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 deal-

ing 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.