \theta)^2 f(x|\theta) dx.$$

The Bayes risk of $\delta(x)$ is $\int_\Theta R(\theta, \delta) d\theta$ and is constant.

For some models, we will use the beta prior for θ and make use of the following general result about this kind of priors (see for example Casella and Berger (1990), p.298):
$X|\theta \sim \mathrm{binomial}(n, \theta)$, $\theta \sim \mathrm{beta}(\alpha, \beta)$, then $\theta|x \sim \mathrm{beta}(x+\alpha, n-x+\beta)$, therefore

$$\hat{\theta} = \mathrm{E}(\theta|x) = \frac{x+\alpha}{\alpha+\beta+n}$$

is the Bayes estimate of θ. The marginal distribution of X is the beta-binomial(n, α, β), whose PDF is given by

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}.$$

5.8.1 Long run evaluation

5.8.1.1 Two-model choice

(a) $M_1 : \theta = \theta_1$ and $M_2 : \theta = \theta_2$, priors for models $P(M_1)$, $P(M_2)$ are given.
The same set-up, like in the frequentist approach, can be carried out in the

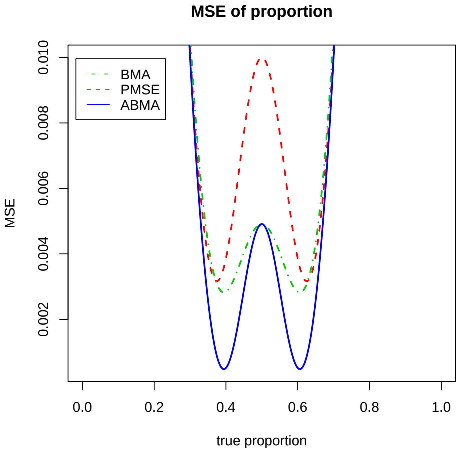

Figure 5.2: Risk of two proportions comparing BPMSE, BMA and ABMA estimators as a function of θ.

Bayesian approach. The degenerate prior $\pi(\theta_1) = \pi(\theta_2) = 1$ can be used. In the framework of hypothesis testing, Bernado and Smith (1994), p.391, refer to (a) as ”simple versus simple test” .
$P(x|M_k) = f(x|\theta_k)\pi(\theta = \theta_k) = f(x|\theta_k)$, k=1,2.
Posterior model probabilities $P(M_k|x)$ are given by
$P(M_k|x) = \frac{P(M_k)f(x|\theta_k)}{\Sigma_{i=1}^2 P(M_i)f(x|\theta_i)}$.
Model 1 is selected if $P(M_1|x) > P(M_2|x)$,

$$\Longleftrightarrow \frac{P(M_1)f(x|\theta_1)}{\Sigma_{i=1}^2 P(M_i)f(x|\theta_i)} > \frac{P(M_2)f(x|\theta_2)}{\Sigma_{i=1}^2 P(M_i)f(x|\theta_i)},$$

$$\Longleftrightarrow \frac{P(M_1)}{f(x|\theta_1)} > \frac{P(M_2)}{f(x|\theta_2)}, \Longleftrightarrow \frac{f(x|\theta_1)}{f(x|\theta_2)} > \frac{P(M_2)}{P(M_1)}.$$

$$\Longleftrightarrow R = log(f(x|\theta_1)) - log(f(x|\theta_2)) > log[\frac{P(M_2)}{P(M_1)}]$$
$$\Longleftrightarrow x[log\frac{\theta_1}{\theta_2} - log[\frac{1-\theta_1}{1-\theta_2}]] + nlog[\frac{1-\theta_1}{1-\theta_2}] > log[\frac{P(M_2)}{P(M_1)}]$$

$$\Longleftrightarrow x > \frac{-nlog[\frac{1-\theta_1}{1-\theta_2}]}{log[\frac{\theta_1(1-\theta_2)}{\theta_2(1-\theta_1)}]} - log[\frac{P(M_2)}{P(M_1)}] = a_n(\theta_1, \theta_2) - log[\frac{P(M_2)}{P(M_1)}] = b_n(\theta_1, \theta_2).$$

$p_1 = P_\theta(X > b_n(\theta_1, \theta_2)) = 1 - P_\theta(X < b_n(\theta_1, \theta_2)) = 1 - F_{B(n,\theta)}(b_n(\theta_1, \theta_2))$.
BMA corresponds to weighting the models with their posterior; the corresponding estimator is $\theta_{\text{BMA}} = \theta_1 P(M_1|x) + \theta_2 P(M_2|x)$.
The BPMSE $\tilde{\theta} = \theta_1$ if M_1 is selected and θ_2 otherwise.
For illustration of the case $P(M_1) \neq P(M_2)$, we take $n = 41$, $P(M_1) = 0.3$, $P(M_2) = 0.7$, $\theta_1 = 0.6$, $\theta_2 = 0.4$.

Figure (5.1) illustrates the performance BPMSE, BMA and ABMA.
BMA and ABMA have similar performance. However, for some regions of the parameter space, BMA tends not to be better than BPMSE.
Only points $\theta = \theta_1 = 0.6$ and $\theta = \theta_2 = 0.4$ are relevant since the true model is one of the two.
Figure (5.2) shows these estimators all together.

(b) Consider the following two models: $M_1 : X \sim \text{Be}(n, \theta_1)$, $P(\theta = \theta_1) = 1$, noninformative prior and $M_2 : X \sim \text{Be}(n, \theta), \theta \sim \text{beta}(\alpha, \beta)$.
Let the selection procedure be choosing the model with higher posterior.

$P(x|M_1) = f(x|\theta_1)$ and $P(x|M_2) = f(x) = \binom{n}{x}\frac{\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}$,

$\hat{\theta}_2 = E(\theta|x) = \frac{x+\alpha}{\alpha+\beta+n}$.
M_1 is chosen if $P(M_1|x) > P(M_2|x)$.

$P(M_1|x) = \frac{P(M_1)f(x|\theta_1)}{P(M_1)f(x|\theta_1)+P(M_2)f(x)}$.

$P(M_2|x) = \frac{P(M_2)f(x)}{P(M_1)f(x|\theta_1)+P(M_2)f(x)}$.

$\tilde{\theta}_{\text{bpmse}} = \theta_1 I_1(P(M_1|x) > P(M_2|x)) + \hat{\theta}_2 I_2(P(M_1|x) \leq P(M_2|x))$.

$p_1 = E_\theta(I_1(P(M_1|x) > P(M_2|x))$ and $p_2 = E_\theta(I_2(P(M_1|x) \leq P(M_2|x))$.

The parameters are: $n = 41$, $\alpha = \beta = 1$, that is $\theta \sim U(0, 1)$, $\theta_1 = 0.5$.

(c) Consider the following two models: $M_1 : X \sim \text{binomial}(n, \theta)$, $\pi(\theta) = 1$ (degenerate prior) and $M_2 : X \sim \text{binomial}(n, \theta), \quad \theta \sim \text{beta}(\alpha, \beta)$.
Similar degenerate priors for model 1 can be seen in Berger (1985, p.132 and p.230), Robert (2001, p.226 and p.404)
Estimators for M_1:
marginal=$f_1(x) = \int_0^1 f(x|\theta)\pi(\theta)d\theta = \int_0^1 f(x|\theta)d\theta$.

$f_1(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{f_1(x)}$, $\hat{\theta_1} = E(\theta|x) = \int_0^1 \theta f(\theta|x)d\theta$.
Figure (5.4) shows the MSE of BPMSE, BMA and ABMA. As can be seen BMA does not dominate BPMSE, but ABMA does.

5.8.1.2 Multi-model choice

(a) Consider also a choice between the following models: $M_k : X \sim \text{binomial}(n, \theta_k)$ for arbitrary K models, with degenerate $\pi(\theta_k) = 1$.
Similar to part (a) with two models, here, instead, we have $K = 30$ and $n = 41$.

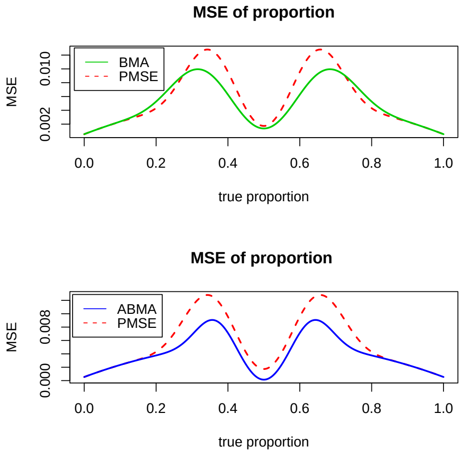

Figure 5.3: Risk of two proportions comparing BPMSE, BMA and ABMA as a function of θ.

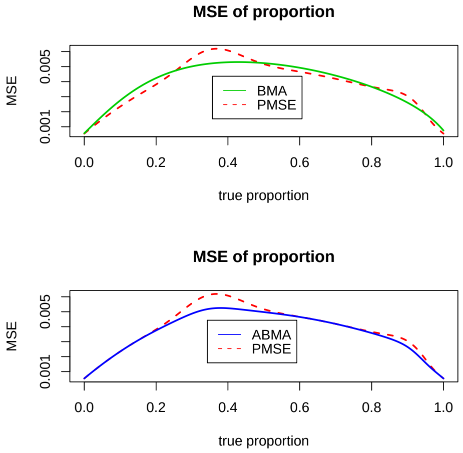

Figure 5.4: Risk of two proportions comparing BPMSE, BMA and ABMA as a function of θ.

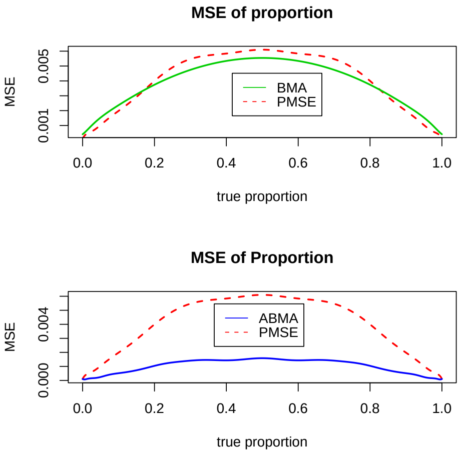

Figure 5.5: Risk of 30 simple models comparing BPMSE, BMA and ABMA as a function of θ.

Figure (5.5) shows the MSE of BPMSE, BMA and ABMA. As can be seen BMA does not dominate BPMSE, but ABMA does.

(b) Consider also a choice between the following models: $M_k : X \sim \text{binomial}(n, \theta_k)$ for arbitrary K models, $\theta_k \sim \text{beta}(\alpha_k, \beta_k)$.
For $K = 30$, $n = 41$, $\alpha_k \in (0.5, 10)$ and $\beta_k \in (1, 20)$, Figure (5.6) shows the MSE of BPMSE, BMA and ABMA. As can be seen BMA does not dominate BPMSE, but ABMA does.

5.8.2 Evaluation with integrated risk

A good feature of integrated risk is that it allows a direct comparison of estimators (since it is a number). Consider a choice between the following models: $M_k : X \sim \text{binomial}(n, \theta_k)$ for arbitrary K models, $\theta_k \sim \text{beta}(\alpha_k, \beta_k)$, $n = 41$, $\alpha \in (1, 50)$, $\beta \in (2, 20)$. For each model (between 10 and 200), the integrated risk is computed and comparisons of estimators is given in Figure (5.7). The ABMA dominates BPMSE, BMA does not.

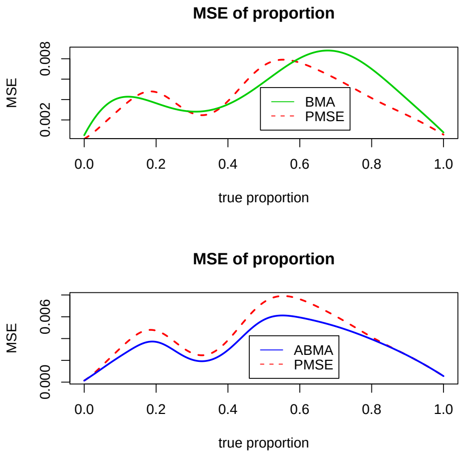

Figure 5.6: Risk of 30 full models comparing BPMSE, BMA and ABMA as a function of θ.

