

# Frequentist Model Average Estimators

Nils Lid Hjort and Gerda Claeskens

*University of Oslo and Texas A&M University*

**ABSTRACT.** The traditional use of model selection methods in practice is to proceed as if the final selected model had been chosen in advance, without acknowledging the additional uncertainty introduced by model selection. This often means underreporting of variability and too optimistic confidence intervals. We build a general large-sample likelihood apparatus in which limiting distributions and risk properties of estimators-post-selection as well as of model average estimators are precisely described, also explicitly taking modelling bias into account. This allows a drastic reduction of complexity, as competing model averaging schemes may be developed, discussed and compared inside a statistical prototype experiment where only a few crucial quantities matter. In particular we offer a frequentist view on Bayesian model averaging methods and give a link to generalised ridge estimators. Our work also leads to new model selection criteria. The methods are illustrated with real data applications.

**KEY WORDS:** *bias and variance balance, growing models, likelihood inference, model average estimators, model information criteria, moderate misspecification*

## 1. Introduction and summary

An impressive range of model selection criteria has been developed and finessed over the past three decades. These have been constructed from different sets of intentions and have been aimed partly at general parametric models while others have been geared towards special types of statistical models, like time series, neural networks or hazard rate regression; some are inspired by Bayesian considerations while others are more traditional frequentistic; some have arisen via asymptotics and optimality properties for large samples while others have been more fine-tuned for moderate sample sizes; etc. A fair list of these model choice schemes have also successfully made the passage from university blackboards to statistical software packages and the mainstream of applied statistical research. Methods like the AIC and the BIC (the Akaike and the Bayesian information criteria), with suitable modifications, along with various stepwise methods for subset selection in regression models, are applied routinely also by non-specialists. For overviews of model selection literature one may consult the monograph Burnham and Anderson (2002) and the introductory sections of Spiegelhalter, Best, Carlin and van der Linde (2002) and Claeskens and Hjort (2003).

**1.1. Estimator-post-selection problems.** It is fair to say, however, that far less work has been carried out, and even less has reached mainstream statistical applications, regarding the many complementary questions related to the consequences of model selection. In statistical practice, one typically applies some off-the-shelf model selection scheme, perhaps supplemented with brief goodness-of-fit checking of residuals, to arrive at some ‘good

model' that is thought to adequately reflect the main aspects of data – after which one proceeds with one's analysis as if this good model had been decided on in advance. It is clear that such analysis 'hides (or ignores) some uncertainty'; reported confidence intervals tend to be too short, an hypothesis rejected at an announced 5% significance level might actually have been tested at a rather higher level, and so on. A central issue is that estimators formed after model selection really are like mixtures of many potential estimators, namely those that would have been computed had the random model selectors landed differently. A second theme is that it is sometimes advantageous to smooth estimators across several models, rather than sticking to only the model that is being reached by a single selection criterion.

There are at least two clear reasons why 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 and partly non-formalised ingredients like 'looking at residuals' and 'trying a suitable transformation'. The second is that these questions of estimator-post-selection behaviour simply are harder to formalise and analyse.

An honourable exception is that of 'Bayesian model averaging' (BMA), where more than a hundred papers have been published over the past decade. If a Bayesian can put down prior probabilities for a list of potential models, along with priors for the parameters of each model, then the Bayesian machinery is in principle capable of delivering the posterior distribution of any interest parameter (provided it retains a precise interpretation and is well defined across the models under study). The tutorial by Hoeting, Madigan, Raftery and Volinsky (1999) discusses pertinent issues of interpretation and implementation via the machinery of Markov chain Monte Carlo, where the chains in question move between models of different dimensions; see also Green (2003) for a review of trans-dimensional McMC theory. With BMA methodology, the extra estimator variability stemming from not knowing the correct model *a priori* is adequately taken into account.

The approach remains problematic, however. First of all there are difficulties associated with the often ad hoc way in which the prior probabilities for a (sometimes long) list of models is set up; see the discussion to Hoeting et al. (1999). Secondly, we raise concern for the fact that the typical application of BMA involves mixing together many conflicting prior opinions regarding interest parameters. If  $\mu$  is some parameter of interest, and  $\mu = \mu(\alpha_j)$  in terms of the parameters  $\alpha_j$  of candidate model  $j$ , with prior  $\pi_j(\alpha_j)$ , this leads to a prior  $\bar{\pi}_j(\mu)$ , say; why would a statistician entertain many different such inside the same problem formulation, and what are the consequences in cases where some of these have clear clashes? Finally, even though BMA 'works', insofar as adequate analysis of data can be carried out after judicious selection of models, prior probabilities for these, and prior densities for parameters in each model, rather little appears to be known about the actual performance or behaviour of the consequent inferences, like estimator precision.

The present article aims at establishing a framework where properties of estimator-

post-selection and estimator average methods can be accurately described. Our framework is general and unified, and involves large-sample likelihood approximations across a list of parametric models. The end result is a machinery for ‘frequentist model averaging’ (FMA), to be partly contrasted with that of BMA. Within this context many natural model averaging strategies can be developed and compared. Our results also shed light on the behaviour of BMA schemes, in fact by leading to precise large-sample results about their behaviour.

*1.2. An illustration: averages over logistic regressions.* To illustrate and pinpoint some of the problems associated with model selection and model averaging, consider the following example. The data set studied is taken from Appendix I in Hosmer and Lemeshow (1989), and concerns factors that may influence the birth weight of babies, in particular, the event that the baby weighs less than 2500 gram. Covariate information for the  $n = 189$  mothers in question included weight just prior to pregnancy ( $x_2$ , in pounds), age ( $x_3$ ), as well as indicators for race ‘black’ ( $x_4$ ) and race ‘other’ ( $x_5$ ); mothers with  $x_4 = 0$  and  $x_5 = 0$  are of race ‘white’. For the purposes of this article we make the assumption that

$$p(x, u) = \Pr\{\text{low birth weight} | x, u\} = \frac{\exp(x^t \beta + u^t \gamma)}{1 + \exp(x^t \beta + u^t \gamma)},$$

where  $x = (1, x_2)^t$  is always to be included in the logistic regression, while subsets of  $u = (x_3, x_4, x_5)^t$  may or may not enter the equation. See also Claeskens and Hjort (2003).

model	-AIC	-BIC	white	SE	black	SE	ratio	SE
0	232.691	239.174*	0.298	0.035	0.256	0.040	0.861	0.060
3	233.123	242.849	0.288	0.035	0.272	0.043	0.945	0.094
4	231.075*	240.800	0.269	0.037	0.412	0.101	1.533	0.423
5	234.101	243.826	0.279	0.041	0.242	0.043	0.868	0.062
34	232.175	249.068	0.264	0.037	0.413	0.101	1.564	0.435
35	234.677	247.644	0.272	0.041	0.259	0.046	0.950	0.097
45	231.259	244.226	0.231	0.044	0.414	0.100	1.794	0.547
345	232.661	248.869	0.230	0.044	0.414	0.100	1.801	0.551

TABLE 1.1. For submodels corresponding to inclusion or not of the covariates  $x_3, x_4, x_5$ , the table lists minus AIC, minus BIC, and then estimates along with estimated standard deviations (computed under the model assumption in question). These are the low birth weight probabilities  $p(\text{white})$ ,  $p(\text{black})$  and the ratio  $p(\text{black})/p(\text{white})$ .

It is convenient to label the eight potential submodels ‘0’, ‘3’, ‘4’, ‘5’, ‘34’, ‘35’, ‘45’, ‘345’, corresponding to inclusion or exclusion of these three extra covariates. We shall take an interest in estimating three parameters, the probability of low birth weight for the average ‘white’ and ‘black’ mothers, and for the ratio of these two. Table 1.1 gives estimates along with associated standard errors for these three estimands, for each of the eight possible models. The table also includes minus AIC and minus BIC, where AIC is twice the maximised log-likelihood minus say  $2k$ , where  $k$  is the number of parameters in

the model, while BIC is twice the maximised log-likelihood minus  $k \log n$ . We see that the AIC selects ‘4’ ahead of ‘45’, while the BIC prefers the narrow model ‘0’ ahead of ‘4’. See also Claeskens and Hjort (2003) for further analysis of these data using the focussed information criterion (FIC), which finds the best model for given interest parameter.

The estimated standard deviations given here have been computed via familiar delta method algebra and approximate normality of the maximum likelihood estimators, and under the typical assumption that the model under consideration is adequate. While the sampling variance perhaps may be adequately estimated here (conditional on the model), there is potential modelling bias, not reflected in the table and not easy to assess. Our article will develop methods that in particular make it possible to answer the following questions:

- (i) If a statistician uses the estimators dictated by the AIC (here, 0.269, 0.412, 1.533 for the three parameters), what are the real variances of these, and what are the biases stemming from the modelling imperfections of the selected logistic equation? How trustworthy are the confidence intervals delivered by standard use?
- (ii) Similarly, if another statistician uses BIC to decide on estimators (here, 0.298, 0.256, 0.861), how big might the modelling biases be, and what are the real variances involved?
- (iii) Are there advantages to taking suitable averages across models, for example weighted averages over those with best AIC, BIC or FIC scores? What are then the biases and variances involved? How can adequate confidence intervals be constructed?
- (iv) When will the simple ‘narrow method’, which here corresponds to disregarding the extra covariates, be more accurate than the ‘full model method’, which includes all five logistic parameters in the inference?
- (v) Could it be advantageous here to trust covariate  $x_2$  fully (along with  $x_1 = 1$ ), but to trust the influence of  $x_3, x_4, x_5$  less, in the sense of shrinking estimated logistic coefficients for these three towards zero?
- (vi) If a BMA regime is used here, what is its (frequentist) behaviour, and how do different BMA schemes compare in performance?
- (vii) Are there FMA schemes with suitable optimality properties?

**1.3. Related work.** As mentioned above, the Bayesian literature so far decidedly outgoliaths its frequentist counterpart concerning model averaging inference and estimator-post-selection performance. Some work in the frequentist directions has however been done over the last decade.

Hurvich and Tsai (1990) pointed out that for linear regression models, coverage rates of confidence intervals for regression parameters, conditional on the selected model, are much smaller than the nominal coverage rates. Such problems have been further addressed by Chatfield (1995) and Draper (1995). Also in a linear regression setting, in the presence of a finite-dimensional nuisance parameter, Kabaila (1995, 1998) considered the effect of

model selection on the construction of confidence intervals as well as on prediction intervals. Pötscher (1991) considered a sequence of nested models, containing an increasing number of parameters  $\theta_1, \dots, \theta_q$  and possibly a nuisance parameter  $\eta$ , in which a backward model selection is performed. He makes the assumption that there is a true model containing parameters  $(\eta, \theta_1, \dots, \theta_{q_0})$ , where  $1 \leq q_0 \leq q$ . Leeb and Pötscher (2000) further build on this subject and obtain distributions of post-model selection estimators under the condition of possibly selecting an incorrect model with fewer than  $q_0$  parameters. Their methods are restricted to linear regression models  $Y = X\theta + \varepsilon$  with independent and identically distributed Gaussian error terms, and for a similar backward selection procedure, employing a  $t$ -test at each stage of the selection procedure. Also in linear models, Sen and Saleh (1987) study the asymptotic distribution after a preliminary test for the presence of part of the regression coefficients, hence dealing with two possible models for the data. Bühlmann (1999) investigates local consistency of post-model selection estimators under a set of conditions which imply all ‘local’ models to have the same dimension asymptotically.

A few non-Bayesian methods for model averaging have been proposed in the literature. There is of course a large literature on model selection methods, which can be considered as hard-threshold averages; see Claeskens and Hjort (2003). George (1986a, 1986b) investigated multiple-shrinkage estimators in the normal model. Also, Foster and George (1994) explicitly analysed performance of estimators-post-selection, in a normal regression context. Rao and Tibshirani (1997) constructed an out-of-bootstrap method which leaves out one training point and constructs bootstrap model weights depending on how well the remaining bootstrap data predict the left out value. They did not provide any asymptotic distribution theory for the model averaging estimator. The adaptive regression by mixing of Yang (2001) splits the data set into two parts, where one part is used for estimation and the other for measuring the quality of predictions, on basis of which model weights are constructed. Buckland, Burnham and Augustin (1997) constructed model averaging weights based on the values of the AIC or BIC scores, further discussed in Burnham and Anderson (2002, Ch. 6). This may accordingly be seen as suggestions for problem (iii) described at the end of Section 1.2 above. The construction of Buckland et al. is somewhat ad hoc, however, and they do not really analyse the performance of the resulting estimator. The results of Section 4 below can be used to accurately describe its behaviour, and also answer other questions raised in their paper.

*1.4. The present article.* In Section 2 we build and discuss a general model selection framework, involving a finite number of parametric extensions around a given parametric basis model. This includes problems of subset selection in general regression models, and, with further amendments, also situations with memory order and averaging order for time series and Markov chain models. Section 3 develops the necessary theory of maximum likelihood estimators inside such a framework, where modelling bias is explicitly present and taken into account for each candidate estimator. Some attention is given to behaviour

of the AIC criterion. In Section 4 we describe a fairly general class of model average estimators, which compromise across a set of candidate models, and derive their limit distributions. These are not normal, but rather non-linear mixtures of normals. This is in particular true for estimator-after-selection schemes. We also pinpoint how the confidence level of confidence intervals becomes lower than the ‘intended level’ when the model selection step is being ignored. Various natural FMA strategies are proposed in Section 5, including Bayesian and empirical Bayesian variants. Then we illustrate our FMA machinery for some applications in Section 6. Section 7 extends the class of compromise estimators further, allowing ‘generalised ridging’, where estimates of potential extensions of a given model are being shrunk in a controlled fashion. This may often lead to smaller variances without significantly increased sizes of modelling bias. Then we turn to a frequentist view of BMA methods in Section 8. Apparently, despite a flurry of BMA activity over the last decade, performance of BMA schemes has not been studied in the classical sense of limiting distributions and large-sample approximations to risks; we do so here. In Section 9 we give some brief analysis of risk behaviour and comparisons, applying and illustrating theoretical results of previous sections. Our article ends with a list of concluding remarks in Section 10, some pointing to further research. All proofs are placed in Section 11.

## 2. A model averaging framework

This section establishes a fruitful general framework for model choice and model average estimators. The motivation is to start with a ‘narrow’ model, perhaps of standard type, and then add on one or more additional parameters to be able to reflect further features of the data generating mechanisms at work. This section partly parallels Section 2 of Claeskens and Hjort (2003), which focusses on model selection, whereas we here also are concerned with model averaging.

*2.1. Models with i.i.d. data.* Suppose independent data  $Y_1, \dots, Y_n$  come from density  $f$ . Inference is sought for a certain parameter estimand  $\mu = \mu(f)$ . We start with a basic narrow model, of the type  $f(y, \theta)$  with a  $p$ -vector of parameters  $\theta$ . The extended models take the form  $f(y, \theta, \gamma)$  with an additional  $q$ -vector of  $\gamma$ -parameters, where  $\gamma = \gamma_0$  corresponds to the narrow model in the sense that  $f(y, \theta) = f(y, \theta, \gamma_0)$ . Thus  $\gamma_0$  is fixed and known. Here one may consider employing suitable submodels, corresponding to having some of the  $\gamma_j$  parameters equal to  $\gamma_{j,0}$  while others are not. Using a bigger model would typically mean less modelling bias but increased estimation variance, and vice versa. At the outset there are  $2^q$  such submodels to consider, one for each subset  $S$  of  $\{1, \dots, q\}$ .

In this framework there is a variety of estimators to consider, from  $\hat{\mu}_{\text{full}} = \mu(\hat{\theta}_{\text{full}}, \hat{\gamma}_{\text{full}})$  using maximum likelihood estimators in the widest model where  $S = \{1, \dots, q\}$  to the simpler  $\hat{\mu}_{\text{narr}} = \mu(\hat{\theta}_{\text{narr}}, \gamma_0)$  which employs maximum likelihood estimation in the narrow model where  $S = \emptyset$ . The general submodel estimator is

$$\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c}), \quad \text{where } S \subset \{1, \dots, q\}, \quad (2.1)$$

found via maximum likelihood in the model that includes exactly the  $\gamma_j$  parameters for  $j \in S$  while keeping the others at  $\gamma_{0,j}$  ( $S^c$  is the complement of  $S$ ). The narrow model corresponds to  $S$  being the empty set. Further special cases would be the nested ones corresponding to  $S = \{1, \dots, k\}$  for  $k = 1, \dots, q$ .

Our intention is to investigate what happens to all the  $\hat{\mu}_S$  estimators, and importantly also averaged versions of these, in the local misspecification framework

$$f_{\text{true}}(y) = f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}). \quad (2.2)$$

The  $\delta_1, \dots, \delta_q$  parameters signify the degrees of the model departures in directions  $1, \dots, q$ , with due influence on the estimand  $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ . Later on we give results for the limiting risk of estimators  $\sum_S c(S)\hat{\mu}_S$ , with random weights summing to one. We assume that  $\mu(\theta, \gamma)$  is smooth in a neighbourhood of  $(\theta_0, \gamma_0)$ . The aim is to understand and assess large-sample approximations to distributions and say mean squared errors of subset and model average estimators, in situations where data come from  $f(y, \theta, \gamma)$  with  $\gamma$  not too far from  $\gamma_0$ , and it is for this reason that we work under the (2.2) scenario. The  $O(1/\sqrt{n})$  framework chosen here is canonical in the sense that it leads to the most fruitful large-sample approximations, with squared model biases and estimator variances as exchangeable currencies, both of size  $O(1/n)$ .

Our framework amounts to studying perturbations around a given narrow model in certain directions, expressed in (2.2) by letting  $\gamma$  vary around  $\gamma_0$ , and various consequences, for different estimators, are highlighted and discussed in the following sections. One may wonder whether yet further consequences of importance would emerge if we in addition perturb the  $\theta$  part of the model around the null value  $\theta_0$ . It turns out that there is no additional gain in considering such scenarios, as judged by what is to be learned from large-sample approximations of estimators and their performances; see Remark 4.1.

**2.2. Subset selection and mixtures of regression models.** With some efforts, the framework above may be generalised to encompass regression situations. Suppose  $Y_i$  for given covariate vectors  $x_i = (x_{i,1}, \dots, x_{i,p})^t$  are independent, with density of the type

$$f_{i,\text{true}}(y | x_i) = f(y | x_i, \beta_0, \sigma_0, \gamma_0 + \delta/\sqrt{n}),$$

most often with a  $p$ -dimensional  $\beta$  parameter, a scale parameter  $\sigma$ , and up to  $q$  further parameters  $\gamma$ . These ‘extra’ parameters could be associated with interactions among the  $x_{i,j}$  covariates or with other regressors. They could also help describe aspects of the variance structure, like a parametric model for the conditional variance in linear regression. It is again required that  $\gamma = \gamma_0$  leads back to the narrow model with only  $\beta$  and  $\sigma$  present. Focus parameters of interest take the form  $\mu = \mu(\beta, \sigma, \gamma)$ , which here corresponds to  $\mu_{\text{true}} = \mu(\beta_0, \sigma_0, \gamma_0 + \delta/\sqrt{n})$ . This could be the median regression surface or the standard deviation function evaluated at a point  $x_0$ , a quotient between two regression coefficients

or between two values of the mean regression function, and so on. In this framework we may consider submodel estimators  $(\hat{\beta}_S, \hat{\sigma}_S, \hat{\gamma}_S)$  via maximum likelihood in the model that employs  $\gamma_j$ s for  $j \in S$ . This leads to the estimator  $\hat{\mu}_S = \mu(\hat{\beta}_S, \hat{\sigma}_S, \hat{\gamma}_S)$  for the focus parameter. In Section 4 we give results for limiting risks of model average estimators  $\sum_S c(S)\hat{\mu}_S$ .

The type of local neighbourhood models described here also have parallels in time series and Markov chains, where it could be advantageous to weight across models with different memory lengths.

### 3. Limit distribution theory

In this section we establish notation necessary for handling analysis in the various submodels, and then sort out behaviour of different maximum likelihood estimators. We also give relevant limit results for log-likelihood ratios, which in particular are needed to understand the performance characteristics of the AIC model choice criterion.

We work throughout under traditional conditions of regularity, sufficient to apply familiar likelihood asymptotics arguments, as laid out in e.g. Lehmann (1983, Ch. 6). Thus the log density admits two continuous partial derivatives in all directions;  $(\theta_0, \gamma_0)$  is an inner point of the parameter space; the variance matrix of the score function statistic is finite and positive definite in a neighbourhood around this null point; and certain derivative operations can be taken under the integral sign. Details and proofs of the lemmas are given in Section 11.

**3.1. Notation for calculus in submodels.** Consider the score function

$$\begin{pmatrix} U(y) \\ V(y) \end{pmatrix} = \begin{pmatrix} \partial \log f(y, \theta_0, \gamma_0) / \partial \theta \\ \partial \log f(y, \theta_0, \gamma_0) / \partial \gamma \end{pmatrix},$$

with a  $p$ -dimensional  $U$  and  $q$ -dimensional  $V$ . Their  $(p+q) \times (p+q)$  variance matrix at the null model is

$$J_{\text{full}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \quad \text{with inverse} \quad J_{\text{full}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

say; in particular, let

$$K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}.$$

Under consideration are models indexed by subsets  $S$  of  $\{1, \dots, q\}$ . We let  $\pi_S$  be the projection matrix mapping  $v = (v_1, \dots, v_q)^t$  to the subvector  $\pi_S v = v_S$  of components  $v_j$  with  $j \in S$ . Hence  $\pi_S$  is of size  $|S| \times q$  with  $|S|$  being the size of  $S$ . For  $V_S = \pi_S V$  we then have

$$J_S = \text{Var}_0 \begin{pmatrix} U(Y) \\ V_S(Y) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01}\pi_S^t \\ \pi_S J_{10} & \pi_S J_{11}\pi_S^t \end{pmatrix}.$$

We shall also need its inverse matrix, which has blocks  $J^{11,S} = (\pi_S K^{-1} \pi_S^t)^{-1} = K_S$ ,  $J^{01,S} = -J_{00}^{-1} J_{01} \pi_S^t K_S$ , and  $J^{00,S} = J_{00}^{-1} + J_{00}^{-1} J_{01} \pi_S^t K_S \pi_S J_{10} J_{00}^{-1}$ .

LEMMA 3.1. Consider the averages  $\bar{U}_n = n^{-1} \sum_{i=1}^n U(Y_i)$  and  $\bar{V}_n = n^{-1} \sum_{i=1}^n V(Y_i)$ . Under the sequence of local alternatives (2.2),

$$\begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix} + \begin{pmatrix} M \\ N \end{pmatrix}, \quad \text{where } \begin{pmatrix} M \\ N \end{pmatrix} \sim N_{p+q}(0, J_{\text{full}}).$$

3.2. Behaviour of maximum likelihood estimators in submodels. Let  $(\hat{\theta}_S, \hat{\gamma}_S)$  denote maximum likelihood estimators in the model that includes  $\gamma_j$  parameters for  $j \in S$ .