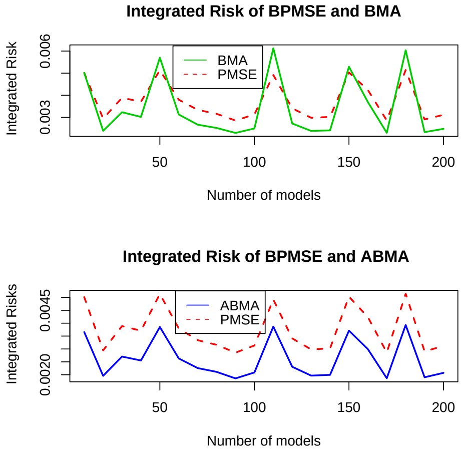

Figure 5.7: Integrated risks comparing BPMSE, BMA and ABMA as a function of the number of models.

Chapter 6

Model Selection and Nuisance Parameter Elimination

6.1 Introduction

Let x be a realization of a random variable X whose distribution is specified by the model M_t. Since M_t is unknown, one usually selects the "best" model from a set of plausible models, say $\mathcal{M} = (M_1, \ldots, M_K)$, i.e. one applies some model selection procedure (*between model selection*).

Consider now a parametric model that is represented by a family $M = \{f_\lambda, \lambda \in \Lambda\}$, and a quantity of interest of the form

$$\triangle = \triangle(X, \lambda). \tag{6.1}$$

$\triangle$ is a pivotal quantity for λ if its distribution does not depend on λ. Suppose that for a known value of the parameter, the properties of $\triangle$ are also known. In practice, λ is rarely known and one has to estimate it using an estimation method, e.g. maximum likelihood estimation. Let $\hat{\lambda}(X)$ be an estimator of λ. Suppose that $\hat{\lambda}(x) \in \Lambda$. This means that the estimate is a member of the parameter space. One can view the family M as a set of competing models. The estimation procedure is viewed as a model selection method and the selected model is $f_{\hat{\lambda}}$ (*within model selection*).

After the parameter has been estimated, the quantity of interest is then

$$\hat{\triangle} = \triangle(X, \hat{\lambda}). \tag{6.2}$$

One can see that a model selection procedure has been performed, followed by inference on the quantity $\hat{\triangle}$. Failure to take into account this selection procedure on the inference about $\hat{\triangle}$ will result in invalid inference. Therefore, the problem is similar to model selection uncertainty that we described in Chapter 3. In this context, we will refer to this as *parameter estimation uncertainty.* The problem is similar for any unknown λ that needs to be estimated; the parameter need not be related to the data generating process (e.g. bandwidth or binwidth selection in nonparametric density estimation).

We consider the problem in the framework of nuisance parameter elimination and compare the properties of $\triangle$ when the parameter is known to those when it is unknown. A typical example is the profile likelihood method, which is of the form (6.1) and then (6.2). We point out that, as long as the estimation of λ is properly taken into account, the procedure is correct. Since profile likelihood estimators can be biased, modified profile likelihood estimators are sometimes used to correct for the bias. However, such modifications can lead to inferior estimators in terms of risk measures, such as mean squared error.

The point of view presented here, of regarding parameter estimation as model selection, leads to an alternative interpretation of certain classical distributions, such as the t, F, Poisson, negative-binomial, beta-binomial and noncentral chi-squared distributions. We can interpret these as examples of model averaging to correct for model selection uncertainty. Finally, a definition of consistency is given for a "*within model selection procedure*".

6.2 Nuisance parameter elimination

In statistical problems, one tries to make inference about an unknown state of nature θ (possibly an entire set of parameters). The entire parameter set is rarely of interest. It is not uncommon to select a paramerisation $\theta = (\psi, \lambda)$ where ψ is the parameter of interest and λ is a incidental or nuisance parameter. The nuisance parameter accounts for aspects of the model that are not of main concern, but important for a realistic statistical modelling. In general, the parameter of interest ψ has small dimension while the nuisance parameter could be high dimensional. The presence of nuisance parameters sometimes makes inference difficult. Frequentist methods have been proposed for eliminating λ. These include marginal likelihood, conditional likelihood and profile likelihood.

In Bayesian analysis, the problem is not difficult, since it involves integrating the joint posterior with respect to λ and using the resulting marginal posterior of ψ. However, the problem arises in eliciting priors for λ and in implementating the procedure. When the likelihood function can be computed exactly, marginal and conditional likelihoods give exact inference for ψ. This often happens for exponential families.

For nonparametric or semi-parametric setting, Qin and Zhang (2005) introduce an empirical likelihood approach with application to genetical quantitative traits analysis.

6.2.1 Profile and modified profile likelihood

Let $L(\theta) = L(\psi, \lambda, x)$ be the likelihood function. If there is no nuisance parameter, one has to maximise the likelihood function in classical situation. The properties of maximum likelihood estimator are well known, in particular under certain conditions $\hat{\psi}$ is asymptotically normal. Suppose there is nuisance parameter. Let $\hat{\lambda}_\psi$ be the maximum likelihood estimate of λ when maximising the likelihood function over λ, ψ considered to be fixed. The profile likelihood is defined by $L_p(\psi) = L(\psi, \hat{\lambda}_\psi)$. Econometricians call the concentrated likelihood. Let $\hat{\psi}$ be the maximum likelihood estimator of ψ obtained by maximising the profile likelihood. The properties of $\hat{\psi}$ are well known in the literature, e.g. Pace and Salvan (1997). By taking the view that $\hat{\lambda}_\psi$ was chosen among an infinite class Λ of possible values (models), we see that one model is selected using the data and to make inference on the quantity of interest ψ. However, the difference is that usually, the moments (e.g. mean, variance) of the parameter of interest are computed (or approximated), since substitution is directly included in the estimate. As long as the properties of the profile likelihood estimator are available, the parameter estimation uncertainty does not exist, although in the literature the modified profile likelihood method is proposed.

Methods for dealing with general form of post-model-selection estimate are then applicable. From the literature, profile likelihood is too concentrated and can be maximised at a wrong value, asymptotic theory does not apply when the number of observations increase with the sample size. Many examples of misleading behaviour of profile likelihood are given in Reid (1988). Modified profile likelihood has been developed in Barndorff-Nielsen (1983,1988) and conditional profile likelihood in Cox and Reid (1987). Difficulties with modified profile include the fact that in some cases, modifications are not enough to establish consistency of the resulting estimators.

6.2.2 Illustrative examples

6.2.2.1 Univariate normal distribution

Let $x_1, \ldots, x_n$ be independent observations from $N(\mu, \sigma^2)$ and let σ^2 be the parameter of interest and μ the nuisance parameter. The maximun likelihood estimate of μ is $\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, then from the profile likelihood, $\hat{\sigma}^2 =$

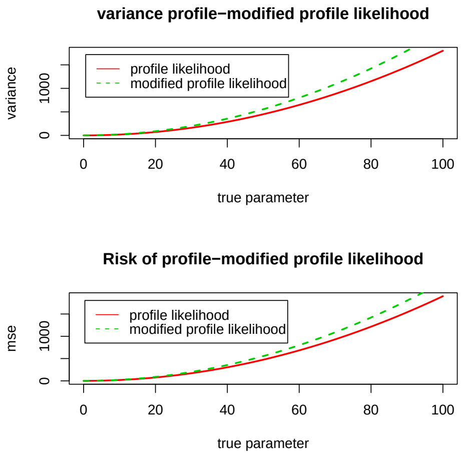

Figure 6.1: Variance and risk functions for the variance estimators using profile likelihood and modified profile likelihood as a function of the true parameter σ^2.

$\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$.
This substitution of μ by $\hat{\mu}$ represents model selection uncertainty. Simple use of modified profile likelihood yields $\hat{\sigma}^2_{mp} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$ which is an unbiased estimate.

$\mathrm{Var}(\hat{\sigma}^2_{mp}) = \frac{2\sigma^4}{n-1} = \mathrm{MSE}(\hat{\sigma}^2_{mp})$,

$\mathrm{Var}(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2} = \frac{(n-1)^2}{n^2}\mathrm{Var}(\hat{\sigma}^2_{mp})$,

$\mathrm{MSE}(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2}$.

Simple computations yield that $\mathrm{Var}(\hat{\sigma}^2_{mp}) \geq \mathrm{Var}(\hat{\sigma}^2)$ and $\mathrm{MSE}(\hat{\sigma}^2_{mp}) \geq \mathrm{MSE}(\hat{\sigma}^2)$, $\forall n \geq 2$ and $\forall \sigma > 0$. This means that although the modified profile likelihood estimator is unbiased, it is uniformly dominated by the profile likelihood estimator, even if, for large samples, both have the same risk. Illustration is given in Figure (6.1) for sample size $n = 10$. Although the act of replacing the unknown μ by its estimator $\hat{\mu}$ represents parameter estimation uncertainty, this causes no problem from the point of view of model selection. The reason is that the computations of the moments, e.g. mean and variances take into account this parameter estimation uncertainty (the moments of $\hat{\sigma}^2$ are known). In this case, the modified profile likelihood is therefore not necessary.

Figure (6.2) illustrates the behaviour of the estimation for the variance for different values of μ. The data used are flow data described in Linhart and

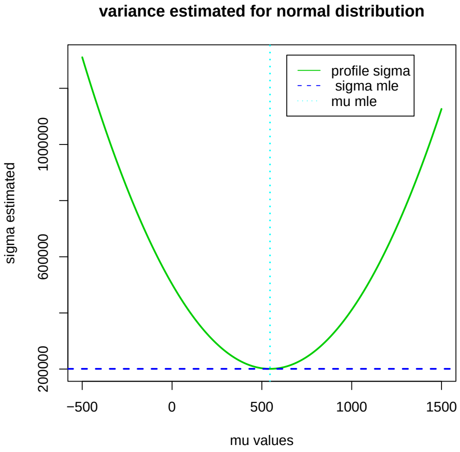

Figure 6.2: Variance estimation for various values of μ.

Zucchini (1986).

6.2.2.2 Multiple linear regression

Consider the normal linear model where $Y_i = x_i'\beta + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2)$ where $x_i' = (x_{i1}, \dots, x_{ip})$, a vector of p regressors, and β an unknown vector of dimension p and we have n observations. The parameter of interest is σ^2 and the nuisance parameter is β. The used profile likelihood yields $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - x_i\hat{\beta})^2$. This does not take into account the fact that the unknown β of dimension p has been estimated, leading in loss of degrees of freedom. The use of modified profile likelihood leads to $\hat{\sigma}^2_{mp} = \frac{1}{n-p}\sum_{i=1}^n (y_i - x_i\hat{\beta})^2$, which is unbiased and is also the estimate for residual sum of squares or the marginal likelihood of σ^2. However, the use of modified profile likelihood is unnecessary from the point of view of model selection since the properties of the profile likelihood $\hat{\sigma}^2$ are known.

6.2.2.3 Gamma models

Consider the gamma distribution with the following parametrisation

$g(x|\mu, \alpha) = \frac{\alpha^\alpha \mu^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\mu^{-1}\alpha x)$, y, α, $\mu > 0$.

Let $a = \frac{\alpha}{\mu}$, g is just the PDF of gamma distribution with scale a and shape α. Then $E(X) = \frac{\alpha}{a} = \mu$ and $V(X) = \frac{\alpha}{a^2} = \frac{\alpha}{a}$. Let the shape α be the parameter of interest and μ the nuisance parameter. The maximum likelihood estimate of μ is $\hat{\mu}_\alpha = \overline{x}$ and the profile likelihood is then given by $L_p(\alpha, \hat{\mu}_\alpha) = \Pi_{i=1}^n g(x_i|\hat{\mu}_\alpha, \alpha)$.

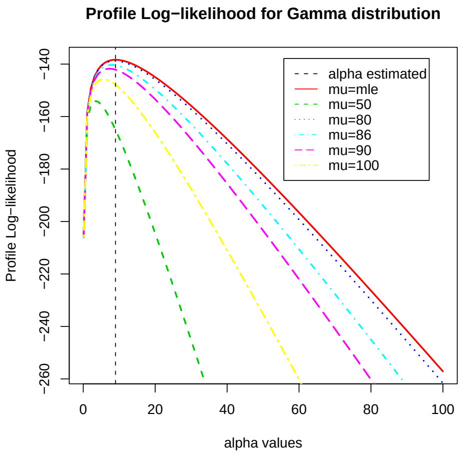

Figure 6.3: Profile log-likelihood for gamma distribution.