LEMMA 3.2. Under the sequence of models  $f_{\text{true}}$  of (2.2),

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} &\xrightarrow{d} \begin{pmatrix} C_S \\ D_S \end{pmatrix} = J_S^{-1} \begin{pmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{pmatrix} \\ &\sim N_{p+|S|}(J_S^{-1} \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} \delta, J_S^{-1}). \end{aligned}$$

Before stating the next lemma it is convenient to introduce some more notation, which also will be needed later. Define first  $W = J^{10}M + J^{11}N = K(N - J_{10}J_{00}^{-1}M)$ . Here  $M \sim N_p(0, J_{00})$ , and it is not difficult to establish that  $M$  and  $W$  are stochastically independent, with  $W$  having a  $N_q(0, K)$  distribution. It follows from Lemma 3.2 and a little algebra that  $\hat{\delta}_S = \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S})$  tends in distribution to  $D_S = K_S \pi_S K^{-1}(\delta + W)$ . In particular,

$$D_n = \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D = \delta + W \sim N_q(\delta, K). \quad (3.1)$$

Next let

$$H_S = K^{-1/2} \pi_S^t K_S \pi_S K^{-1/2} \quad \text{and} \quad \omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}, \quad (3.2)$$

where the partial derivatives indicated are evaluated at the null model  $(\theta_0, \gamma_0)$ . Note that  $\omega$  is determined by the specifics of the focus parameter  $\mu$ . The  $H_S$  is a  $q \times q$  projection matrix, being symmetric and idempotent, and is orthogonal to  $I - H_S$ . We define  $H_\emptyset$  as the null matrix of size  $q \times q$ .

LEMMA 3.3. Assume  $\mu(\theta, \gamma)$  has continuous partial derivatives in a neighbourhood of  $(\theta_0, \gamma_0)$ . Then the maximum likelihood estimator of  $\mu$  in the  $S$  model has limiting distribution of the form

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} M + \omega^t (\delta - K^{1/2} H_S K^{-1/2} D),$$

where the partial derivatives indicated are evaluated at the null model  $(\theta_0, \gamma_0)$ . The limiting variable is normal, with mean  $\omega^t(I - K^{1/2} H_S K^{-1/2})\delta$  and variance  $(\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta} + \omega^t K^{1/2} H_S K^{1/2} \omega$ .

**3.3. AIC calculus.** The Akaike information criterion is equal to

$$\text{AIC}_{n,S} = 2 \sum_{i=1}^n \log f(Y_i, \hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c}) - 2|S|,$$

again with  $|S|$  being the number of elements in  $S$ . Its typical use is to pick out the model with the largest value of this criterion. In order to understand the behaviour of this criterion in the present framework, we start out with the likelihood-ratio statistic, expanding it to the second order, using familiar arguments. This leads to

$$\begin{aligned} G_{n,S} &= 2 \sum_{i=1}^n \log \{f(Y_i, \hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})/f(Y_i, \theta_0, \gamma_0)\} \\ &\doteq n \left( \frac{\bar{U}_n}{\bar{V}_{n,S}} \right)^t J_S^{-1} \left( \frac{\bar{U}_n}{\bar{V}_{n,S}} \right) \xrightarrow{d} \left( \frac{J_{01}\delta + M}{\pi_S J_{11}\delta + N_S} \right)^t J_S^{-1} \left( \frac{J_{01}\delta + M}{\pi_S J_{11}\delta + N_S} \right). \end{aligned}$$

(Here and later we use for notational simplicity  $X_n \doteq X'_n$  to indicate that the difference between the two variables tends to zero in probability; thus they have the same limit distribution, if it exists.) This is a noncentral chi-squared with  $p + |S|$  degrees of freedom. Furthermore,

$$\begin{aligned} G_{n,S} - G_{n,\emptyset} &= n(\bar{V}_{n,S} - J_{10,S}J_{00}^{-1}\bar{U}_n)^t J^{11,S}(\bar{V}_{n,S} - J_{10,S}J_{00}^{-1}\bar{U}_n) \\ &\xrightarrow{d} (K_S^{-1}\delta + N_S - J_{10,S}J_{00}^{-1}M)^t K_S (K_S^{-1}\delta + N_S - J_{10,S}J_{00}^{-1}M), \end{aligned}$$

which is a noncentral  $\chi^2_{|S|}(\delta^t K_S^{-1}\delta)$ .

Using a combination of previous arguments,  $\hat{\delta}_S = \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S})$  is at most  $o_p(1)$  away from  $\sqrt{n}K_S\pi_S(\bar{V}_n - J_{10}J_{00}^{-1}\bar{U}_n)$ , and similarly  $D_n \doteq \sqrt{n}K(\bar{V}_n - J_{10}J_{00}^{-1}\bar{U}_n)$ , which implies  $\hat{\delta}_S \doteq K_S\pi_S K^{-1}D_n$ . The important consequence is that  $\hat{\gamma}_S$ , the estimator based on the  $S$  subset model, can be expressed, within the first order local asymptotic framework, as a function of  $\hat{\gamma}_{\text{full}}$ . It also follows that the AIC criterion can be expressed in terms of  $D_n$  as

$$\text{AIC}_{n,S} = G_{n,S} - G_{n,\emptyset} - 2|S| = D_n^t K^{-1/2} H_S K^{-1/2} D_n - 2|S| + o_p(1). \quad (3.4)$$

**3.4. Results for the regression framework.** The methods and results above generalise to the regression type framework of Section 2.2 without too many difficulties, with the appropriate modifications and regularity conditions. An important ingredient is

$$J_{n,\text{full}} = \frac{1}{n} \sum_{i=1}^n \text{Var}_0 \begin{pmatrix} \partial \log f(Y_i | x_i, \beta_0, \sigma_0, \gamma_0) / \partial \beta \\ \partial \log f(Y_i | x_i, \beta_0, \sigma_0, \gamma_0) / \partial \sigma \\ \partial \log f(Y_i | x_i, \beta_0, \sigma_0, \gamma_0) / \partial \gamma \end{pmatrix} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix},$$

say, where  $J_{n,00}$  is of size  $(p+1) \times (p+1)$  and  $J_{n,11}$  of size  $q \times q$ . This matrix is assumed to converge to a suitable  $J_{\text{full}}$  as  $n$  increases. There are natural analogues of Lemmas 3.1–3.3

as well as for the AIC calculus results. Concrete regularity conditions would depend on the regression models studied. They would typically include assumptions of the Lindeberg–Lyapunov type  $n^{-1/2} \max_{i \leq n} \|x_i\| \rightarrow 0$ , and which are fulfilled in situations where the  $x_i$ s come from some covariate distribution with finite second moment.

#### 4. Estimators-after-selection and compromise estimators

The estimator employed by a statistician using a model selection criterion really takes the form  $\hat{\mu} = \hat{\mu}_{\hat{S}}$ , where  $\hat{S}$  is the (random) set picked out by the selection procedure, for example, the one exhibiting the largest  $AIC_{n,S}$  number. The behaviour of a large class of such mixed-situations estimators, which we may think of as frequentist model average estimators, is studied in this section. Our results are in particular used to pinpoint the over-optimistic nature of traditionally employed confidence interval, w.r.t. coverage probability.

**4.1. Compromise estimators.** To be able to single out submodels with more influence than others it is natural to employ  $\hat{\delta}_S = \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S})$  in a suitable form. We saw in Section 3.3 that the behaviour of  $\hat{\delta}_S$  is essentially determined by that of  $D_n = \hat{\delta}_{\text{full}}$  of (3.1). Which submodel is picked out by the AIC method, for example, is determined by  $D_n$ , see (3.4). This motivates studying the large class of compromise estimators, those taking the form

$$\hat{\mu} = \sum_S c(S | D_n) \hat{\mu}_S, \quad (4.1)$$

where the sum is potentially over all subsets of  $\{1, \dots, q\}$ , including the empty subset, which corresponds to the narrow model. The weight functions  $c(S | d)$  are required to sum to 1 for each  $d$ , since otherwise the estimator is not consistent. A special case would be  $\hat{\mu} = \sum_{k=0}^q c(\{1, \dots, k\} | D_n) \hat{\mu}_k$ , say, indicating a mixture of estimators  $\hat{\mu}_k$  constructed from the model with  $S = \{1, \dots, k\}$ . Estimators formed after using the AIC criterion for nested submodels would be of this type, for example.

For a general compromise estimator of type (4.1), and with  $H_S$  as in (3.2), define  $G(d) = K^{-1/2} \{ \sum_S c(S | d) H_S \} K^{1/2}$  and

$$\hat{\delta}(D) = G(D)^t D = K^{1/2} \left\{ \sum_S c(S | D) H_S \right\} K^{-1/2} D. \quad (4.2)$$

Then  $G(d)$  is a  $q \times q$  matrix of functions in  $d = (d_1, \dots, d_q)^t$ , and  $\hat{\delta}(D)$  is to be seen as an estimator of  $\delta$  based on  $D$ . Recall that  $D_n = \hat{\delta}_{\text{full}}$  tends to  $D \sim N_q(\delta, K)$  by (3.1).

**THEOREM 4.1.** *As long as the weight functions  $c(S | d)$  sum to 1 for each  $d$  and have at most a countable number of discontinuities,  $\sqrt{n}(\hat{\mu} - \mu_{\text{true}})$  tends under the (2.2) assumption in distribution to*

$$\Lambda = \sum_S c(S | D) \Lambda_S = \left( \frac{\partial \mu}{\partial \theta} \right)^t J_{00}^{-1} M + \omega^t \{ \delta - \hat{\delta}(D) \}.$$

Its mean and variance are  $\omega^t \{\delta - E\hat{\delta}(D)\}$  and  $\tau_0^2 + \omega^t \text{Var } \hat{\delta}(D) \omega$ , with mean squared error  $E\Lambda^2 = \tau_0^2 + R(\delta)$ , in which

$$R(\delta) = E(\omega^t \hat{\delta} - \omega^t \delta)^2 = \omega^t E\{\hat{\delta}(D) - \delta\} \{\hat{\delta}(D) - \delta\}^t \omega, \quad (4.3)$$

where  $\tau_0^2 = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta}$  and  $\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$ .

Densities of various  $\Lambda$ s are displayed in Figure 5.1, illustrating in particular the non-normal nature of the limit distributions.

**REMARK 4.1.** The theorem spells out what happens to compromise estimators under the (2.2) scenario. A reviewer has wondered whether this description is adequate, if the underlying framework also allows perturbations of the  $\theta$  part of the model. To investigate this issue, consider  $f_n(y) = f(y, \theta_0 + \eta/\sqrt{n}, \gamma_0 + \delta/\sqrt{n})$  instead of (2.2), where  $\eta = (\eta_1, \dots, \eta_p)^t$ , along with  $\mu_n = \mu(\theta_0 + \eta/\sqrt{n}, \gamma_0 + \delta/\sqrt{n})$ . Then Lemmas 3.1–3.3 may be generalised, leading to parallel statements involving say  $\tilde{C}_S$  and  $\tilde{D}_S$ , which now also depend on  $\eta$ , with a consequent expression for say  $\tilde{\Lambda}_S$ , the limit variable for  $\sqrt{n}(\hat{\mu}_S - \mu_n)$ . It turns out that  $\tilde{\Lambda}_S$  has the same distribution as before, independent of  $\eta$ . This shows that the description given in Theorem 4.1 continues to be adequate even when the  $\theta$  part of the model is being locally perturbed. ■

**REMARK 4.2.** The theorem was stated in a form focussing on  $D_n$  of (3.1) and its limit form  $D \sim N_q(\delta, K)$ . It is convenient, also for interpretational purposes, to rephrase in terms of

$$Z_n = \hat{K}^{-1/2} D_n = \hat{K}^{-1/2} \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0). \quad (4.4)$$

and its limit form  $Z = K^{-1/2} D \sim N_q(a, I)$ , via the link  $a = K^{-1/2}\delta$ . Note that  $Z_n \rightarrow_d N_q(a, I)$ . Here  $\hat{K}$  is any reasonable estimator of  $K$ ; it suffices that it is consistent for  $K$  under the null model  $\gamma = \gamma_0$ . Also, the weights of the compromise estimator  $c(S | D_n)$  may be seen as functions of  $Z_n$  rather than of  $D_n$ . For such compromise estimators  $\sum_S \bar{c}(S | Z_n) \hat{\mu}_S$ , the limiting distribution has risk  $\tau_0^2 + \bar{R}(a)$ , where  $\bar{R}(a) = E(\omega^t K^{1/2} \hat{a} - \omega^t K^{1/2} a)^2$  and  $\hat{a}(Z) = \sum_S \bar{c}(S | Z) H_S Z$ . This is viewed as an estimator of  $a$  on the canonical scale where  $Z \sim N_q(a, I)$ . Using this notation  $AIC_{n,S} = Z_n^t H_S Z_n - 2|S| + o_p(1)$ . ■

**REMARK 4.3.** Note that  $\pi_S^t K_S \pi_S$  is the  $q \times q$  matrix with the elements of  $K_S$  placed according to the indexes of the subset  $S$  and with zeroes elsewhere. It is also worthwhile recording the simpler structure that results in the special case of a diagonal  $K$  matrix. Then  $H_S$  is diagonal with values 1 for  $j \in S$  and 0 for  $j \notin S$ . Accordingly,

$$\hat{a}(z) = (W_1(z) z_1, \dots, W_q(z) z_q)^t \quad \text{where} \quad W_j(z) = \sum_S \bar{c}(S | z) I\{j \in S\} \quad (4.5)$$

in such situations. The limiting risk is  $\tau_0^2 + E[\sum_{j=1}^q \omega_j k_j^{1/2} \{W_j(Z) Z_j - a_j\}]^2$ . Eq. (4.5) also shows that different-looking compromise strategies may well have the same performance, for large  $n$ . Let for illustration  $q = 3$  with a diagonal  $K$ , with eight weight functions,

say  $\bar{c}_{000}(Z_n), \dots, \bar{c}_{111}(Z_n)$  with 0 and 1 indicating exclusion and inclusion of  $\gamma_1, \gamma_2, \gamma_3$  in the model. Then the performance of the procedure is determined by the three functions  $W_1 = \bar{c}_{100} + \bar{c}_{101} + \bar{c}_{110} + \bar{c}_{111}$ ,  $W_2 = \bar{c}_{010} + \bar{c}_{110} + \bar{c}_{011} + \bar{c}_{111}$ ,  $W_3 = \bar{c}_{001} + \bar{c}_{101} + \bar{c}_{011} + \bar{c}_{111}$ . ■

Theorem 4.1 spells out the drastic reduction in complexity by comparing model choice and estimation strategies in the large-sample limit experiment. Performances of such regimes are characterised fully by (4.3), in other words by a simpler estimation problem in a standard situation involving a multinormal  $D \sim N_q(\delta, K)$  with known variance matrix. Two viewpoints can be taken here. The first is that components  $\delta_1, \dots, \delta_q$  are being estimated simultaneously on the basis of  $D$ , with loss function  $\{\sum_{j=1}^q \omega_j(\hat{\delta}_j - \delta_j)\}^2$ . The alternative viewpoint is that only the one-dimensional parameter  $\psi = \omega^t \delta = \omega^t K^{1/2} a$  matters, and that this parameter has to be estimated under quadratic loss by estimators of the form

$$\hat{\psi} = \omega^t \hat{\delta}(D) = \omega^t K^{1/2} \left\{ \sum_S c(S | D) H_S \right\} K^{-1/2} D. \quad (4.6)$$

It is instructive to see the role of the parameter of interest  $\mu = \mu(\theta, \gamma)$ ; what is in the end a good model selection strategy or a regime for smoothing between models does depend on the parameter under study. This is perhaps only to be expected, but the point is often overlooked, in that the most popular model choice methods work independently of the inference to take place afterwards. See Claeskens and Hjort (2003) for applications where different estimands correspond to different optimal submodels.

Importantly, the theory developed above goes through also for the regression model cases, under mild regularity conditions of the type described in Section 3.4.

**4.2. Dwindling confidence.** The traditional use of model selection methods in practice is to proceed as if the finally selected model had been chosen a priori. Thus a typical confidence interval, taking intended coverage probability 90 per cent as an example, would take the form

$$\mu \in \hat{\mu}_{\hat{S}} \pm 1.645 \hat{\tau}_{\hat{S}} / \sqrt{n}, \quad (4.7)$$

where  $\hat{S}$  represents the chosen model and  $\hat{\tau}_{\hat{S}} / \sqrt{n}$  is an estimator of the standard deviation for  $\hat{\mu}_{\hat{S}}$ , without model uncertainty for  $S$ . From Lemma 3.3 and Theorem 4.1,  $\hat{\tau}_{\hat{S}}$  estimates  $\tau_S = (\tau_0^2 + \omega^t K^{1/2} H_S K^{1/2} \omega)^{1/2}$ . Such procedures ignore the uncertainties involved in the model selection step of the analysis and are consequently too optimistic about the confidence level attained by such intervals; similar comments apply to tests and other forms of inference. This is e.g. visible when one compares the optimistic standard deviation estimates of Table 1.1, for the AIC-chosen  $\mu$ -estimators, with the real ones, as found in Table 9.1.

Consider any selection-estimator of this type, where the model selection is being determined exactly or asymptotically via  $Z_n$  of (3.3). These correspond to compromise estimators (4.1) for which  $\mathcal{R}^q$  is partitioned into regions  $R_S$ , where  $c(S | z) = 1$  for  $z \in R_S$

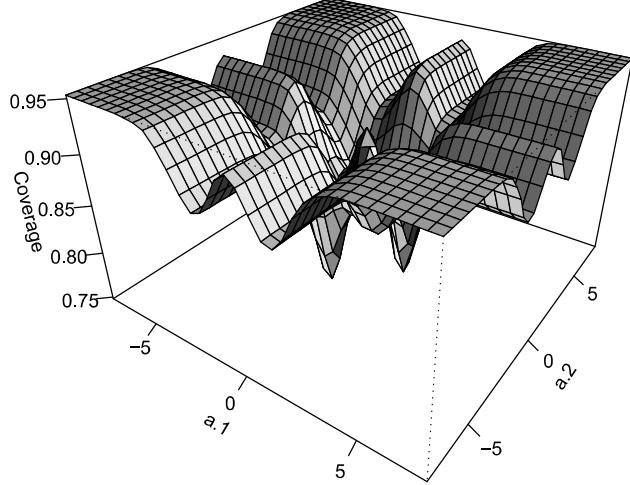


FIGURE 4.1. True coverage probability when ignoring AIC choice between four models, for  $q = 2$ , when  $\omega = (1, 1)^t$  and  $K = \text{diag}(1, 1)$ .

and zero outside. Now study

$$V_n = \sqrt{n}(\hat{\mu}_{\hat{S}} - \mu_{\text{true}})/\hat{\tau}_{\hat{S}}.$$

By previous efforts,  $V_n \rightarrow_d V = \Lambda/\tau(Z)$ , say, where  $\tau(z)^2 = \tau_0^2 + \omega^t K^{1/2} H_S K^{1/2} \omega$  for  $z \in R_S$ . Also, from the proof of Theorem 4.1,  $\Lambda | z$  is normal with variance  $\tau_0^2$  and mean  $\omega^t \{\delta - G(z)^t K^{1/2} z\}$ . Thus the real coverage probability of an interval like (4.7) goes for growing  $n$  to

$$p(a) = \Pr_a\{|V| \leq 1.645\} = \sum_S \int_{R_S} \Pr\{|\mathbb{E}(\Lambda | z) + \tau_0 N|/\tau_S \leq 1.645\} \phi(z - a) dz,$$

in which  $N$  denotes a standard normal variable and again  $a = K^{-1/2}\delta$ .

For  $q = 1$  these probabilities are easily calculated via numerical integration. For the AIC-selected estimator, one chooses the narrow estimator when  $|Z| \leq \sqrt{2}$  and the full one when  $|Z| > \sqrt{2}$ , and some algebra leads to  $p(a)$  being equal to

$$\int_{|z| \leq \sqrt{2}} \Pr\{|\rho a + N| \leq 1.645\} \phi(z - a) dz + \int_{|z| > \sqrt{2}} \Pr\left\{\frac{|\rho(z - a) + N|}{(1 + \rho^2)^{1/2}} \leq 1.645\right\} \phi(z - a) dz,$$

in terms of  $\rho = \omega K^{1/2}/\tau_0$ . This is often significantly smaller than the intended level 0.90. Figure 4.2 below displays the true coverage probability as a function of  $a$ , for AIC model choice between narrow and wide models. We have also carried out such computations for

the case of  $q = 2$ , using simulations. Figure 4.1 presents the coverage deficiency for AIC choice amongst four models in a situation where  $\omega = (1, 1)^t$  and  $K = \text{diag}(1, 1)$ . In the limit as  $\|a\| \rightarrow \infty$  correct coverage is obtained.

**4.3. Better confidence.** We have seen that the traditionally employed construction (4.7) leads to too optimistic intervals, in that the real coverage probability is lower than the intended level. Aware of this phenomenon, Buckland et al. (1997) have suggested a method for taking the extra model uncertainty into account, and which in particular leads to modified confidence intervals. Their method has later been embraced by Burnham and Anderson (2002, Section 4.3), particularly in conjunction with the smoothed AIC weights for  $c(S | D_n)$ , see Section 5.2. The method amounts to using  $\hat{\mu} \pm u \hat{s}_e$  as confidence intervals, with  $u$  the appropriate normal quantile and formula (9) in Buckland et al. for the estimated standard error  $\hat{s}_e$ . Rephrased to fit our framework,

$$\hat{s}_e = \sum_S c(S | D_n) (\hat{\tau}_S^2/n + \hat{b}_S^2)^{1/2},$$

in which  $\hat{\tau}_S$  is a consistent estimator of  $\tau_S = (\tau_0^2 + \omega^t K^{1/2} H_S K^{1/2} \omega)^{1/2}$  and  $\hat{b}_S = \hat{\mu}_S - \hat{\mu}$ . The resulting coverage probability  $p_n$  is not studied accurately in the references mentioned, but it is claimed that it will be close to the intended  $\Pr\{-u \leq N(0, 1) \leq u\}$ . Our methods make it possible to study  $p_n$  accurately, however. One has  $p_n = \Pr\{-u \leq B_n \leq u\}$ , where  $B_n = (\hat{\mu} - \mu_{\text{true}})/\hat{s}_e$ . This variable has a well-defined limit distribution, since  $\sqrt{n}\hat{s}_e \rightarrow_d \hat{s}_e = \sum_S c(S | D) \{\tau_S^2 + (\Lambda_S - \Lambda)^2\}^{1/2}$ , simultaneously with  $\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \rightarrow_d \Lambda$ , by an extension of arguments used in Section 11 to prove Theorem 4.1. Furthermore,  $\Lambda_S - \Lambda = \omega^t \{\hat{\delta}(D) - K^{1/2} H_S K^{-1/2} D\}$ . Thus

$$B_n \rightarrow_d B = \frac{\Lambda}{\hat{s}_e} = \frac{\Lambda_0 + \omega^t \{\delta - \hat{\delta}(D)\}}{\sum_S c(S | D) \{\tau_S^2 + [\omega^t \{\hat{\delta}(D) - K^{1/2} H_S K^{-1/2} D\}]^2\}^{1/2}}.$$

This variable is a normal, for given  $D$ , but is clearly not standard normal when averaged over the distribution of  $D$ , and neither is it centred at zero, so the coverage probability  $p_n$  is biased.

To illustrate this, consider the  $q = 1$  case, with compromise estimator  $\hat{\mu} = \{1 - W(Z_n)\}\hat{\mu}_{\text{narr}} + W(Z_n)\hat{\mu}_{\text{full}}$ , for which  $\Lambda_\emptyset = \Lambda_0 + \omega K^{1/2} a$  and  $\Lambda_{\text{full}} = \Lambda_0 + \omega K^{1/2}(a - Z)$ , writing  $\Lambda_0 = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} M$ . Here  $B = \Lambda/\hat{s}_e$  takes the form

$$\frac{\Lambda_0 + \omega K^{1/2}\{a - W(Z)Z\}}{\{1 - W(Z)\}\{\tau_0^2 + \omega^2 K W(Z)^2 Z^2\}^{1/2} + W(Z)\{\tau_0^2 + \omega^2 K + \omega^2 K\{1 - W(Z)\}^2 Z^2\}^{1/2}},$$

with  $\Lambda_0 \sim N(0, \tau_0^2)$  and independent of  $Z \sim N(a, 1)$ . The limiting coverage probability may then be computed via numerical integration, as  $p(a) = \int \Pr\{-u \leq B \leq u | z\} \phi(z - a) dz$ . See Figure 4.2.