Figure (6.3) illustrates various choice of μ for the profile log likelihood and a particular choice of μ, $\hat{\mu}$ has been made. This means that a difference choice of log profile could have arisen if another set of data were used. Each choice of μ corresponds to a different value for $\alpha_{\max}$. This represents the uncertainty on plug-in $\hat{\mu} = 8.96$. The data used are storms data described in Chapter 2. The point is that even if, for some cases, the profile likelihood does not take into account the number of degrees of freedom lost in estimating the nuisance parameter, as long as the properties of the profile likelihood estimator are known, this does not suffer from parameter estimation uncertainty. Adjusted methods like the modified profile likelihood are unnecessary, unless they perform better than the profile likelihood estimator.

6.3 Deriving basic distributions

Let $X \sim f(x|\theta)$ a parameter family with Θ the parameter space (finite or infinite). This parameter can be any fixed value in the parameter space. Let $\hat{\theta}$ be an estimator of θ. For simplicity, assume that the range of the estimator is a subset of the parameter space. This means that the estimator takes on values in the parameter space, therefore, the event $\hat{\theta} = \theta$ is possible, since θ is fixed. For the model selection approach, the model space $\mathcal{M}$ is the same as parameter space Θ. The true model M_0 is then included in the set of models. This is a case where the true model really belongs to the set, and so there is no model misspecification. The estimation method for θ is now called model selection procedure, consisting of selecting one model in the model space $\mathcal{M}$. $\hat{\theta}$ is now denoted by $\hat{M}$. Now consider, $\triangle = \triangle(X, \theta)$ and replace $\triangle$ by an estimator $\hat{\triangle} = \triangle(X, \hat{\theta})$. One is selecting a model and making inference on the quantity of interest with the same data. One should then take into account the variability due to that. We recognise the parameter estimation uncertainty problem. Assume that for a fixed parameter value, the distribution of $\triangle$ is known and depends on θ, G and the distribution of $\hat{\theta}$ is also known, H and their densities by g and h. The idea is to obtain the distribution of $\hat{\triangle} = \triangle(X, \hat{\theta})$ by averaging over possible values of $\triangle$ using the density function of $\hat{\theta}$ as weights. This density function is referred as *parameter selection probabilities.* The advantage of this model averaging scheme is that there is no model misspecification. Let V and v be respectively the distribution function and density of $\hat{\triangle}$. Mathematically, this is given by

$$V(t) = \int_{\Theta} G(t, \theta)h(\theta)d\theta \qquad \text{or} \tag{6.3}$$

$$u(t) = \int_{\Theta} g(t, \theta)h(\theta)d\theta \qquad \text{if } g \text{ is continuous.} \tag{6.4}$$

To see that this distribution is well defined, it is straightforward to see that the quantity $V(t) = \int_{\Theta} G(t, \theta)h(\theta)d\theta$ is a monotone function of t, is increasing from 0 to 1, and is a distribution function. When the parameter space Θ is discrete or finite, the distribution of $\hat{\triangle}$ is given by

$$u(t) = \Sigma_{\Theta} g(t, \theta_k)h_k, \tag{6.5}$$

where $h_k = P(\hat{\theta} = \theta_k)$, model selection probability for model k, that is the probability that the estimator takes on a particular value θ_k.

6.3.1 Standard normal distribution and t-distribution

Let $X = (X_1, \dots, X_n)$ be a random sample from $N(\mu, \sigma^2)$ with σ unknown. Let drawing inference on μ be our objective. The parameter for estimation is then $\theta = \sigma$. Consider $\triangle(X, \theta) = \frac{\sqrt{n}(\overline{X} - \mu)}{\theta}$ where $\overline{X}$ is the sample mean. For θ known, $\triangle(X, \theta)$ is standard normal distributed. When θ is unknown, we want to find the distribution of $\triangle(X, \hat{\theta})$ where $\hat{\theta}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$. The estimator $\hat{\theta}$ can take any strictly positive value, in our terminology, we call each of possible value a model. We know that

$$\frac{(n-1)\hat{\theta}^2}{\theta^2} \sim \chi_{n-1}. \tag{6.6}$$

Let $Y = \frac{\hat{\theta}}{\theta}$. Then realisation of Y implies also realisation of $\hat{\theta}$. From (6.6), it follows that the distribution of Y is then given by $2pyf_p(py^2)$,where $p = n - 1$

and f_p is the density of chi-squared with p degrees of freedom. This distribution is the parameter selection probability. For a realisation y of Y, the distribution of $\triangle(X, \hat{\theta})$, using the transformation Y is given by $\phi(ty)y$ where ϕ is the PDF of standard normal. The distribution of $\triangle(X, \hat{\theta})$ is obtained by weighting this distribution by parameter selection probabilities. We have the following:

$u(t) = \int_0^\infty \{\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2y^2}{2}} y\}\{\frac{1}{\Gamma(\frac{p}{2})2^{p/2}}(py^2)^{p/2-1}e^{-py^2/2}2py\}dy$

$= \frac{2p}{\sqrt{2\pi}\Gamma(\frac{p}{2})2^{p/2}} \int_0^\infty y^2 e^{-\frac{t^2y^2}{2}}(py^2)^{p/2-1}e^{-py^2/2}dy$

$= \frac{2pp^{p/2-1}}{\sqrt{2\pi}\Gamma(\frac{p}{2})2^{p/2}} \int_0^\infty e^{-\frac{y^2}{2}(t^2+p)}(y^2)^{p/2}dy$, let $\delta = y^2$,

(1)$= \frac{p^{p/2}}{\sqrt{2\pi}\Gamma(\frac{p}{2})2^{p/2}} \int_0^\infty e^{-\frac{\delta}{2}(t^2+p)}\delta^{\frac{p+1}{2}-1}d\delta.$

Since $\int_0^\infty x^{\alpha-1}e^{-x/\beta}dx = \Gamma(\alpha)\beta^\alpha$, for $x > 0$, $\beta > 0$, $\alpha > 0$,

(1)$= \frac{p^{p/2}}{\sqrt{2\pi}\Gamma(\frac{p}{2})2^{p/2}}\Gamma(\frac{p+1}{2})(\frac{2}{t^2+p})^{\frac{p+1}{2}}$,

$= \frac{p^{p/2}}{\sqrt{2\pi}\Gamma(\frac{p}{2})2^{p/2}p^{\frac{p+1}{2}}}\Gamma(\frac{p+1}{2})(\frac{2}{1+t^2/p})^{\frac{p+1}{2}}$, therefore,

$$u(t) = \frac{\Gamma(\frac{p+1}{2})}{(p\pi)^{1/2}\Gamma(\frac{p}{2})(1+t^2/p)^{(p+1)/2}}, \qquad -\infty < t < \infty.$$

This yields, as expected, the PDF of t-distribution with p degrees of freedom. The mean is 0 and the variance is $\frac{p}{p-2}$. As expected the variance of t-distribution is larger than that of standard normal. This is the price to pay for estimation, although the variance tends to 1 as p increases.

Graphical illustration is given in Figure (6.4) with $p = 11$ degrees of freedom.

6.3.2 Equality of variance and F distribution

Let $X_1, \dots, X_n$ be a random sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_m$ be a random sample from an independent $N(\mu_2, \sigma_2^2)$. We are interested in comparing the variability of the two populations, namely, comparing the equality of variance. Let the quantity of interest be the ratio $\varrho = \frac{\sigma_2}{\sigma_1}$. Relevant information about this quantity is contained in $\hat{\rho} = \frac{\hat{\sigma}_2}{\hat{\sigma}_1}$ where $\hat{\sigma}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$ and $\hat{\sigma}_2 = \frac{\sum_{i=1}^m (Y_i - \overline{Y})^2}{m-1}$, $\overline{X}$ and $\overline{Y}$ are sample means. Consider the quantity: $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}\rho$. F can be written as $F = \triangle(Z, \hat{\theta})$ where $Z = (X_1, \dots, X_n, Y_1, \dots, Y_m)$ and $\theta = \sigma_2^2$, the parameter σ_1^2 is implicit and will vanish because $\frac{(n-1)\hat{\sigma}_1^2}{\sigma_1^2} \sim \chi^2_{n-1}$ and $\frac{(m-1)\hat{\sigma}_2^2}{\sigma_2^2} \sim \chi^2_{m-1}$. For each

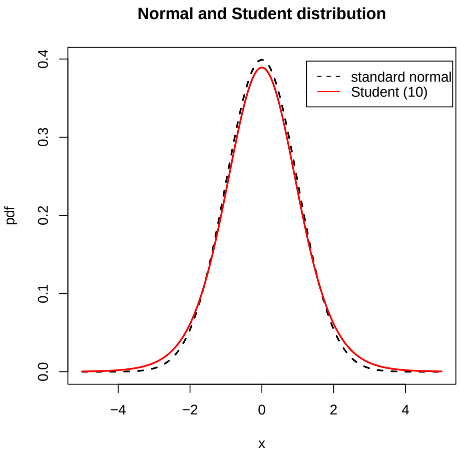

Figure 6.4: Densities comparing normal distribution (fixed parameter) with student (estimated parameter).

possible value of θ, that is for each model, the distribution of $\triangle(Z,\theta)$ is known and is for each $t>0$, $f_{n-1}(ty)y$ where f_{n-1} is the density of chi-squared with $n-1$ degrees of freedom and $y=\frac{\hat{\theta}}{\theta}$ and the weights are given by $(m-1)f_{m-1}((m-1)y)$, f_{m-1} is the density of chi-squared with $m-1$ degrees of freedom, corresponding to each possible value of the realisation of $\hat{\theta}$, that is each model. Let denoting $p=n-1$ and $q=m-1$. The density of $\triangle(Z,\hat{\theta})$ is then given by
$u(t)=\int_0^\infty\{pf_p(pty)y\}\{qf_q(qy)\}dy$
$=pq\int_0^\infty f_p(pty)yf_q(qy)dy$

$=pq\int_0^\infty y\frac{1}{\Gamma(\frac{p}{2})2^{p/2}}(pty)^{p/2-1}e^{-pty/2}\frac{1}{\Gamma(\frac{q}{2})2^{q/2}}(qy)^{q/2-1}e^{-qy/2}dy$

$=\frac{pqp^{p/2-1}t^{p/2-1}q^{q/2-1}}{\Gamma(\frac{p}{2})2^{p/2}\Gamma(\frac{q}{2})2^{q/2}}\int_0^\infty y^{(p+q)/2-1}e^{-qy/2(1+\frac{p}{q}t)}dy.$

Using the fact that $\int_0^\infty x^{\alpha-1}e^{-x/\beta}dx=\Gamma(\alpha)\beta^\alpha$, for $x>0$, $\beta>0$, $\alpha>0$,

$u(t)=\frac{pqp^{p/2-1}t^{p/2-1}q^{q/2-1}}{\Gamma(\frac{p}{2})2^{p/2}\Gamma(\frac{q}{2})2^{q/2}}\Gamma(\frac{p+q}{2})(\frac{1}{q/2(1+\frac{p}{q}t)})^{(p+q)/2}$

$=\frac{p^{p/2}t^{p/2-1}q^{q/2}2^{(p+q)/2}}{\Gamma(\frac{p}{2})2^{(p+q)/2}\Gamma(\frac{q}{2})q^{p/2}q^{p/2}}\Gamma(\frac{p+q}{2})(\frac{1}{1+\frac{p}{q}t})^{(p+q)/2}$

$=\frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})}(\frac{p}{q})^{p/2}\frac{t^{p/2-1}}{(1+\frac{p}{q}t)^{(p+q)/2}},\ t>0.$

We recognise this expression as the distribution function of $F(p,q)$, Fisher's F distribution with p and q degrees of freedom. Since, the $u(t)$ is independent of

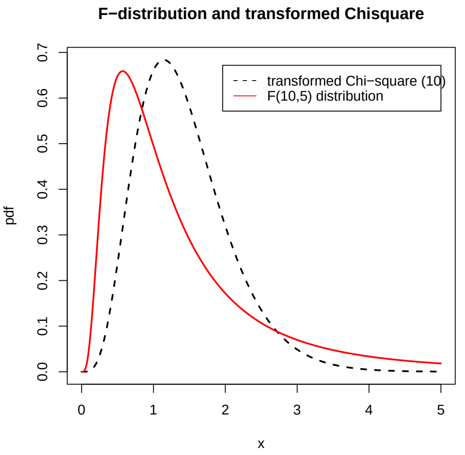

Figure 6.5: Densities comparing transformed chi-squared (fixed parameter) with F (estimated parameter).