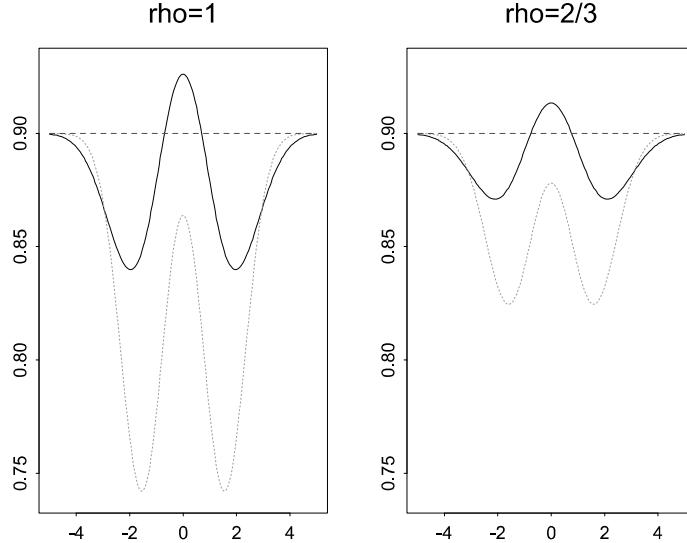


FIGURE 4.2. Exact limiting coverage probability  $p(a)$  for three confidence interval procedures, in two situations, corresponding to  $\rho = \omega K^{1/2}/\tau_0$  equal to 1 and  $2/3$ , for  $q = 1$ . The three methods are the AIC-based version of (4.7) (dotted line); the smoothed AIC method of Section 5.2 using  $\hat{s}_n$  described above as standard error (solid line); and finally the general (4.8) method, which gives correct 0.90 coverage for each method (dashed line).

Consider instead

$$\begin{aligned} \text{low}_n &= \hat{\mu} - \hat{\omega}^t \{D_n - \hat{\delta}(D_n)\} / \sqrt{n} - u\hat{\kappa} / \sqrt{n}, \\ \text{up}_n &= \hat{\mu} - \hat{\omega}^t \{D_n - \hat{\delta}(D_n)\} / \sqrt{n} + u\hat{\kappa} / \sqrt{n}, \end{aligned} \quad (4.8)$$

where  $\hat{\omega}$  and  $\hat{\kappa}$  are consistent estimators of  $\omega$  and  $\kappa = \tau_{\text{full}} = (\tau_0^2 + \omega^t K \omega)^{1/2}$ , and  $u$  is a normal quantile. We observe that the coverage probability  $p_n = \Pr\{\text{low}_n \leq \mu_{\text{true}} \leq \text{up}_n\}$  is the same as  $\Pr\{-u \leq T_n \leq u\}$ , where

$$T_n = [\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) - \hat{\omega}^t \{D_n - \hat{\delta}(D_n)\}] / \hat{\kappa}.$$

But there is simultaneous convergence in distribution

$$(\sqrt{n}(\hat{\mu} - \mu_{\text{true}}), D_n) \rightarrow_d (\Lambda_0 + \omega^t \{\delta - \hat{\delta}(D)\}, D),$$

essentially by the arguments used to prove Theorem 4.1. It follows that  $T_n \rightarrow_d \{\Lambda_0 + \omega^t \{\delta - \hat{\delta}(D)\}\} / \kappa$ , which is simply a standard normal. Thus, with  $u = 1.645$ , for example, the (4.8) interval has asymptotic confidence level precisely the intended 90% level.

**4.4. Example: Exponential within Weibull.** Let  $Y_1, \dots, Y_n$  come from the Weibull distribution with cumulative  $1 - \exp\{-(\theta y)^\gamma\}$ , with  $\gamma$  in the vicinity of  $\gamma_0 = 1$ . With some efforts one finds the information matrix, with inverse;

$$J = \begin{pmatrix} \gamma^2/\theta^2 & (1-r)\theta \\ (1-r)/\theta & c^2/\gamma^2 \end{pmatrix}, \quad J^{-1} = \frac{1}{\pi^2/6} \begin{pmatrix} c^2\theta^2/\gamma^2 & -(1-r)\theta \\ -(1-r)/\theta & \gamma^2 \end{pmatrix},$$

in which  $r = 0.5772\dots$  is the Euler–Mascheroni constant and  $c^2 = \pi^2/6 + (1 - r)^2$ . We consider estimators of the median  $\mu = (\log 2)^{1/\gamma}/\theta$  of the form

$$\widehat{\mu} = \{1 - W(Z_n)\}\widehat{\mu}_{\text{narr}} + W(Z_n)\widehat{\mu}_{\text{full}} = \{1 - W(Z_n)\}\frac{\log 2}{\widehat{\theta}_{\text{narr}}} + W(Z_n)\frac{(\log 2)^{1/\widehat{\gamma}_{\text{full}}}}{\widehat{\theta}_{\text{full}}},$$

where, following our recipe,  $Z_n = \sqrt{n}(\widehat{\gamma}_{\text{full}} - 1)/\widehat{K}^{1/2}$  with  $\widehat{K}$  estimating  $K = 6\gamma^2/\pi^2$ . Also,  $\omega = \mu\{-(1 - r) + \log q\}/\gamma^2$ , in terms of  $v = \log 2$ , and we find

$$\tau_0 = v^{1/\gamma}/(\theta\gamma), \quad (K\omega^2)^{1/2} = q^{1/\gamma}/(\theta\gamma)| - (1 - r) + \log v|\sqrt{6}/\pi.$$

When  $\gamma = 1 + \delta/\sqrt{n}$ , the limit distribution of  $\sqrt{n}(\widehat{\mu} - \mu_{\text{true}})$  is  $\Lambda = \Lambda_0 + \omega K^{1/2}\{a - W(Z)\}$ , where  $\Lambda_0 \sim N(0, \tau_0^2)$  and is independent of  $Z \sim N(a, 1)$ , and  $a = \delta/K^{1/2}$ .

We have carried out simulations in this example, for estimation of the median and other quantiles, using hard and smoothed AIC estimators, and yet further of the compromise estimators described in Section 5. The density of  $T_n$  above was seen to be quite close to its limiting standard normal density, for even moderate  $n$ . The coverage probability for the (4.8) intervals is consequently close to the intended level.

## 5. Some model average estimation schemes

In this section we go through a partial list of particularly attractive FMA methods. Different FMA schemes are characterised by their  $\delta$ -estimator and  $\psi$ -estimator counterparts  $G(D)^t D$  and  $\omega^t G(D)^t D$  in the limit experiment, as shown in the previous section. It is therefore often fruitful to construct FMA regimes via arguments inside the context of the limit experiment.

**5.1. The AIC selection-estimator.** For the AIC method with all  $2^q$  subsets allowed, let  $R_S$  be the set of  $D$  such that  $\text{AIC}_S(D)$  is larger than all other  $\text{AIC}_{S'}(D)$ , where

$$\text{AIC}_S(D) = D^t K^{-1/2} H_S K^{-1/2} D - 2|S| = Z^t H_S Z - 2|S|. \quad (5.1)$$

Then for  $D \in R_S$ ,  $c(S | D) = 1$  while the other  $c(S' | D) = 0$ . For the case of  $K$  being a diagonal matrix with diagonal elements  $k_j$ , we have  $\text{AIC}_S(D) = \sum_{j \in S} (D_{n,j}^2/k_j - 2)$ . This shows that, to the first order of large-sample approximation, precisely those  $j$  are included in the selected set for which  $D_{n,j}^2/\widehat{k}_j = n(\widehat{\gamma}_{\text{full},j} - \gamma_{0,j})^2/\widehat{k}_j > 2$ .

**5.2. A smoothed AIC-based estimator.** Buckland et al. (1997) make a general model averaging suggestion that amounts to taking weights  $c^*(S | \text{data})$  proportional to  $\exp(\ell_S - |S|)$ , where  $\ell_S$  is the maximised log-likelihood at model  $S$ . Thus comparing weights for models of the same complexity corresponds to likelihood ratio methods, and the penalisation of these otherwise ad hoc constructed terms stem from the analogy with the AIC method. By (3.4), the smoothed AIC weights may be represented as

$$\frac{\exp(\frac{1}{2}\text{AIC}_{n,S})}{\sum_{\text{all } S'} \exp(\frac{1}{2}\text{AIC}_{n,S'})} = \frac{\exp(\frac{1}{2}D_n^t K^{-1/2} H_S K^{-1/2} D_n - |S|)}{\sum_{\text{all } S'} \exp(\frac{1}{2}D_n^t K^{-1/2} H'_S K^{-1/2} D_n - |S'|)} + o_p(1). \quad (5.2)$$

It follows from the theory developed in Section 4 that the large-sample distributions of compromise estimators are the same, whether one uses the left hand side ratio or the right hand side ratio as weights. Note also that there is some independent motivation for using such weights from a Bayesian analogy, where  $\exp(\frac{1}{2}\text{BIC}_S)/\sum_{S'} \exp(\frac{1}{2}\text{BIC}_{S'})$  is known to be an approximation to the posterior probability of model  $S$  being correct; see Schwarz (1978), as well as discussion in Burnham and Anderson (2002, Section 6.4). Results developed in Section 8 below lead to other and potentially better approximations.

Using the theory developed in Section 4, the limiting distribution is a suitable convex mixture of normals, and the limiting squared error can be computed via (4.3). Buckland et al. partly motivated their method by considering correlations between different estimators, but without estimating these correlations accurately. We may show, using arguments of Section 4, that the limiting correlation between submodel estimators  $\hat{\mu}_S$  and  $\hat{\mu}_{S'}$  is

$$\rho(S, S') = \frac{\tau_0^2 + \omega^t K^{1/2} H_S H_{S'} K^{1/2} \omega}{(\tau_0^2 + \omega^t K^{1/2} H_S K^{1/2} \omega)^{1/2} (\tau_0^2 + \omega^t K^{1/2} H_{S'} K^{1/2} \omega)^{1/2}}. \quad (5.3)$$

We may also derive the limiting correlation between any two compromise estimators via similar arguments. Its size depends in particular on the relative sizes of  $\tau_0$  and  $(\omega^t K \omega)^{1/2}$ .

**5.3. The FIC selection-estimator.** The AIC method selects one winning model, regardless of the intended use for this model. In contrast, Claeskens and Hjort (2003) develop a focussed information criterion that specifically aims at finding the best candidate model for a given focus parameter  $\mu$ . While the AIC method chooses  $S$  to maximise  $\text{AIC}_S(D)$  of (5.1), the FIC goes for  $S$  to minimise

$$\text{FIC}_S(D) = (\omega^t D - \hat{\psi}_S)^2 + 2\omega^t K_S \omega_S, \quad \text{where } \hat{\psi}_S = \omega^t K^{1/2} H_S K^{-1/2} D.$$

This is the limit experiment version of the FIC. In practice one plugs in estimates of  $\omega$ ,  $K$ ,  $K_S$  and  $H_S$ . Suppose for example that the choice is only between the narrow and the full models. Then AIC selects the full model provided  $D^t K^{-1} D \geq 2q$ , while the corresponding FIC criterion for selecting the full model is  $(\omega^t D)^2 \geq 2\omega^t K \omega$ .

It is also attractive to smooth across estimators using the information carried by the FIC scores, and we suggest using

$$c(S | D) = \exp\left(-\frac{1}{2}\kappa \frac{\text{FIC}_S}{\omega^t K \omega}\right) / \sum_{S'} \exp\left(-\frac{1}{2}\kappa \frac{\text{FIC}_{S'}}{\omega^t K \omega}\right) \quad \text{with } \kappa \geq 0. \quad (5.4)$$

Here  $\kappa$  is an algorithmic parameter, bridging from uniform weighting ( $\kappa$  close to zero) to the hard-core FIC (which is the case of large  $\kappa$ ). Of course the added  $\frac{1}{2}$  is somewhat redundant, but the form (5.4) is suggested by connections to certain empirical Bayes arguments that can be developed using the theory of Section 9. The  $\omega^t K \omega$  factor appearing in the scaling for  $\kappa$  is the constant risk of the minimax estimator  $\hat{\delta} = D$ . The point of the scaling is that

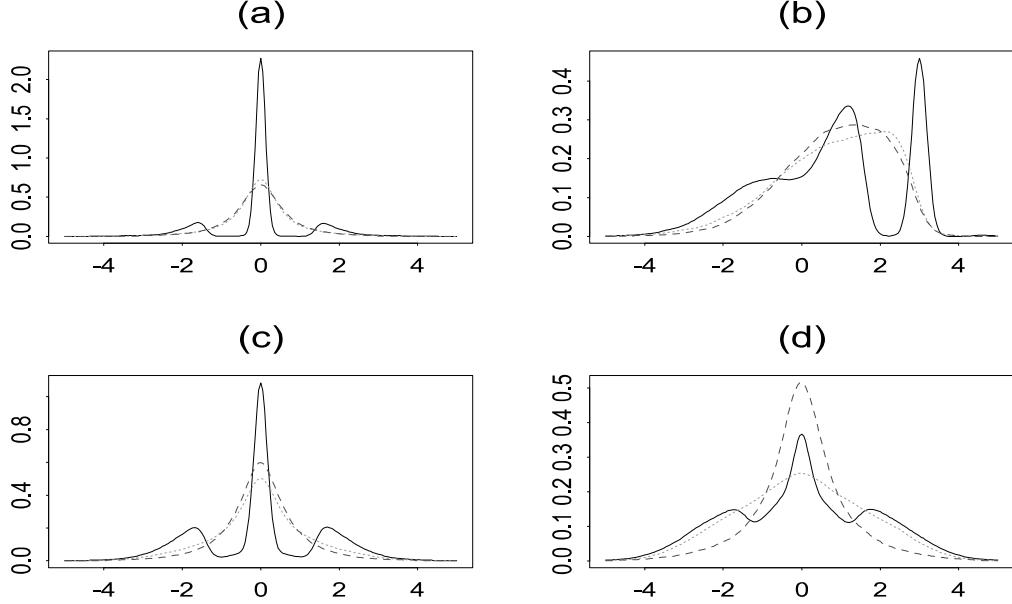


FIGURE 5.1. Density of the limiting distribution  $\Lambda$  of  $\sqrt{n}(\hat{\mu} - \mu_{\text{true}})$ , for three compromise estimators, at four positions in the parameter space. The situation studied has  $q = 2$ ,  $K = \text{diag}(1, 1)$ ,  $\omega = (1, 1)^t$  and  $\tau_0 = 0.5$ , and the four positions are (a)  $(0, 0)$ , (b)  $(1.5, 1.5)$ , (c)  $(1, -1)$ , (d)  $(2, -2)$  for  $a = (a_1, a_2)$ . The estimators are post-AIC (smooth line), smoothed AIC (dotted line), and smoothed FIC with  $\kappa = 1$  in (5.4) (dashed line).

$\kappa$  values used in different data contexts now can be compared directly. One may show here that for the one-dimensional case  $q = 1$ , the value  $\kappa = 1$  makes the weights of (5.4) agree with those of for the smoothed AIC.

To illustrate the limiting distribution of some compromise estimators, we display in Figure 5.1 the density of  $\Lambda$  for a situation with  $q = 2$  extra parameters, where  $K = \text{diag}(1, 1)$ ,  $\omega = (1, 1)^t$ , and  $\tau_0 = 0.5$ . The non-normal nature is evident, not only for non-smooth methods like the AIC, but also for smoothed versions thereof. For each of the four positions in the parameter space considered here, the smoothed FIC wins over the others in terms of mean squared error.

**5.4. Minimising estimated risk.** Consider estimators of the form  $\sum_S c(S)\hat{\mu}_S$ , with non-random weights summing to 1. From previous results, the limiting distribution in question is that of  $\Lambda = \sum_S c(S)\Lambda_S$ , with  $\Lambda_S$  as in Lemma 3.3. One finds  $E\Lambda = \omega^t(I - Q)^t\delta$ , where  $Q = \sum_S c(S)K^{-1/2}H_SK^{1/2}$ , and furthermore  $\text{Var } \Lambda = \tau_0^2 + \omega^tQ^tKQ\omega$ , using the covariance extension of Lemma 3.3 which was also used in connection with (5.3). Thus the limiting risk of the estimator is  $\tau_0^2 + R(\delta)$ , where  $R(\delta) = \omega^t\{(I - Q)^t\delta\delta^t(I - Q) + Q^tKQ\}\omega$ . This also agrees with (4.3).

Various model average estimators may now be constructed, along the following lines. Estimate the risk  $R(\delta)$ , for example by inserting  $D$  for  $\delta$  above, or alternatively the unbiased  $DD^t - K$  for  $\delta\delta^t$ . Then select weights  $c(S)$  to minimise this estimated risk. Different

versions emerge from this, depending also on the list of submodels one wishes to smooth across. A simple special case worth recording is that of smoothing optimally between the two extreme models,  $\hat{\mu} = (1 - c)\hat{\mu}_{\text{narr}} + c\hat{\mu}_{\text{full}}$ . Using  $\hat{\psi} = \omega^t D$  as estimator of  $\psi = \omega^t \delta$  in the full model, for the current purpose of estimating the optimal weights, the result for the limit experiment situation is

$$\hat{\mu} = \frac{\omega^t K \omega}{\hat{\psi}^2 + \omega^t K \omega} \hat{\mu}_{\text{narr}} + \frac{\hat{\psi}^2}{\hat{\psi}^2 + \omega^t K \omega} \hat{\mu}_{\text{full}}. \quad (5.5)$$

With real data, one in addition plugs in estimates of  $\omega$  and  $K$  and uses  $\hat{\psi} = \hat{\omega}^t D_n$ .

**5.5. Smoothing across singletons.** An attractive challenge is to form data-based averages over estimators  $\hat{\mu}_{\{j\}}$ , corresponding to the simple one-parameter model extensions of the narrow model. These take the form  $\sum_{j=0}^q c(j | D_n) \hat{\mu}_{\{j\}}$ , where  $j = 0$  corresponds to the narrow model estimator, and might be thought of as resembling first-stage Taylor expansion estimators. These  $\mu$ -estimators are further related to  $\delta$ -estimators of the form  $\hat{\delta} = G(D)^t D$  in the limit experiment, where  $G(D)$  is as in (4.2) but engaging only  $H_S$  matrices for  $S$  being empty or a singleton.

For brevity we present only one of these methods, which has been seen to perform well in some limited simulation exercises of the authors. This method emerges from Bayesian and empirical Bayesian considerations, starting with a prior which has  $\delta = 0$  with some probability  $p_0$  and with probability  $p_j$  has  $\delta_j$  from a normal and the other  $\delta_i$ s equal to zero, and where  $\sum_{j=0}^q p_j = 1$ . The estimator is

$$\hat{\mu} = (1 - \hat{\rho})\hat{\mu}_{\text{narr}} + \hat{\rho} \sum_{j=1}^q \frac{\exp(\frac{1}{2}\hat{\rho}k^{jj}\hat{T}_j^2)}{\sum_{i=1}^q \exp(\frac{1}{2}\hat{\rho}k^{ii}\hat{T}_i^2)} \hat{\mu}_{\{j\}}, \quad \text{where } \hat{\rho} = \frac{\hat{\tau}^2}{1 + \hat{\tau}^2}, \quad (5.6)$$

with  $\hat{\tau} = (D_n^t \hat{K}^{-1} D_n - q)_+^{1/2}$  and  $\hat{T}_j = (\hat{k}^{jj})^{-1} e_j^t \hat{K}^{-1} \hat{\delta}_{\text{full}}$ , in terms of the diagonal elements of  $\hat{K}^{-1}$  and the  $j$ th unit vector  $e_j$ . Details of this construction, along with useful variations, are available in a technical report from the authors.

**5.6. An empirical Bayes model smoother.** The following arguments motivate a particular estimator-smoother, with data-dependent weights  $c(S | D_n)$  in (4.1). The idea is to start with a Bayesian mixture prior, of a more general type than that used in Section 5.5, then work out the necessary details pertaining to the posterior, and finally estimate the required spread parameter from the marginal distribution of data.

**REMARK 5.1.** We use this opportunity to make the following general point. Our theory has been developed by the desire to handle averages of subset estimators  $\hat{\mu}_S$  of the form (2.1), i.e. for subsets of the original  $(\theta, \gamma)$  or  $(\theta, \delta)$  parameterisation of the fullest model. Mathematically we are free to reparametrise from  $\delta$  to the canonical  $a = K^{-1/2} \delta$  scale, however, and instead work with subset estimators  $\mu_S^* = \mu(\theta_S^*, a_S^*, 0_{S^c})$  and averages  $\mu^* = \sum_S c^*(S | Z_n) \mu_S^*$ . The advantage is a cleaner orthogonal structure, since

$Z \sim N_q(a, I)$ . The theory of Sections 3 and 4 would go through with minor changes. We illustrate this here, since the mixture strategy becomes easier to develop and describe. ■

Focus first on one of the  $a_j$  components, and let it be zero with probability  $p_0$  and a  $N(0, \sigma^2)$  with probability  $p_1$ . Then  $a_j | z_j$  is zero with probability  $\tilde{p}_0(z_j)$  and from a  $N(\rho z_j, \rho)$  with probability  $\tilde{p}_1(z_j)$ , where  $\rho = \sigma^2 / (1 + \sigma^2)$ . Furthermore,

$$\tilde{p}_1(z_j) = \frac{p_1 \phi(z_j, 1 + \sigma^2)}{p_0 \phi(z_j, 1) + p_1 \phi(z_j, 1 + \sigma^2)} = \frac{p_1 (1 + \sigma^2)^{-1/2} \exp(\frac{1}{2} \rho z_j^2)}{p_0 + p_1 (1 + \sigma^2)^{-1/2} \exp(\frac{1}{2} \rho z_j^2)}$$

with  $\tilde{p}_0(z_j) = 1 - \tilde{p}_1(z_j)$  and  $\phi(z, v^2)$  the  $N(0, v^2)$  density evaluated at  $z$ . If now  $a_1, \dots, a_q$  are given independent priors of this type, which is reasonable in that the  $a_j$ s have been transformed towards orthogonality and the same scale, then  $E(a_j | z) = \rho \tilde{p}_1(z_j) z_j$  for  $j = 1, \dots, q$ . By the general recipe established in Remark 4.3, we should have  $\hat{a}_j = W_j(z) z_j$  for  $W_j(z) = \sum_{S: j \in S} c^*(S | z)$ . But this fits in with the compromise regime that uses

$$c^*(S | z) = \rho \prod_{j=1}^q \tilde{p}_0(z_j)^{I\{j \notin S\}} \tilde{p}_1(z_j)^{I\{j \in S\}} \quad \text{for non-empty } S$$

and  $c(\emptyset | z) = 1 - \rho + \rho \prod_{j=1}^q \tilde{p}_0(z_j)$ . A fruitful variation is the empirical Bayes construction which inserts an estimate  $\hat{\sigma}$  for  $\sigma$  in the  $c^*(S | z)$  formulae above. Such an estimate may emerge from likelihood analysis based on the marginal distribution of  $(Z_1, \dots, Z_q)$ . One may also use a hyper-prior for  $\sigma$  in a two-stage Bayesian fashion. It suffices for the present purposes to devise a simple moment estimator, however, using that  $\sum_{j=1}^q Z_j^2$  has mean  $q + qp_1\sigma^2$ . We therefore propose  $\hat{\sigma}^2 = (\sum_{j=1}^q Z_j^2 - q)_+ / (qp_1)$ , where the positive part notation indicates that  $\hat{\sigma} = 0$  in the case of  $\sum_{j=1}^q Z_j^2 \leq q$ . Such an event suggests that none of the  $a_j$ s are significantly non-zero, and the scheme selects the narrow model.