θ, $\triangle(Z,\hat{\theta})$ is in fact a pivotal quantity for θ. Graphical illustration is given in Figure (6.5) with $p = 10$, $q = 5$ and $y = 7$.

6.3.3 Poisson distribution as weighted binomial

For a known θ^*, $\triangle^* = \triangle(X,\theta^*) \sim \text{binomial}(\theta^*, p)$ and $\hat{\theta}$ as estimator of θ follows Poisson(λ), model selection probabilities. The probability density function of $\hat{\triangle} = \triangle(X,\hat{\theta})$ is given by:

$$u(t) = \textstyle\sum_{\theta=0}^{\infty}\{\binom{\theta}{t}p^t(1-p)^{\theta-t}\}\{\frac{e^{-\lambda}\lambda^{\theta}}{\theta!}\},$$

$$= \tfrac{e^{-\lambda}(\lambda p)^t}{t!}\textstyle\sum_{\theta=t}^{\infty}\tfrac{((1-p)\lambda)^{\theta-t}}{(\theta-t)!},$$

Since $\triangle^* = \triangle(X,\theta^*) \sim \text{binomial}(\theta^*, p)$ for $\theta > t$ otherwise the value of the probability is 0.

$$u(t) = \tfrac{e^{-\lambda}(\lambda p)^t}{t!}\textstyle\sum_{\theta=t}^{\infty}\tfrac{((1-p)\lambda)^{\theta-t}}{(\theta-t)!}, \text{ let } z = \theta - t,$$

$$= \tfrac{e^{-\lambda}(\lambda p)^t}{t!}\textstyle\sum_{z=0}^{\infty}\tfrac{((1-p)\lambda)^{z}}{z!}, \text{ given that } \textstyle\sum_{x=0}^{\infty}\tfrac{a^x}{x!} = e^a,$$

$$= \tfrac{e^{-\lambda}(\lambda p)^t}{t!}e^{(1-p)\lambda} = \tfrac{e^{-\lambda}(\lambda p)^t}{t!}e^{-p\lambda}, \qquad t = 0, 1, \ldots$$

Therefore, $\triangle(X,\hat{\theta}) \sim \text{Poisson}(\lambda p)$. Graphical illustration is given in Figure (6.6) with $\lambda = 10$, $\theta^* = 15$ and $p = 0.5$.

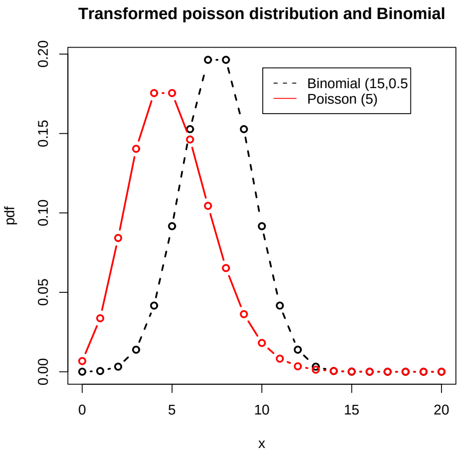

Figure 6.6: Densities comparing binomial (fixed parameter) with transformed poisson (estimated parameter).

6.3.4 Negative binomial as weighted gamma and Poisson distribution

$\triangle^* = \triangle(X, \theta^*) \sim$ Poisson(θ^*) and $\hat{\theta} \sim$ gamma(α, β), with α an integer, model selection probabilities.
The distribution of $\hat{\triangle}$ is given by:

$u(t) = \int_0^\infty \{\frac{e^{-\theta}\theta^t}{t!}\}\{\frac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}e^{-\theta/\beta}\}d\theta$

$= \frac{1}{t!\Gamma(\alpha)\beta^\alpha}\int_0^\infty \theta^{t+\alpha-1}e^{-\theta(1+1/\beta)}d\theta$, using $\int_0^\infty x^{\alpha-1}e^{-x/\beta}dx = \Gamma(\alpha)\beta^\alpha$, for $x > 0$, $\beta > 0$, $\alpha > 0$,

$=\frac{\Gamma(t+\alpha)(\frac{1}{1+1/\beta})^{t+\alpha}}{t!\Gamma(\alpha)\beta^\alpha} = \frac{\Gamma(t+\alpha)(\frac{1}{1+1/\beta})^{t}(\frac{1}{1+1/\beta})^{\alpha}}{t!\Gamma(\alpha)\beta^\alpha}$

$= \frac{\Gamma(t+\alpha)(\frac{\beta}{1+\beta})^{t}(\frac{1}{1+\beta})^{\alpha}\beta^\alpha}{t!\Gamma(\alpha)\beta^\alpha} = \frac{\Gamma(t+\alpha)(\frac{\beta}{1+\beta})^{t}(\frac{1}{1+\beta})^{\alpha}}{t!\Gamma(\alpha)}.$

Let $\pi = \frac{1}{1+\beta}$, $r = \alpha$, the distribution of $\hat{\triangle}$ is then given by:
$u(t) = \binom{t+r-1}{t}\pi^r(1-\pi)^t$, $t = 0, 1, 2, \ldots$, $0 < \pi < 1$, $r > 0$.
We recognise this as the PDF of negative binomial with parameter π and r. Special case include exponential(β) which is gamma$(1, \beta)$ and chi-squared with p degrees of freedom which is gamma$(\alpha = p/2, \beta = 2)$. Graphical illustration is given in Figure (6.7) with $\theta^* = 10$, $r = \alpha = 25$, $\beta = 2$.

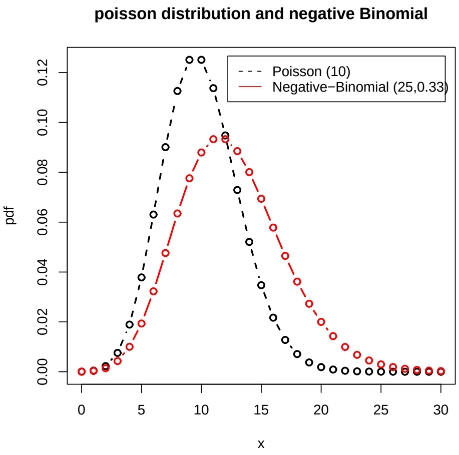

Figure 6.7: Densities comparing Poisson (fixed parameter) with negative binomial (estimated parameter).

6.3.5 Beta-binomial distribution as weighted beta and binomial distributions

$\triangle^* = \triangle(X, \theta^*) \sim \text{binomial}(n, \theta^*)$ and $\hat{\theta} \sim \text{beta}(\alpha, \beta)$, model selection probabilities. The distribution of $\hat{\triangle} = \triangle(X, \theta)$ is given by:

$u(t) = \int_0^1 \{\binom{n}{t}\theta^t(1-\theta)^{n-t}\}\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\}d\theta$

$= \binom{n}{t}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{t+\alpha-1}(1-\theta)^{n-t+\beta-1}d\theta.$

We have that $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

$u(t) = \binom{n}{t}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(t+\alpha)\Gamma(n-t+\beta)}{\Gamma(n+\alpha+\beta)}$, $t = 0, 1, \ldots, n$.

This is the PDF of beta-binomial(n, α, β).
Graphical illustration is given in Figure (6.8) with $n = 15$, $\theta^* = 0.5$, $\alpha = 3$ and $\beta = 5$.

6.3.6 Noncentral chi-squared as weighted central chi-squared and Poisson distributions

$\triangle^* = \triangle(X, \theta^*) \sim \chi_{p+2\theta^*}$ and $\hat{\theta} \sim \text{Poisson}(\lambda)$, model selection probabilities. The distribution of $\hat{\triangle} = \triangle(X, \hat{\theta})$ is given by:

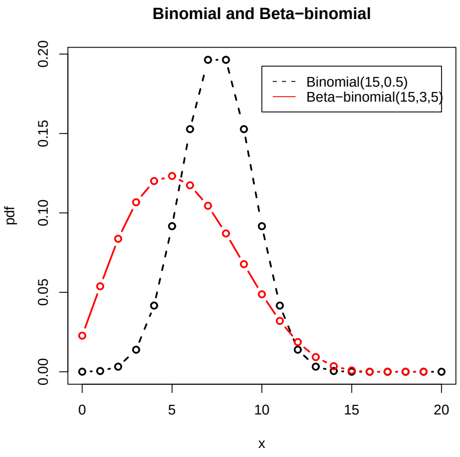

Figure 6.8: Densities comparing binomial (fixed parameter) with beta-binomial (estimated parameter).

$$u(t) = \sum_{\theta=0}^{\infty}\{\frac{1}{\Gamma(\frac{p+2\theta}{2})2^{\frac{p+2\theta}{2}}}t^{(\frac{p+2\theta}{2})-1}e^{-t/2}\}\{\frac{e^{-\lambda}\lambda^{\theta}}{\theta!}\},$$

$$u(t) = \sum_{\theta=0}^{\infty}\{\frac{1}{\Gamma(\frac{p}{2}+\theta)2^{\frac{p}{2}+\theta}}t^{(\frac{p}{2}+\theta)-1}e^{-t/2}\}\{\frac{e^{-\lambda}\lambda^{\theta}}{\theta!}\},\ 0 < t < \infty\ .$$

We recognize u as PDF of noncentral chi-squared with p degrees of freedom and noncentrality parameter λ. Graphical illustration is given in Figure (6.9) with $\theta^{*} = 1$, $\lambda = 5$ and $p = 10$.

Finally, it is important to note that the parameters λ or θ need not be necessarily associated to a family of distributions. An example is the density (histogram and kernel estimation) and nonparametric regression where the binwidth and the bandwidth are selected from the data. When a data-dependent binwith or bandwidth is used, one should take into account in computing the properties of the estimated density.

6.4 Consistency of a “within model selection criterion”

This section introduces the notion of consistency of a within model selection criterion (estimation inside a parametric model). Namely, estimation is viewed as model selection problem and an estimation method is a model selection criterion.

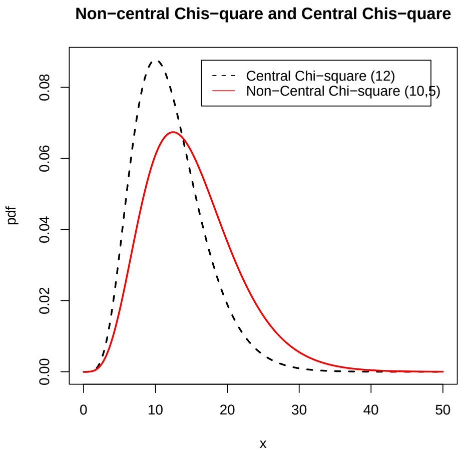

Figure 6.9: Densities comparing central chi-squared (fixed parameter) with non-central (estimated parameter).

A within model selection procedure (e.g. MLE) is consistent if

$$\lim_{n\to\infty} P(F_{\hat{\theta}}(t) = F_{\theta_0}(t)) = 1, \forall t \in \mathbb{R}, \text{i.e.} \lim_{n\to\infty} P(\hat{\theta} = \theta_0) = 1, \tag{6.7}$$

where $F_{\hat{\theta}}$ is estimated distribution function and F_{θ_0} the true distribution function. However, for Θ infinite, (6.7) is difficult to achieve (since the probability could tend to 0), therefore there is a need for another definition.
The only requirement for consistency will be that for large samples size, the two distributions are close enough. A within model selection procedure is:

1. Pointwise weakly consistent if
$$\forall \epsilon > 0, \quad \lim_{n\to\infty} P(d(F_{\hat{\theta}}(t), F_{\theta_0}(t)) < \epsilon) = 1, \quad \forall t \in \mathbb{R}.$$
2. Pointwise strongly consistent (almost sure pointwise) if
$$P(\lim_{n\to\infty} F_{\hat{\theta}}(t) = F_{\theta_0}(t)) = P(\lim_{n\to\infty} d(F_{\hat{\theta}}(t), F_{\theta_0}(t)) = 0) = 1, \quad \forall t \in \mathbb{R}.$$
3. Uniformly consistent (almost sure) if
$$P(\lim_{n\to\infty} \{\sup_{t\in\mathbb{R}} d(F_{\hat{\theta}}(t), F_{\theta_0}(t))\} = 0) = 1. \tag{6.8}$$

E.g. Glivenko-Cantelly theorem with $F_{\hat{\theta}}$ be the empirical CDF and $d = ||$ in (6.8). It is important to note the difference between classical consistency of an estimator and consistency of an estimation procedure, viewed as model selection criterion. For within consistency, only the convergence of the distribution function is of concern.