Several variations of these arguments could be considered. One may e.g. use a vague hyper-prior for the  $\sigma$  parameter. Another alternative is to estimate both  $\sigma$  and  $p_1$  in the above construction based on the marginal distribution of  $(Z_1, \dots, Z_q)$ , which obviates the need to specify  $p_1$  in advance.

## 6. Illustrations and applications

**6.1. Computational aspects.** Frequentist model averaging analysis can be easily performed using standard statistical software. All numerical results presented in this article are obtained using the software packages S-Plus and R.

Obtain parameter estimates in the different models and either compute the model selection criterion value for each of these models in order to form the indicator variable of the optimal model, or directly construct the general model averaging weights of choice. From the estimate in the biggest model we construct  $\hat{\delta}$ . Nonlinear optimisation algorithms, such as `nlm()` in R, provide us immediately with a matrix of second order partial derivatives,

leading to the matrix  $\hat{J}_{\text{full}}$ . Next we construct the projection matrices  $\pi_S$ , and use these to define  $\hat{K}_S$  and  $\hat{H}_S$ , for each model  $S$ . Partial derivatives of  $\mu$  w.r.t.  $\theta$  parameters and  $\gamma$  parameters, at either  $(\hat{\theta}_{\text{narr}}, \gamma_0)$  or  $(\hat{\theta}_{\text{full}}, \gamma_0)$ , are needed for the computation of  $\hat{\omega}$  and  $\hat{\tau}_0$ . These are sometimes easy to derive mathematically, and can otherwise be computed using numerical derivatives.

As far as our theoretical results are concerned we may use any  $J_{\text{full}}^*$  estimator for the crucial matrix  $J_{\text{full}}$  of Section 3.1 (along with the consequent estimators for  $K, K_S, H_S$ ), as long as it is consistent under our  $\gamma_0 + \delta/\sqrt{n}$  framework. In particular, it may be computed under ‘narrow’ or ‘full’ circumstances. Narrow estimation is sometimes easiest, via explicit formulae or via simulation of score vectors under the null model. To guard against cases where  $\delta$  is some distance away from zero, however, it will be more satisfactory and robust to use full-model estimation; see also parallel discussion of this in Claeskens and Hjort (2003).

We have found it useful in practice to simulate the limit distribution  $\Lambda$  of Theorem 4.1, for the average estimator scheme being used, at  $\delta$  corresponding to its estimate  $\hat{\delta}_{\text{full}}$ . A density estimate of say 10,000 such  $\Lambda$  copies is informative, and leads to estimated bias and standard deviation for the compromise estimator being used, as well as to approximative confidence intervals.

**6.2. Averaging over logistic regression models.** Time has come to revisit the 189 babies of Section 1.2. Here we illustrate our general methodology by exhibiting results for each of the three focus parameters  $p(\text{white})$ ,  $p(\text{black})$  and their ratio  $p(\text{black})/p(\text{white})$ , for six different FMA regimes. These are the AIC-selected estimator, the smooth-AIC of Section 5.2, the FIC-selected, the smooth-FIC of (5.4) with  $\kappa = 1$ , the smoothing across singletons which makes data-dictated compromises between the four models ‘0’, ‘3’, ‘4’, ‘5’, and finally the simple compromise between narrow and full models as in (5.5).

	(a)	(b)	(c)	(d)	(e)	(f)
For $p(\text{white})$ :						
estimate	0.269	0.261	0.263	0.258	0.281	0.242
stdev	0.051	0.047	0.050	0.045	0.039	0.048
lower	0.174	0.168	0.173	0.165	0.173	0.191
upper	0.343	0.322	0.338	0.315	0.302	0.350
For $p(\text{black})$ :						
estimate	0.412	0.368	0.412	0.365	0.323	0.380
stdev	0.112	0.107	0.115	0.106	0.107	0.096
lower	0.257	0.216	0.257	0.215	0.203	0.190
upper	0.618	0.559	0.614	0.553	0.549	0.501
For the ratio $p(\text{black})/p(\text{white})$ :						
estimate	1.534	1.440	1.564	1.501	1.159	1.651
stdev	0.668	0.610	0.647	0.582	0.516	0.567
lower	0.681	0.640	0.712	0.779	0.843	0.384
upper	2.792	2.563	2.758	2.619	2.414	2.299

TABLE 6.1. For each of the three focus parameters associated with the study of low birth weights described in Section 1.2, the table gives parameter estimate and estimated standard deviation, along with lower and upper points for 90% confidence intervals for the FMA strategies: (a) AIC, (b) smooth-AIC, (c) FIC, (d) smooth-FIC, (e) smoothing across singletons, and (f) compromise between narrow and full model.

We record here that for  $p(\text{white})$ ,  $\hat{\omega} = (-0.245, 0.032, 0.065)$  and  $\hat{\tau}_0 = 0.477$ ; for  $p(\text{black})$ ,  $\hat{\omega} = (0.429, -0.185, 0.073)$  and  $\hat{\tau}_0 = 0.550$ ; while for the ratio parameter,  $\hat{\omega} = (3.783, -10.057, -0.190)$  with  $\hat{\tau}_0 = 0.495$ . We observe that the averaging across singletons method leads to low standard deviation and short confidence intervals. This is particularly noticeable for the ratio parameter. The standard deviations and the confidence bounds come from 10,000 simulations of the appropriate  $\Lambda$  distributions.

**6.3. Averaging over covariance structure models.** There are no inherent problems with applying our methodology to situations where data are multi-dimensional. To illustrate this we report on a brief investigation of multinormal data where different models for the covariance structure, in the absence of clear a priori preferences, are being averaged over to form estimators of quantities of interest. We note that there are several areas of statistics where covariance modelling is of interest, and sometimes perhaps of primary concern, as with factor analysis, and where variations of our methods might be fruitful.

Assume one has observed  $d$ -dimensional vectors  $Y = (X_1, \dots, X_d)^t$  from the multinormal  $N_d(\xi, \Sigma)$ , where different models for the structure of  $\Sigma$  are being considered. As a specific example, we use data from the so-called Adelskalenderen of speedskating. This is the list of the best speedskaters ever, as ranked by their personal bests over the four distances 500, 1500, 5000, 10000 m, via the classical point-sum  $X_1 + X_2 + X_3 + X_4$ , where  $X_1$  is the 500 m time,  $X_2$  is the 1500 m time divided by 3,  $X_3$  is the 5k time divided by 10, and  $X_4$  the 10k time divided by 20. The correlation structure of the 4-vector  $Y$  is important when relating, discussing and predicting performances on different distances. While there is a long list of parameters  $\mu = \mu(\xi, \Sigma)$  that on occasions will ignite the fascination of speedskating fans, for this discussion we single out as focus parameters the generalised standard deviation measures  $\mu_1 = \{\det(\Sigma)\}^{1/8}$  and  $\mu_2 = \{\text{Tr}(\Sigma)\}^{1/2}$ , the average correlation  $\mu_3 = \frac{1}{6} \sum_{i < j} \text{corr}(X_i, X_j)$ , and the so-called maximal correlation  $\mu_4$  between  $(X_1, X_2, X_3)$  and  $X_4$ . The latter is the maximal correlation between a linear combination of  $X_1, X_2, X_3$  and  $X_4$ , and is e.g. of interest at championships when one tries to predict the final outcomes, after the completion of the three first distances. It is also equal to  $(\Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01}/\Sigma_{11})^{1/2}$ , in terms of the blocks of  $\Sigma$ , of size  $3 \times 3$  for  $\Sigma_{00}$  and so on. Below we analyse the top of the Adelskalenderen, with the best  $n = 250$  skaters ever, as per the end of the 2002 season. The vectors  $Y_1, \dots, Y_n$  are by definition ranked, but as long as one discusses estimators that are permutation-invariant we may view the data vectors as a random sample from the population of the top skaters of the world.

A minimal plausible model for  $\Sigma$  is  $M_0$ , which has equicorrelation and equal variances.

Model  $M_1$  assumes equicorrelation though allows the variances to be different, while  $M_2$  has no pre-imposed structure on the correlations but does assume equal variances. The fullest model  $M_3$  is the unstructured covariance matrix with 10 parameters. To place this setting into the framework developed in earlier sections, let

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho\phi_2 & \rho\phi_3(1 + \nu_{13}) & \rho\phi_4(1 + \nu_{14}) \\ \rho\phi_2 & \phi_2^2 & \rho\phi_2\phi_3(1 + \nu_{23}) & \rho\phi_2\phi_4(1 + \nu_{24}) \\ \rho\phi_3(1 + \nu_{13}) & \rho\phi_2\phi_3(1 + \nu_{23}) & \phi_3^2 & \rho\phi_3\phi_4(1 + \nu_{34}) \\ \rho\phi_4(1 + \nu_{14}) & \rho\phi_2\phi_4(1 + \nu_{24}) & \rho\phi_3\phi_4(1 + \nu_{34}) & \phi_4^2 \end{pmatrix}.$$

The parameter  $\theta = (\sigma^2, \rho)$  is present in all of the models while subsets of  $\gamma = (\phi_2, \phi_3, \phi_4, \nu_{13}, \nu_{14}, \nu_{23}, \nu_{24}, \nu_{34})$  are present in some of the models. Here  $\gamma_0 = (1, 1, 1, 0, 0, 0, 0, 0)$ . We use the criteria AIC and FIC to select an appropriate covariance structure.

For the models described above we get the following parameter estimates. Note that FIC depends on the parameter under focus and hence gives different values for different  $\mu_k$ s. On this occasion, the FIC for parameter  $\mu_2$  points to model  $M_1$ , while FIC selects model  $M_3$  for all other parameters, as does the AIC. Also presented in the table are the model averaged estimates using smoothed AIC and FIC weights, using weights as with (5.2) and (5.4), where for the latter  $\kappa = 1$ . Confidence intervals are constructed using (4.8) with observed value of  $D_n = \hat{\delta}_{\text{full}}$  equal to  $(-1.624, -0.308, 5.948, -16.482, -22.861, -7.306, -15.478, -0.892)^t$ . At nominal level 90% we find for the  $\mu_2$  parameter  $(2.390, 2.611)$  for FIC,  $(2.254, 2.474)$  for both AIC and smooth-AIC, and  $(2.339, 2.559)$  for smooth-FIC.

	$M_0$	$M_1$	$M_2$	$M_3$	sm-AIC	sm-FIC
$\hat{\mu}_1$	1.146	1.101	0.844	0.816	0.816	0.816
$\hat{\mu}_2$	2.364	2.389	2.461	2.364	2.364	2.381
$\hat{\mu}_3$	0.225	0.271	0.388	0.262	0.262	0.263
$\hat{\mu}_4$	0.324	0.378	0.751	0.810	0.810	0.796

TABLE 6.2. Six different estimates of the parameters  $\mu_1, \mu_2, \mu_3, \mu_4$ . These correspond to models  $M_0, M_1, M_2, M_3$ , and to AIC-smoothed and FIC-smoothed averages thereof, as per Sections 5.2–5.3.

Following the computational steps in Section 6.1, we use simulation to compute the standard deviation of the estimators, for post model selection estimation by AIC and FIC, as well as for the smoothed versions. For the  $\mu_2$  parameter, for example, we have the following estimated standard deviations, for the different methods: the same value 1.029 for post-AIC and smooth-AIC, while for the post-FIC and smooth-FIC the value is 1.009.

**6.4. Variable selection and model smoothing in linear regression.** Assume that observations  $Y_i$  are to be regressed w.r.t. regressors  $x_{i,1}, \dots, x_{i,p}$  and possibly w.r.t. a further subset of additional regressors  $u_{i,1}, \dots, u_{i,q}$ . Which subset of these ought to be included, and which ways are there of averaging over all models? The natural framework is that of  $Y_i = \alpha + x_i^t \beta + u_i^t \gamma + \varepsilon_i$  for  $i = 1, \dots, n$ , where the  $\varepsilon_i$ s are independent and  $N(0, \sigma^2)$ . Suppose that the  $u_i$ s have been made orthogonal to the  $x_i$ s, in the sense that  $n^{-1} \sum_{i=1}^n x_i u_i^t = 0$ .