Chapter 7

Bootstrap after Model Selection

7.1 Introduction

This chapter is concerned with potential application of bootstrap methods in the context of model selection. We first ask ourself whether bootstrap methods, which are applied to solve a number of mathematically intractible problems, can also be used to assess the properties of post-model-selection estimators (PMSEs). We explain that this is different, and more complex, than the normal use of bootstrap. In general PMSEs are highly variable, biased, and their distribution is a multimodal. The bootstrap estimates of the properties of PMSEs reflect these properties. This might tempt one to believe that the bootstrap is an appropriate method here. However, by means of a concrete theoretical example, we illustrate that the bootstrap can provide poor estimators of the properties of PMSEs. We identify the reason for the failure as follows: the bootstrap can be an inaccurate estimator of the model selection probabilities.

There is an additional issue concerning the existence of certain properties, such as the moments, of PMSEs. For some selection procedures, e.g. forward variable selection, it is not clear that the moments even exist. Of course the bootstrap will always supply an estimate of the moments of PMSEs, even in cases where these do not exist. For informal selection criteria such as exploratory data analysis, bootstrap may not be applicable.

7.2 The complexity of PMSEs

The PMSEs depend on the selection process S and the set $\mathcal{M}$ of competing models. The difficulty of computing the properties of these estimators therefore depends on S and $\mathcal{M}$. The complexity increases with the complexity of model selection procedure and the dimension of $\mathcal{M}$. The general probabilistic framework we described does not mention the difficulty due to the model selection procedure.

Whereas in theory, in this framework, the distribution and the moments of PMSEs can be derived, there are many types of selection process where it may not be possible. By selection process, we mean any method leading to a choice of a model. We do not necessarily mean only parsimonious selection procedures. For instance, consider the following iterative model selection procedure: identification, parameter estimation, model diagnostic. If the model does not fit, we reconsider another model until we obtain a "good" model. It is really difficult to derive the properties of the resulting post-model-selection estimator. Other complex procedures include: multiple testing, any iterative process, exploratory data analysis (EDA), in regression analysis (forward selection, backward elimination, F-to-enter). In such complex selection procedures, the problem becomes more difficult.

A natural question may be: why not try computer intensive methods? Consider the PMSE $\tilde{\theta}(X|S,\mathcal{M})$ and the naive estimator $\hat{\theta}_{\tilde{k}}(\mathrm{X})$ of θ. It is important to recognise that after data have been observed, these two estimates are equal, that is $\tilde{\theta}(x|S,\mathcal{M}) = \hat{\theta}_{\tilde{k}}(x)$. The post-model-selection estimate and naive estimate are the same. We have then two random variables that yield the same estimate. This means that in practice, after obtaining estimate of the naive model, one needs valid inference. The problem is then to obtain properties of $\tilde{\theta}(X|S,\mathcal{M})$. Moments of $\tilde{\theta}(X|S,\mathcal{M})$ involve knowlege of the unknown true model. We turn then to the problem of making inference on a complex random variable. There are two general approaches of evaluating the variability of an estimate: The traditional analytic approach and the resampling approach. For the analytic approach, the variability of $\tilde{\theta}(X|S,\mathcal{M})$ is evaluated by deriving an explicit theoretical formula that approximates its distribution or other characteristics such as second order moments. If this theoretical formulae contains unknown quantities like θ, one can substitute the unknown parameter by its estimate. For PMSEs, the theoretical formula is really difficult to derive. In the text, we will maintain the notation with S and $\mathcal{M}$ in order to distinguish classical bootstrap and model selection bootstrap.

7.3 Bootstrap model selection

Consider an unknown probability distribution F that has given the observed data $x = (x_1, \ldots, x_n)$ by random sampling. Suppose that we want to estimate a parameter of interest $\theta = h(F)$ based on x, say $\hat{\theta} = v(x)$. The natural question is about the accuracy of such estimate. The bootstrap was introduced by Efron (1979) as a computer based method for estimating the standard error of $\hat{\theta}$. The key idea is to resample from the original data, directly or by a fitted model to recreate replicates of x, and then assess the variability of $\hat{\theta}$. It does not require theoretical calculations and is possible regardless how mathematically complicated $\hat{\theta}$ (or the mapping v) is. Bootstrap methods can then be used to produce statistical inferences about an unknown quantity.

The estimate of the quantity of interest θ is given by

$$\tilde{\theta}(x|S,\mathcal{M}) = v(x|S,\mathcal{M}) = \sum_{k=1}^{K} I_k(x|S,\mathcal{M}) v_k(x), \tag{7.1}$$

where $\hat{\theta}_k = v_k(x)$, v_k being a mapping generating each estimator in each model, not necessarily the same mapping for each model. The mapping v can be quite complicated as it depends on the selection procedure S and the set of model under consideration. Analytical expression of the distribution of $\tilde{\theta}(x|S,\mathcal{M})$ as well as its moments are difficult to formalise due mostly to the fact that the $I_k(X|S,\mathcal{M})$ are likely to be correlated with the $\hat{\theta}_j$. Let x^{*j} and $\hat{\theta}^{*j}$ be respectively the j^{th} bootstrap sample and bootstrap estimate. $\hat{\theta}^{*j}$ is expressed as

$$\tilde{\theta}^{*j}(x^{*j}|S,\mathcal{M}) = v(x^{*j}|S,\mathcal{M}) = \sum_{k=1}^{K} I_k(x^{*j}|S) v_k(x^{*j}). \tag{7.2}$$

The bootstrap algorithm is described as follows.

1. Select m independent bootstrap samples $x^* = (x^{*1}, \dots, x^{*m})$, each consisting of n data values drawn with replacement from x (nonparametric bootstrap) or drawing m samples of size n from the parametric estimate $\hat{F}_{para}$, where $\hat{F}_{para}$ is an estimate of F derived from a parametric model for the data (parametric bootstrap).

2. Evaluate the bootstrap replication corresponding to each bootstrap sample, $\tilde{\theta}^{*j}(x^{*j}|S,\mathcal{M})$, $j = 1, \dots, m$.

3. Estimate the standard error and bias of $\tilde{\theta}(x|S,\mathcal{M})$.

$$\begin{aligned} \hat{se}(\tilde{\theta}(x|S,\mathcal{M})) &= \{\tfrac{1}{m-1}\textstyle\sum_{j=1}^{m}[\tilde{\theta}^{*j}(x^{*j}|S,\mathcal{M}) - \tilde{\theta}^{*}(x^{*}|S,\mathcal{M})]^2\}^{1/2}, \\ \hat{\text{bias}}(\tilde{\theta}(x|S,\mathcal{M})) &= \tilde{\theta}^{*}(x^{*}|S,\mathcal{M}) - \tilde{\theta}(x|S,\mathcal{M}), \end{aligned} \tag{7.3}$$

where $\tilde{\theta}^{*}(x^{*}|S,\mathcal{M}) = \frac{1}{m}\sum_{j=1}^{m}\tilde{\theta}^{*j}(x^{*j}|S,\mathcal{M})$.

The bootstrap can also be used for constructing confidence interval. For the classical situation, suppose that $\hat{\theta} \sim N(\theta, \sigma^2)$ with known standard error, the standard confidence interval with coverage probability $1 - \alpha$ is

$$[\hat{\theta} - \phi_{1-\alpha}\sigma, \hat{\theta} - \phi_{\alpha}\sigma],$$

where ϕ_α indicates the $100\alpha^{th}$ percentile point of standard normal distribution. However, the assumption of normality fails in many cases. $\tilde{\theta}(x|S, \mathcal{M})$ is likely to be a typical case. An approximate $1 - \alpha$ percentile interval is

$$[\tilde{\theta}^*(x^*|S, \mathcal{M})_{\alpha/2}, \tilde{\theta}^*(x^*|S, \mathcal{M})_{1-\alpha/2}], \tag{7.4}$$

where $\tilde{\theta}^*(x^*|S, \mathcal{M})_\beta$ is the $100\beta^{th}$ empirical percentile of bootstrap samples. As one can see, the problem is more difficult than classical situation in that it involves model selection step in each resample. The important issue is whether this bootstrap has theoretical justification. More work is required to find conditions under which bootstrap could work in this case.

7.4 Properties of bootstrap model selection

The validity of bootstrap is usually established by showing that the conditional distribution of $\tilde{\theta}^*(X^*|S, \mathcal{M})$ given X is approximatively the same as the distribution of $\tilde{\theta}(X|S, \mathcal{M})$. Unlike the classical bootstrap with only one model, the first difficulty here is that the true model is not known, that is the model from which the sampling was performed. One can only assume a true model in order to establish the validity of bootstrap and there is still uncertainty about this choice. Let D be a "distance" measure, and $\hat{G}$ be the distribution of $D(\tilde{\theta}(X|S, \mathcal{M}), \theta)$, and G^* conditional distribution of $D^*(\tilde{\theta}^*(X^*|S, \mathcal{M}), \theta)$ given X, then $G \simeq G^*$. In terms of probability, let P, E and P^*, E^* be denoting probabilities and expectation under the true distribution F and its estimate $\hat{F}$, for n large, we have

$$\|P(D(\tilde{\theta}(X|S, \mathcal{M}), \theta) \leq t) - P^*(D^*(\tilde{\theta}^*(X^*|S, \mathcal{M}), \tilde{\theta}(X|S, \mathcal{M})) \leq t)\| \longrightarrow_{a.s} 0,$$

$\|.\|$ being an appropriate norm. The problem here is that it will be difficult to show that $\tilde{\theta}(X|S, \mathcal{M}))$ converges almost surely to θ, or even to look for conditions under which this convergence is possible. The difference between classical bootstrap and model selection bootstrap is that bootstrap for classical situation is more stable than in model selection since estimators in the later case are expected to come from various models. This is the reason why the variability is likely to be higher.

7.5 Naive bootstrap approximation for model selection

The naive approach involves working with $\tilde{M}(X|S) = M_{\tilde{k}}$ or in the bootstrap world $\tilde{M}^*(X^*|S) = M^*_{\tilde{k}}$, in other words, no selection procedure is included in the bootstrap estimates. When only bootstrap is used to get properties of the naive model, it is possible to compare both estimators. We conjecture that these two approaches are equivalent in the following circonstances:

1. For $\tilde{M}^{*j}(x^{*j}|S) = M^*_{\tilde{k}}$, $j = 1, \ldots, m$, that is the selected model is the same for each resample, then $\tilde{\theta}^{*j}(x^{*j}|S, \mathcal{M}) = \tilde{\theta}^*(x^*|S, \mathcal{M})$ and is the naive model.

2. When estimates for different selected models are very closed for each resample.

In either of the scenarios (1) and (2), the bootstrap variance of the PMSE is close to the naive variance. It is important to know that only bootstrap properties of $\tilde{\theta}(X|S, \mathcal{M})$ and $\tilde{\theta}_{\tilde{k}}(X)$ can be compared. For e.g., theoretical properties of naive estimators can not be compared to bootstrap properties of PMSEs.

7.6 Failure of post-model-selection bootstrap

7.6.1 Bootstrap model selection probability estimates

To mimic the bootstrap world to that of real world, our set probability framework is applicable in each bootstrap sample. Namely, the probabilistic argument should apply, that is the notion of partitioning bootstrap sample space into $\mathcal{X}^*$ for model M_k. An important ingredient there is $P^*(\mathcal{X}^*_k)$, the probability of landing on sample space $\mathcal{X}^*$. This probability can be viewed as assessing the uncertainty of the selected model. This can be interpreted as a confidence level for the model in supporting the data and is obtained as

$$P^*(\mathcal{X}^*_k) = P^*(\tilde{M}^*(X^*|S, \mathcal{M}) = M_k) = E^*(I_k(X^*|S, \mathcal{M})). \tag{7.5}$$

This probability aims to estimate model selection probability. An example of failure of bootstrap to converge to the same distribution like PMSEs is given in Hjort and Claeskens (2003), p.897 for testing a model with mean 0 against a model with non-null mean for normal distribution. From the probability framework, finite sample distribution of PMSEs are mixtures of conditional distribution by model selection probabilities. For e.g., Knight (1999) provides an example in linear regression where residual bootstrapping fails. This example of failure of bootstrap for model selection will result on the failure bootstrap to accurately estimate model selection probabilities. Therefore the failure of bootstrap to accurately approximate the finite sample distribution. In general, the failure of bootstrap in classical situation (without model selection) will result in a failure when taking model selection into account.

7.6.2 One-way ANOVA theoretical example

Consider the one way ANOVA situation where data Y_{ij} are observed according to model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \qquad i = 1, \dots, k, \qquad j = 1, \dots, n_i, \tag{7.6}$$

where the θ_i are unknown parameters and ϵ_{ij} are errors.
Assumptions are the following

1. $\mathrm{E}\epsilon_{ij} = 0$, $\mathrm{Var}\epsilon_{ij} = \sigma_i^2 < \infty$, for all i, j. $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ for all i, i', j, and j' unless $i = i$'and $j = j$'; the ϵ_{ij} are independent.

2. $\sigma_i^2 = \sigma^2$ for all i (homoscedasticity).

Note that we do not assume normality. For simplicity, let assume $k = 2$, $n = n_1 + n_2$. Suppose that our parameter of interest is θ_1 and we perform a test

$H_0 : \theta_1 = \theta_2$ (restricted model M_0) against $H_0 : \theta_1 \neq \theta_2$ (unrestricted model M_1).

For all i, let $\hat{\theta}_i = Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$, the sample mean in each population;

$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2$, the sample variance for each population;

and the pooled variance by: $S_p^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) S_i^2$;

$\hat{\theta}_p = \frac{1}{n} \sum_{i=1}^{k} n_i \hat{\theta}_i$ a pooled estimate.

Consider the following selection procedure
If H_0 is rejected, use $\hat{\theta}_1$ as an estimate of θ_1 otherwise use the pooled estimate. The selection procedure S is pre-test and $\mathcal{M} = \{M_0, M_1\}$. The test statistic is given by $T = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{S_p^2 (\frac{1}{n_1} + \frac{1}{n_2})}}$.
The PMSE $\tilde{\theta}_1$ is given by

$$\tilde{\theta}_1 = \hat{\theta}_1 I_1 (T < Z_{1-\frac{\alpha}{2}}) + \hat{\theta}_p I_p (T \geq Z_{1-\frac{\alpha}{2}}), \tag{7.7}$$

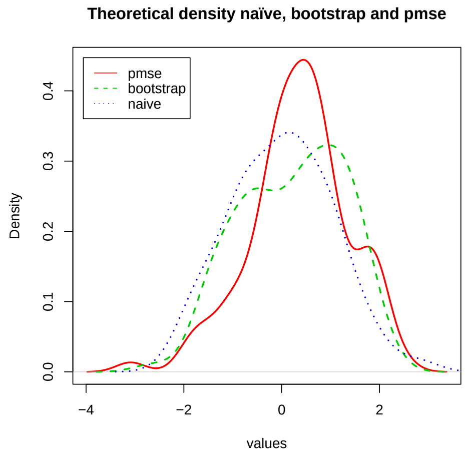

Figure 7.1: Densities of naive, PMSE and bootstrap estimator.

where I_1 and I_p are respectively indicator functions for the choice of θ_1 and θ_p, and $Z_{1-\frac{\alpha}{2}}$ is the $100(1-\frac{\alpha}{2})^{th}$ the quantile of N(0,1).
We state the following theorem proved by Kulperger and Ahmed (1992), p.2076.

Theorem 7.6.1 *Under the assumptions*

1. *$n_1/n = n_2/n$ converges to $\tau \in (0, 1)$,*

2. *$\theta_2^{(n)} = \theta_1 + \varrho/\sqrt{n}$, ϱ is a fixed number, (local alternatives),*

*then $\sqrt{\tau n}(\tilde{\theta}_1 - \theta_1)$ tends in distribution to $H(Y_1, Y_2, \tau, \nu)$,
where $H(Y_1, Y_2, \tau, \nu) = \sigma Y_1 + \sigma\sqrt{1-\tau}S(\tau, \nu)I(|S(\tau, \nu)| < Z_{1-\frac{\alpha}{2}})$.
As n becomes larger, and Y_1 and Y_2 are independent standard normal random variables and $S(\tau, \nu) = \sqrt{\tau}Y_2 - \sqrt{1-\tau}Y_1 + \sqrt{\tau(1-\tau)}\nu$ and $\nu = \varrho/\sigma$.*

The distribution of the bootstrap estimator is given in Kulperger and Ahmed (1992), p.2080.

Theorem 7.6.2 *Under the assumptions stated in Lemma 1 in Kulperger and Ahmed (1992), p.2077,*

$$\sqrt{\tau n}(\tilde{\theta}_1 - \hat{\theta}_1) \text{ tends in distribution to } H(Z_1, Z_2, \tau, \frac{Y_2}{\sqrt{1-\tau}} - \frac{Y_1}{\sqrt{\tau + \nu}}), \quad (7.8)$$

where Y_1, Y_2, Z_1 and Z_2 are independent standard normal.

But the main issue is to compare the naive approach to the bootstrap and Figure (7.1) illustrates how these three estimators perform. One can see that the density of bootstrap-after-model-selection estimator is not close to that of PMSE. The key issue on the validity of bootstrap is the consistency of the PMSEs, which is difficult to check. However, Kilian (1998) advocates the success of bootstrap procedure for selecting the order of a autoregressive models. Alonso et al. (2004) introduce bootstrap methods for accounting for the variability for autoregressive order selection. Bootstrap model selection estimated probabilities and how they can be used to weighting competing models are discussed in Buckland et al. (1997), Burnham and Anderson (2002). Although warning of no guarantee of the use of bootstrap, we give some practical issues described in the literature.

7.7 Practical issues

As one may expect, practical difficulties in such bootstrap can be enormous. Shao (1996) studies bootstrap model selection and shows that in linear regression, the bootstrap procedure is inconsistent in the sense that the probability of selecting the optimal subset does not converge to 1 as the sample size grows. This means that straightforward application of the model selection procedure is not consistent. He corrects this inconsistency by modifying the sampling method.

For pairs bootstrap, he proposes to sample less than n observations and this modified bootstrap becomes consistent if the bootstrap observations $m \to \infty$ and $m/n \to 0$. Changing the bootstrap sample to correct for consistency has been shown to be successful in other problems by Hall (1990) and Swanepoel (1986). The choice of m remains difficult and Shao (1996) recommends to choose as to minimise the length of the bootstrap confidence interval. For residuals bootstrap, Shao (1996) modifies the sample procedure by increasing the variability among bootstrap observations and suggests to multiply the residuals by a factor $\sqrt{n/m}$ where $m \to \infty$ and $m/n \to 0$.

Shao (1996) generalises the results to more complex models as nonlinear regressions, generalised linear models, and autoregressive time series. It is important to note that bootstrapping residuals is more appropriate for nonstochastic regressors and pairs bootstrap are suitable for stochastic regressors. However, in the case of residual bootstrap, the model form which the residuals are obtained will be more often selected. Nonparametric bootstrap assumes that observations (or residuals) are independent and identically distributed. For generalised linear models, the variance depends on the expectation, Buckland et al. (1997) propose a parametric bootstrap.

7.8 An illustrative example

Let consider again the estimation of design storms. Now, we use model selection criterion. As noted in Linhart and Zucchini (1986), it is necessary to choose a model selection criterion that takes into account the upper tail of the distribution. The selection criterion should also depend on the design horizon to take into account the relevant portion of the distribution. They use the following model selection criterion (discrepancy)

$$d(F_\lambda, G; h) = max\{|G(x)^h - F_\lambda(x)^h| : \qquad x \in \mathbf{R}\}. \tag{7.9}$$

Since the true G is unknown, the following empirical discrepancy were used

$$d_n(F_\lambda, G; h) = max\{|(\frac{i}{n+1})^h - F_\lambda(x_{(i)})^h| : i = 1, \ldots, n\}, \tag{7.10}$$

where the $x_{(i)}$'s are the order statistics and the design horizon were chosen to be 1, 5, 10.

This is a situation where one can not apply standard selection criteria like AIC. The discrepancy of model and the selected model for each horizon are given in Table (7.1). These discrepancies are closed to those computed with 200 bootstrap replications given in Linhart and Zucchini (1986). For the full data set, selected model for AIC is Gumbel.

For $h = 1$, selected model is lognormal, Weibull for $h = 5$ and exponential for $h = 10$. Table (7.2) compares model selection approach with the bootstrap of the naive model for each horizon. Table (7.3) gives bootstrap estimated of model selection probabilities. For $h = 1$, the selected model is lognormal, one would expect that it would be the model with higher estimated probability. But the Gumbel is the model with highest probability. For $h = 2$, the model with highest

Distribution	$h = 1$	$h = 5$	$h = 10$
gamma	0.13	0.31	0.47
normal	0.15	0.32	0.49
lognormal	**0.12**	0.30	0.45
exponential	0.40	0.32	**0.39**
Weibull	0.15	**0.29**	0.40
Gumbel	0.13	0.32	0.49
Selected Model	lognormal	Weibull	exponential

Table 7.1: Discrepancies and selected models for each horizon.

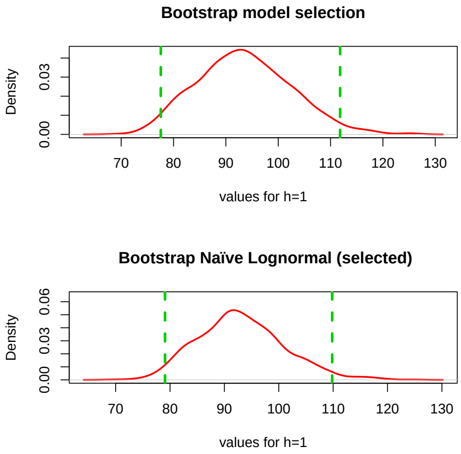

Figure 7.2: Comparison density of naive lognormal and bootstrap model selection for $h = 1$ with 0.95 percentile intervals.

estimated probability is lognormal, whereas the selected model is Weibull. For $h = 5$, exponential distribution is both the selected model and that with estimated higher probability.

For $h = 1$, the naive model is lognormal, the standard error for the naive model is 7.97 and that of model selection bootstrap is 9. Both bias are close for the two approaches, as well as similar confidence intervals. These observations can be seen on the histogram, Figure (7.2) where both are quite close.

For $h = 5$, the standard error of bootstrap is 44.36, large than that of naive (selected) Weibull 14.99. The absolute bias is also higher for model selection bootstrap. The later also has larger confidence interval. Graphical illustration is given in Figure (7.3), where, whereas the histogram of naive Weibull is unimodal, that of bootstrap model selection is bimodal.

For $h = 10$, the bootstrap model selection also has higher absolute bias, higher standard error and larger confidence interval than the naive (selected) exponential. The bootstrap model selection histogram (Figure (7.4)) is bimodal, whereas unimodal for the naive exponential model.

One can see that the behaviour of the bootstrap-after-model-selection estimators are similar to that of PMSEs: likely to have more bias, larger variance, wider confidence interval, multimodality. This is a general fact about bootstrap-after-model-selection estimator. This can lead one to think that this is then an approximate solution to the problem. The point is how good is this approximation. As explained above, we can not trust this approximation.

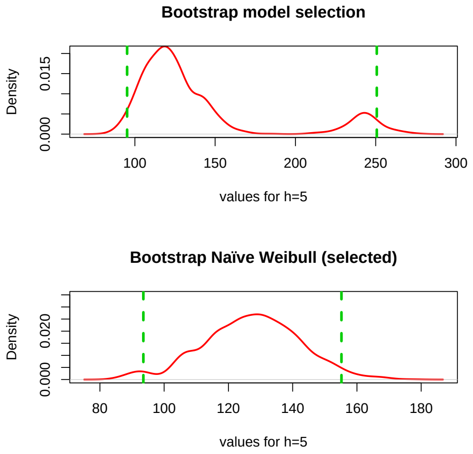

Figure 7.3: Comparison density of naive Weibull and bootstrap model selection for $h = 5$ with 0.95 percentile intervals.

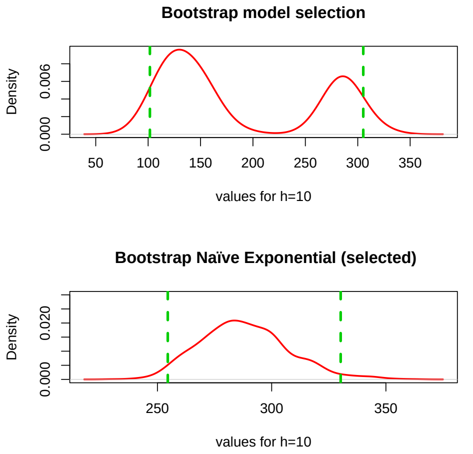

Figure 7.4: Comparison density of naive exponential and bootstrap model selection for $h = 10$ with 0.95 percentile intervals.

Horizon	Selected Model	Characteristic	Naive	Bootstrap
$h = 1$	Lognormal	Estimate	93.73	93.73
		Mean	93.06	93.55
		Standard error	7.97	9.00
		Bias	-0.67	-0.18
		Lower ci	79.08	77.57
		Upper ci	109.83	111.83
$h = 5$	Weibull	Estimate	130.23	130.23
		Mean	127.36	138.07
		Standard error	14.99	44.36
		Bias	-2.87	7.84
		Lower ci	93.53	95.18
		Upper ci	155.18	250.67
$h = 10$	Exponential	Estimate	288.74	288.74
		Mean	288.04	186.66
		Standard error	19.37	75.34
		Bias	-0.7	-102.08
		Lower ci	254.53	101.68
		Upper ci	330.20	305.26

Table 7.2: Comparison bootstrap model selection and naive approach.

Distribution	$h = 1$	$h = 5$	$h = 10$
gamma	0.04	0.01	0.01
normal	0.02	0.01	0.00
lognormal	0.33	**0.36**	0.18
exponential	0.00	0.14	**0.35**
Weibull	0.20	0.23	0.21
Gumbel	**0.41**	0.25	0.25
High model probability	Gumbel	lognormal	exponential

Table 7.3: Bootstrap model selection probability estimates.

Chapter 8

Summary and Conclusion

8.1 Summary

Much of classical inference is based on the assumption that the model being fitted is known, and that only the parameters are unknown. It is based on this assumption that the standard formulae for confidence intervals, p-values, etc. are correct. Such formulae are no longer valid if one uses the same data to first select a model. By selection we include the case in which models are developed via tests of hypotheses, and, in general, all procedures covered by the term "iterative model building". Although model selection is universally applied in statistical analyses, its consequences are poorly understood even by statisticians. This thesis has examined a number of these consequences.

We discussed the *model uncertainty problem*. A method to deal with this is to use model averaging rather than any single model. The method of model averaging was discussed in detail, both from the Bayesian and the frequentist points of view. This involves using weighted averages of estimates based on the different models.

We argue that some issues regarding Bayesian model averaging (BMA) have not been clearly described in full. We demonstrated that BMA estimators based on the currently applied weights are such that their long-run (frequentist) properties (e.g. minimaxity, admissibility) are difficult to assess; it is possible to examine their conditional properties (given the data). No prior and a well-defined data-generating mechanism have been assigned to BMA. Therefore, in general, it is hard to show whether BMA methodology constitutes a fully Bayesian approach. We proposed an alternative way of weighting the models, a "fully Bayesian model averaging" (FBMA) approach, which leads to model averaging estimators whose long-run properties are well-defined. The key ingredient is the use of an average of prior distributions and parametric models, the weights being the corresponding model probability.

To reduce the enormous computational effort required to apply BMA it has been suggested that some models be eliminated in a “preselection” step. Suggestions include Occam’s window, Markov chain Monte Carlo model composition and stochastic search variable selection. We stress the fact that the long-run performance of BMA estimators will be affected if *data-based* model search methods are applied. This introduces an additional source of uncertainty which we call *model space selection uncertainty*. The application of preselection changes the estimator, and therefore its properties. It is necessary to take that source of additional uncertainty into account. For *posterior* analysis, i.e. conditioned on the data, such search strategies present no problem.

We next examined the *model selection uncertainty* problem. The usual procedure that is applied in practice is to select a single model (using some data-based selection criterion) and then to apply the model ignoring the fact that the data has already been used for the selection. It is known that the resulting *post-model-selection estimator* (PMSE) is in general biased, and that ignoring this fact leads to invalid inference. We examined the problem from the point of view of decision theory. We point out that *all* selection procedures partition the sample space, and so the act of selecting a particular model conditions the resulting estimator on the fact that the sample is restricted to a subset of the sample space. We show that, even in the context of a specified problem, no single model selection procedure is better (has smaller risk) than any other in all circumstances. Secondly we show that the model selection uncertainty problem is not solved even if one uses a consistent model selection procedure, i.e. a procedure that will always select the “correct model” asymptotically. The reason is that from the asymptotic efficiency view, their normalised risks grow without bound as the sample size increases.

We point out an important theoretical issue concerning the existence of moments of PMSEs. This is especially desirable for widely applied selection methods, such as the typical methods used for variable selection in regression. The well-established and widely-accepted approach to statistical modeling called “iterative model building” constitutes a very complex model selection procedure. For such criteria, the existence of moments is difficult to establish. However, these difficulties should be weighed against the fact that, unless these are found, inference following such selection cannot be regarded as valid.

Model averaging and model selection are generally regarded as two entirely different methodologies. The former estimator is based on a weighted average of estimators from the different models; the latter is based on a single selected model.

We show that, suitably represented, it is possible to regard these two methodologies as different manifestations of the same problem. In this framework, we point out that, in terms of risk function, neither of the above approaches can be expected to consistently outperform the other. We propose a method that accounts for the selection procedure in classical model averaging by using *adjusted Akaike weights* (AAW) and *adjusted likelihood weights* (ALW), the adjusting factor being model selection probabilities. The resulting model averaging estimator dominates the PMSEs.

In the Bayesian context, we argue that as long as one is concerned with posterior evaluation (Bayes risks, posterior variance, etc.) of an estimator, i.e. conditional on the data, model selection uncertainty is not an issue (only model uncertainty matters); the data are held fixed. In this case it is valid to perform post-model selection inference. That is, the model selection uncertainty problem is eliminated by performing a conditional inference, e.g. Bayesian model selection inference. However, if interest is focused on frequentist performance of estimators (e.g. frequentist risk) then the problem of model selection uncertainty exists and can be very severe. BMA estimators are not better than Bayesian post-model-selection estimator (BPMSE). We propose a model averaging procedure, *adjusted Bayesian model averaging* (ABMA). The proposed weights are functions of the *prior model selection probabilities* and the approach is better than the BPMSE.

We argue that parameter estimation can sometimes also be regarded as model selection. Therefore, a two-step estimation is viewed as model selection followed by inference. This (unusual) point of view provides illustrations of cases in which model selection uncertainty has been taken into account. A number of well-known distributions can be interpreted in this way.

It would be most convenient if bootstrap methods could be used to estimate the properties of post-model-selection estimators. We illustrate that, unfortunately they do not necessarily do so.

We also point out that the properties of model averaging and post-model-selection estimators can only be derived under an assumed true model. However, there is uncertainty about the choice of this model and it is precisely this uncertainty that led to model averaging or model selection. Under such an assumption, one would simply use that model without applying model selection or model averaging. The same issue also arises when assessing the properties of the bootstrap-after-model-selection estimator. However, the properties of an estimator are well defined if one computes with respect to the model from which this estimator is derived. It is this circularity that makes the problem so difficult to deal with. For model averaging, the fact of not knowing the model that generated the model averaging estimator leads to the difficulty of interpreting it.

Finally, we point out that the use of informal model selection criteria such as exploratory data analysis and iterative model building renders the model selection uncertainty problem difficult to take into account. More generally, any (frequentist) statistical inference after a preliminary inspection of the data (e.g. graphical inspection as informal selection procedure) is suspect and may suffer of model selection uncertainty. Without entering in the controversy about the merits of frequentist and Bayesian methods, from the model selection uncertainty point of view, we recommend the use of Bayesian methods for statistical data analysis as long as one is not interested in the frequentist properties of the resulting estimator. In this case, the remaining uncertainty includes the choice of the statistical model, the prior and the loss (or utility) function.

8.2 Suggestions for future research

The performance of the proposed adjusted Akaike model averaging, adjusted likelihood model averaging, adjusted Bayesian model averaging and fully Bayesian model averaging have only been examined here in a few simple special cases. Our intention has been to establish their feasibility. These methods need to be examined in much greater detail, in a much greater variety of situations, and, especially, in the context of applications.

References

Akaike, H. (1970), Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* **22**, 203-217.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, eds. B. Petrov and F. Csáki, Budapest: Akadémiai Kiadó, 267-281.

Akaike, H. (1978), On the likelihood of a time series model. *The Statistician* **27**, 217-235.

Akaike, H. (1981), Likelihood of a model and information criteria. *Journal of Econometrics* **16**, 3-14.

Alonso, A., Peña, D., and Romo, J. (2004), Introducing model uncertainty in time series bootstrap. *Statistica Sinica* **14**, 155-174.

Ahmed, S. E., and Basu, A. K. (2000), Least squares, preliminary test and Stein-type estimation in general vector AR(p) models. *Statistica Neerlandica* **54**, 47-66.

Bancroft T.A. (1944), On bias in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* **15**, 190-204.

Barbieri, M. M., and Berger, J. (2004), Optimal predictive model selection. *Annals of Statistics* **32**, 870-897.

Barnard, G. A. (1963), New methods of quality control. *Journal of the Royal Statistical Society,* series A **126**, 255.

Barndorff-Nielsen, O. E. (1983), On a formula for the distribution of the maximum likelihood estimator. *Biometrica* **70**, 343-365.

Barndorff-Nielsen, O. E. (1988), *Parametric statistical models and likelihood.* Lecture Notes in Statistics **50**, Heidelberg: Springer-Verlag.

Bates, J. M., and Granger, C. W. J. (1969), The combination of forecasts. *Operational Research Quality* **20**, 451-468.

Bayarri M. J., and Berger, J. O. (2004), The interplay of Bayesian and frequentist analysis. *Statistical Science* **19**, 58-80.

Berger, J. (1985), *Statistical decision theory and Bayesian analysis*, 2nd edition, New York: Springer.

Bernado, J. M., and Rueda, R. (2002), Bayesian Hypothesis testing: A reference approach. *International Statistical Review* **70**, 351-372.

Bernado, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Breiman, L. (1992), The little bootstrap and other methods for dimensionality selection in regression: X-Fixed predictor error. *Journal of the American Statistical Association* **87**, 738-754.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), Model selection: An integral part of inference. *Biometrics* **53**, 603-618.

Bunea, F. (2004), Consistent covariate selection and post model selection inference in semiparametric regression. *The Annals of Statistics* **32**, 898-927.

Burnham, P. K., and Anderson, D. R. (2002), *Model Selection and Multimodel Inference, a practical Information-Theoretic Approach*, 2nd Edition, New York: Springer-Verlag.

Candolo, C., Davison, A. C., and Demétrio, C. G. B. (2003), A note on model uncertainty in linear regression. *The Statistician* **158**, 165-177.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and empirical Bayes methods for data analysis*, 1st edition, London: Chapman and Hall.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Belmont, California: Wadsworth and Brooks/Cole.

Chatfied, C. (1995), Model Uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society,* series B **158**, 419-466.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2001) The practical implementation of Bayesian model selection (with discussion). In *Model selection,* (P. Lahiri, ed.) 65-134. IMS, Beachwood, OH.

Claeskens, G., and Hjort, N. L. (2003), The focused information criterion. *Journal of the American Statistical Association* **98**, 900-916.

Clyde, M. A. (1999), Bayesian model averaging and model search strategies. In *Bayesian Statistics* **6** (Bernado, J. M., Berger, J. O., Dawid, A.P., and Smith, A. F. M., eds) 157-185, Oxford Univ. Press. 81-94.

Clyde, M., Desimone, H., and Parmigiani, G. (1996), Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91**, 1197-1208.

Clyde, M., and George, E. I. (2000), Flexible empirical Bayes estimation for Wavelets. *Journal of the Royal Statistical Society,* series B **62**, 681-698.

Clyde, M., and George, E. I. (2004), Model Uncertainty. *Statistical Science* **19**, 81-94.

Cox, D. R., and Reid, N. (1993), Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society,* series B **49**, 1-39.

Danilov, D., and Magnus, J. R. (2004), On the harm that ignoring pretesting can cause. *Journal of Econometrics* **122**, 27-46.

Davison, A. C. (2003), *Statistical Models*, Cambridge: Cambridge University Press

Dijkstra, T.K., ed. (1988), *On Model Uncertainty and Its Statistical Implications.* Proceedings of a workshop held in Groningen, The Netherlands, September 25-26, 1986, Berlin: Springer-Verlag.

Dijkstra, T. K., and Veldkamp, J. H. (1988), Data-driven selection of regressors and the bootstrap. *Lecture Notes in Economics and Mathematical Systems* **307**, 27-46.

Draper, D. (1995), Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society,* series B **57**, 45-97.

Efron, B. (1979), Bootstrap methods: another look at the jacknife. *Annals of Statistics* **7**, 1-26.

Efron, B. (2004), The estimation of prediction error:covariance penalties and cross-validation. *Journal of the American Statistical Association* **99**, 619-642.

Efron, B., and Morris, C. (1971), Limiting the risk of Bayes and empirical Bayes estimators- Part I: The Bayes case. *Journal of the American Statistical Association* **66**, 807-815.

Efron, B., and Morris, C. (1972), Limiting the risk of Bayes and empirical Bayes estimators- Part I: The empirical Bayes case. *Journal of the American Statistical Association* **67**, 130-139.

Furnival, G. M., and Wilson, R. W. (1974), Regression by leaps and bounds. *Technometrics* **16**, 499-511.

George, E. I. (1999), Bayesian model averaging. In *Encyclopedia of Statistical Sciences Update* **3**, New York: Wiley.

George, E. I., and McCulloch, R. E. (1993), Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

Giles, J. A., and Giles, D. E. A. (1993), Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys* **7**, 145-197.

Giles, D. E. A., and Srivastava, V. K. (1993), The exact distribution of a least squares regression coefficient after a preliminary t-Test. *Statistics and Probability Letters* **16**, 59-64.

Good, I. J. (1952), Rational decisions. *Journal of the Royal Statistical Society,* series B **14**, 107-114.

Hall, P. (1990), Using the bootstrap to estimate mean square error and select smoothing parameters in nonparametric problems. *Journal of Multivariate Analysis* **32**, 177-203.

Hannan, E. J., and Quin, B. G. (1979), The determination of the order of an autoregression. *Journal of the Royal Statistical Society,* series B **41**, 190-195.

Hjort, N. L., and Claeskens, G. (2003), Frequentist Model Average Estimators. *Journal of the American Statistical Association* **98**, 879-899.

Hjorth, J. (1994), *Computer intensive statistical methods:Validation, model selection, and bootstrap*, London: Chapman and Hall.

Hoeting J., Madigan D., Raftery A., and Volinsky C. (1999), Bayesian model averaging: A tutorial. *Statistical Science* **4**, 382-417.

Huntsberger, D. V. (1955), A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics* **26**, 734-43.

Judge, G., and Bock, M. (1978), *The statistical implications of pre-test and Stein rule estimators in Econometrics*, Amsterdam: North Holland.

Judge, G., and Bock, M. (1983), *Biased estimation*, In Z Griliches, and M. D. Intriligator (Eds.), Handbook of Econometrics, Volume I, Chapter 10, Amsterdam: North Holland.

Judge, G., and Yancy, T. A. (1986), *Improved methods of inference in econometrics*, Amsterdam: North Holland.

Kabaila, P. (1995) Effet of model selection on confidence regions and prediction regions. *Econometric Theory* **11**, 537-549.

Kabaila, P. (1998) Valid confidence intervals in regression after variable selection. *Econometric Theory* **14**, 463-482.

Kadane, J. B., and Lazar, N. A. (2004), Methods and Criteria for model selection. *Journal of the American Statistical Association* **99**, 279-290.

Kapetanios, G. (2001), Incorporating lag order selection uncertainty in parameter inference for AR models. *Economics Letters* **72**, 137-144.

Kass, R. E., and Raftery, A. E. (1995), Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

Kilian, L. (1998), Accounting for lag order uncertainty in autoregressions: the endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis* **19**, 531-548.

Knight, K. (1999), *Epi-convergence in distribution and stochastic equi-semicontinuity.* Working Paper, Department of Statistics, University of Toronto.

Kulperger, R. J., and Ahmed S. E. (1992), A bootstrap theorem for a preliminary test estimator. *Communications in Statistics: Theory and Methods* **21**, 2071-2082.

Lauritzen, S. L. (1996), *Graphical models*, Oxford: Clarendon Press.

Leamer, E. E. (1978), *Specification searches*, New York: Wiley.

Le Cam, L. M., and Yang, G. L. (2000), *Asymptotics in statistics. Some basic concepts*, 2nd Edition, New York: Springer.

Lehmann, E. L. (1983), *Theory of point estimation*, New York: Wiley.

Lehmann, E. L., and Casella, G. (1998), *Theory of point estimation.* Springer Texts in Statistics, New York: Springer Verlag.

Lehmann, E. L., and Casella, G. (2001), *Theory of point estimation*, 2nd Edition, New York: Springer Verlag.

Leblanc, M., and Tibshirani, R. (1996), Combining estimates in regression and classification. *Journal of the American Statistical Association* **91**, 1641-1650.

Leeb, H., and Pötscher, B. M. (2005), Model selection and inference: Fact and fiction. *Econometric Theory* **21**, 21-59.

Linhart,H., and Zucchini, W. (1986), *Model selection*, New York: John Wiley and Sons.

Longford, N. T. (2005),. Editorial: Model selection and efficiency-is 'which model ...?' the right question? *J. R. Statist. Soc. A* **168**, Part3, 469-472.

Lovell, M. C. (1983), Data mining. *Review of Economics and Statistics* **65**, 1-12.

Madigan, D., and Raftery, A. E. (1994), Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1535-1546.

Madigan, D., and York, J. (1995), Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215-232.

Magnus, J. R. (1999), The traditional pretest estimator. *Econometrica* **67**, 639-643.

Magnus, J. R. (2002), Estimation of the mean of a univariate normal distribution with known variance. *The Econometrics Journal* **5**, 225-36.

Magnus, J. R., and Durbin J. (1999), Estimation of regression coefficient of interest when other regression coefficients are of no interest. *Theory of Probability and Its Applications* **44**, 293-308.

Mallows, C. L. (1973), Some comments on Cp. *Technometrics* **15**, 661-675.

Marriott, J. M., Spencer N. M., and Pettitt A. N. (2001), A Bayesian approach to selecting covariates for prediction. *Scandinavian Journal of Statistics* **28**, 87-97.

McQuarrie, A. D. R., and Tsai, C.L. (1998), *Regression and time series model selection*, Singapore: World Scientific.

Miller, A. J. (2002), *Subset selection in regression*, 2nd Edition, London: Chapman and Hall.

Mittelhammer, R. C. (1984), Restricted least squares, pre-test, OLS, and Stein rule estimators: Risk comparison under model misspecification. *Journal of Econometrics* **25**, 151-164.

Mosteller, F. (1948), A k-sample slippage test for an extreme population. *Annals of Mathematical Statistics* **19**, 58-65.

Neyman, J., and Scott, E. L. (1948), Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

Nguefack, G., and Zucchini, W. (2005), *Inference after model selection in linear regression.* Internal Report, Institut für Statistik und Ökonometrie, Universität Göttingen.

Pace, L., and Salvan, A. (1997), *Principles of statistical inference from a neo-Fisherian perspective,* Singapour: World Scientific.

Pierce, D. A. (1982), The asymptotic effect of substituing estimators for parameters in certain types of statistics. *The Annals of Statistics* **10**, 475-478.

Pötscher, B. M. (1991), Effects of model selection on inference. *Econometric Theory* **7**, 163-185.

Pötscher, B. M. (1995), Comment on ”Model Uncertainty, data mining and statistical inference” by Chatfied, C. *Journal of the Royal Statistical Society,* series B **158**, 461.

Pötscher, B. M., and Novak A. J. (1998), The distribution of estimators after model selection: large and small sample results. *Journal of Statistical Computing and Simulation* **60**, 1035-1065.

Qin, J., and Zhang, B. (2005), Marginal likelihood, conditional likelihood and empirical likelihood: Connections and applications. *Biometrika* **92**, 251-270.

Raftery, A. E. (1999), Bayes factors and BIC: Comment on ”A critique of the Bayesian Information Criterion for model selection”. *Sociological Methods and Research* **27**, 411-427.

Raftery, A. E., Madigan, D., and Hoeting, J. (1992), Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179-191.

Raftery, A. E., and Zheng, Y. (2003), Discussion: Performance of Bayesian model averaging: Comment on ”Frequentist model averaging estimators”. *Journal of the American Statistical Association* **98**, 931-938.

Randles, R. H. (1982), On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* **10**, 462-474.

Reid, N. (1988), Saddlepoint methods and statistical inference (with discussion). *Statistical Science* **3**, 213-238.

Robert, C. P. (2001), *Bayesian Choice*, 2nd edition, New York: Springer.

Roberts, H. V. (1965), Probabilistic prediction. *Journal of the American Statistical Association* **60**, 50-62.

Roehrig, C. S. (1984), Optimal critical regions for pre-test estimators using a Bayes risk criterion. *Journal of Econometrics* **25**, 3-14.

Schwarz, G. (1978), Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

Sclove, S. L, Morris C., and Radhakrishnan, R. (1972), Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* **43**, 1481-1490.

Sen, P. K. (1987), Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* **7**, 1019-1033.

Sen, P. K., and Saleh, A. K. M. E. (1987), On the preliminary test and shrinkage M-estimation in linear models. *Annals of Statistics* **15**, 1580-1592.

Shao, J. (1996), Bootstrap model selection. *Journal of the American Statistical Association* **91**, 655-665.

Shao, J. (1997), An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* **7**, 221-264.

Shen, X., Huang, H.C., and Ye, J. (2004), Inference after model selection. *Journal of the American Statistical Association* **93**, 120-131.

Shen, X., and Ye, J. (2002), Adaptative model selection. *Journal of the American Statistical Association* **97**, 210-221.

Spiegelhalter, D. J., Dawid, A., Lauritzen, S., and Cowell, R. (1993), Bayesian analysis in expert systems (with discussion). *Statistical Science* **8**, 219-283.

Stein, C.M. (1981), Estimation of the mean of a multivariate normal distribution. *Annals of statistics* **9**, 1135-1151.

Strimmer, K., and Rambaut, A. (2001), Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc Lond B Biol Sci* **269**, 137-142.

Swanepoel, J. W. H. (1986), A note on proving that the (modified) bootstrap works. *Communications in statistics, Part A-Theory and Methods* **15**, 3193-3203.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A (1997), Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society,* series C **46**, 433-448.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Cambridge: Cambridge University Press.

Wasserman, L. (2000), Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92-107.

Ye, J. (1998), On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93**, 120-131.

Zucchini, W. (2000), An introduction to model selection. *Journal of Mathematical Psychology* **44**, 41-61.

Zucchini, W., and Adamson, P. T. (1984), *Assessing the risk of deficiencies in streamflow.* WRC Report No. 91/2/84, Water Research Commision, South Africa.

Curriculum Vitae

Personal Data

Last name: Nguefack Tsague

First name: Georges Lucioni Edison

Date of birth: December 05 1971

Place of birth: Fongo-Tongo, Cameroon

Nationality: Cameroonian

Address: Christophorusweg 12/108, 37075, Göttingen, Germany

E-mail: gnguefack@yahoo.fr or gnguefa@uni-goettingen.de

Education

Oct 2003- Feb 2006 : Ph.D., Institute for Statistics and Econometrics,
Center for Statistics, University of Göttingen, Germany.

Sept 2000- June 2001: M.A in Economics, University of Namur (FUNDP), Belgium.

Sept 1991- June 1997: M.A in Statistics (ISSEA); B.Sc in Mathematics (University of Yaoundé I), Yaoundé, Cameroon.

Sept 1990- June 1991: G.C.E Advanced Level in Mathematics, Notre Dame College, Dschang, Cameroon.

Sept 1989- June 1990: Probatoire, Notre Dame College, Dschang, Cameroon.

Sept 1983- June 1988: G.C.E Ordinary Level, Big-Ben College, Dschang, Cameroon.

Sept 1978- June 1983: First School Leaving Certificate, Ecole Publique de Fossong-Tchuentchué, Dschang, Cameroon.

Employment

Oct 2001- Sept 2003: Teaching assistant, University Carlos 3, Department of Statistics, Madrid, Spain.

July 1997- Sept 2000: National Institute of Statistics (INS); Ministry of Economy and Finance (MINEFI), Yaoundé, Cameroon.