Then

$$J_{n,\text{full}} = \sigma^{-2} \text{diag}(2, \Sigma_{00}, \Sigma_{11}) \quad \text{with} \quad J_{n,\text{full}}^{-1} = \sigma^{-2} \text{diag}\left(\frac{1}{2}, \Sigma_{00}^{-1}, \Sigma_{11}^{-1}\right),$$

where  $\Sigma_{00} = n^{-1} \sum_{i=1}^n x_i x_i^t$  and  $\Sigma_{11} = n^{-1} \sum_{i=1}^n u_i u_i^t$ . Inside this framework we may now study model selection and model averaging for different focus parameters, using methods developed in earlier sections. For the arguably most important case of  $\mu = E(Y | x, u)$  at some given location  $(x, u)$ , FMA estimators take the form

$$\hat{\mu}(x, u) = \sum_S c(S | D_n)(x^t \hat{\beta}_S + u_S^t \hat{\gamma}_S) = x^t \beta^* + u^t \gamma^*,$$

where the  $\beta_j^*$ s and  $\gamma_k^*$ s involved are non-linear regression coefficient estimates. Methods and results of earlier sections can be used to settle on weighting schemes here, along with proper analysis of performances.

## 7. Risk comparison

We are now in position to compare various model selection-estimation and model averaging methods in terms of performance.

*7.1. Comparing risks in the limit experiment.* In a given situation the risk function  $n E(\hat{\mu} - \mu_{\text{true}})^2$  can be a quite complicated quantity, particularly when the estimator in question uses non-linear weight schemes and when the underlying models are difficult. There is a drastic reduction in complexity as  $n$  grows, however, as spelled out in Section 4, in that the limiting risk  $\tau_0^2 + R(\delta)$  depends on only a few crucial quantities. This allows broad comparisons to be made in a fairly easy fashion, by computing risk functions for situations and estimation schemes of interest, in their reduced limit experiment form.

It is often convenient to discuss performance in terms of  $\bar{R}(a)$  instead of  $R(\delta)$ , since  $a = K^{-1/2}\delta$  is scale-independent with  $Z \sim N_q(a, I)$ . Note in this connection Remark 5.1 about reparametrisation, which makes it possible to have  $K$  diagonal if one works with submodels represented by subsets of  $(a_1, \dots, a_q)^t$ . Also note that when  $K$  is diagonal,

$$\bar{R}(a) = \sum_{i,j} \omega_i \omega_j k_i^{1/2} k_j^{1/2} [V_{i,j}(a) + \{M_i(a) - a_i\} \{M_j(a) - a_j\}] \quad (7.1)$$

in terms of the means  $M_i(a)$  of  $W_i(Z)Z_i$  and covariance  $V_{i,j}(a)$  of  $W_i(Z)Z_i$  with  $W_j(Z)Z_j$ . This follows from Remark 4.3, and shows that even complicated risk functions may be computed easily via simulation. Before we go on to a briefly annotated list of estimators we mention one more fact, namely that for the simplest case of  $q = 1$  model extension,

$$\bar{R}(a) = K \omega^2 R^*(a) \quad \text{in terms of} \quad R^*(a) = E\{W(Z)Z - a\}^2. \quad (7.2)$$

This is the one-dimensional risk function for the estimator  $W(Z)Z$  for  $a$  in the standard experiment where  $Z \sim N(a, 1)$ . Here  $1 - W(Z_n)$  and  $W(Z_n)$  are the weights given to  $\hat{\mu}_{\text{narr}}$  and  $\hat{\mu}_{\text{full}}$ . Such  $R^*(a)$  functions are displayed in Figure 7.1 for various competing schemes.

(i) *Narrow estimation.* Here  $\bar{c}(\emptyset | Z) = 1$  and  $\bar{c}(S | Z) = 0$  for other subsets, reflecting the optimistic belief, or blissful ignorance, that  $a = 0$ . The limiting risk is  $\bar{R}_{\text{narr}}(a) = (\omega^t \delta)^2 = (\omega^t K^{1/2} a)^2$ , which is unbounded and quickly becomes big in size. The risk is satisfactorily small when  $\|a\|$  is small (and in that case for all estimands  $\mu$ ), or in cases where  $a$  is nearly orthogonal to  $K^{1/2}\omega$  (which depends on the estimand).

(ii) *Wide model estimation.* Here  $\bar{c}(S | Z) = 1$  for the full set, which leads to a constant minimax risk,  $\bar{R}_{\text{full}}(a) = \omega^t K \omega$ . This is satisfactory performance in some situations, but the estimator is often too guardedly pessimistic, losing out to methods that take into account that  $a$  could be small in size or have low correlation with  $K^{1/2}\omega$  (making in that case  $\psi = \omega^t \delta$  small in size).

(iii) *Hard and smooth AIC-selection estimators.* The following comments are valid for the non-nested case. For simplicity of illustration also take  $K$  to be diagonal, in which case the AIC scores can be written  $\sum_{j \in S} (Z_j^2 - 2)$  in terms of  $Z_j = D_j/k_j^{1/2}$ . This entails a quite simple structure for the  $R_S$  regions, as the winning  $S$  is  $\{j : |z_j| > \sqrt{2}\}$ . Turning this around, one sees that

$$R_S = \{z : |z_j| > \sqrt{2} \text{ for each } j \in S \text{ and } |z_j| \leq \sqrt{2} \text{ for each } j \notin S\}.$$

In particular,  $R_\emptyset = [-\sqrt{2}, \sqrt{2}]^q$  is the set inside which the method selects the narrow model. This also leads to  $W_j(z) = I\{|z_j| > \sqrt{2}\}$  in (4.5), making it possible to calculate  $\bar{R}(a)$  of (7.1) explicitly. We have done this in some further numerical comparison work, not reported on here due to limitations of space. The smoothed AIC scheme described in Section 5.2 uses weights proportional to  $\exp\{\frac{1}{2} \sum_{j \in S} (z_j^2 - 2)\}$ , as opposed to the hard thresholding  $I\{|z_j| > \sqrt{2}\}$  involved in ordinary AIC.

(iv) *Hard and smooth FIC-selection estimators.* The limiting risk functions for the two compromise estimators which use respectively the AIC and the FIC become  $\tau_0^2 + \bar{R}_{\text{aic}}(a)$  and  $\tau_0^2 + \bar{R}_{\text{fic}}(a)$ , where  $\bar{R}_{\text{fic}}(a) = E\{\omega^t K^{1/2} Z I_{\text{fic}}(Z) - \omega^t K^{1/2} a\}^2$  with an analogous definition for  $\bar{R}_{\text{aic}}(a)$ . Here  $I_{\text{fic}}(z) = I\{(\omega^t K^{1/2} z)^2 \geq 2\omega^t K \omega\}$  and  $I_{\text{aic}}(z) = I\{z^t z \geq 2q\}$ . Investigations reported on in Claeskens and Hjort (2003) show that the FIC method often does better than the AIC. Also, it typically pays off to smooth the FIC weights as in (5.4).

(v) *Average-across-singletons estimator.* When  $K$  is diagonal, the method developed in Section 5.5 has

$$\bar{R}(a) = E\left[\sum_{j=1}^q \omega_j k_j^{1/2} \left\{ \frac{\exp(\frac{1}{2} \hat{\rho} Z_j^2)}{\sum_{i=1}^q \exp(\frac{1}{2} \hat{\rho} Z_i^2)} \hat{\rho} Z_j - a_j \right\}\right]^2,$$

where  $\hat{\rho} = \hat{\tau}^2/(1 + \hat{\tau}^2)$  and  $\hat{\tau} = (\|Z\|^2 - q)_+^{1/2}$ .

(vi) *Other empirical Bayes schemes.* The main method of Section 5.5 may be analysed via the appropriate

$$W_j(z) = \hat{\rho} \frac{p_1(1 + \hat{\sigma}^2)^{-1/2} \exp(\frac{1}{2} \hat{\rho} z_j^2)}{p_0 + p_1(1 + \hat{\sigma}^2)^{-1/2} \exp(\frac{1}{2} \hat{\rho} z_j^2)}, \quad \text{with } \hat{\rho} = \frac{\hat{\sigma}^2}{1 + \hat{\sigma}^2},$$

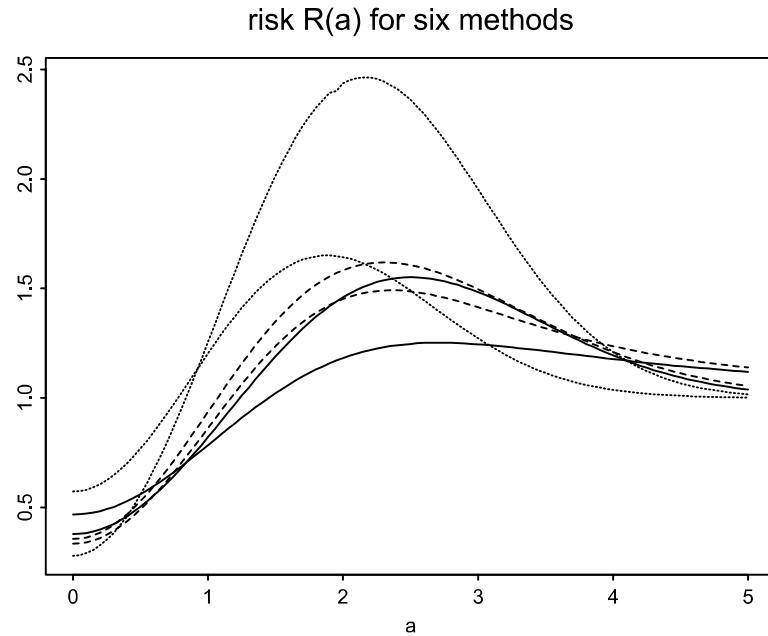


FIGURE 7.1. Risk functions  $R(a)$  associated with six FMA methods, for  $q = 1$ , as in (7.2); these are symmetric around zero and are displayed here for  $a \in [0, 5]$ . The AIC one (dotted line) starts at 0.572 with max-risk 1.650. The pre-test approach which uses 0.05 as test level (dotted line) starts at 0.279 with high max-risk 2.464. The smoothed AIC (solid line) starts at 0.378 with max-risk 1.551. The empirical Bayes singletons method (dashed line) starts at 0.333 with max-risk 1.491. The method corresponding to (5.5) (solid line) starts at 0.467 with low max-risk 1.252. Finally the method of Section 5.6 (dashed line), with  $p_0 = 0.25$ , starts at 0.335 with max-risk 1.619.

along with simulation-based computation of  $\bar{R}(a)$ , as per (7.1). Alternatives may be analysed similarly.

We have studied risk functions for various procedures, for the one- and two-dimensional cases, but cannot report in any depth here due to limitations of space. Some brief remarks are as follows. (a) It pays to smooth the hard-core AIC and FIC methods, as in Sections 5.2–5.3, and sometimes with a  $\kappa$  bigger than 1 in (5.4). These risk functions are smaller than the constant minimax risk  $\omega^t K \omega$  in a decent neighbourhood around zero, then increase, and level off towards the minimax value as  $\|a\|$  grows. (b) The singletons method of (5.6) does quite well in a reasonable neighbourhood around zero and along axes, where one  $|\delta_j|$  is big but the others small, but its risk may become large when more than one  $|\delta_j|$  becomes big. (c) The empirical Bayes scheme of Section 5.6, along with similarly inspired versions, does quite well in terms of low max-risk and being smaller than  $\omega^t K \omega$  in a broad neighbourhood around zero.

**7.2. Risk comparison in a simulated Poisson setting.** Here we illustrate the mean squared error (mse) behaviour of model averaged and post-model selection estimators for Poisson regression. The situation we study has the narrow model containing an intercept

only, whereas in the widest model four variables are included. In other words, counts  $Y_i$  are independent and Poisson with parameters  $\xi_i$ , where  $\xi(u_i) = \exp(\beta_0 + \sum_{j=1}^4 \gamma_j u_{i,j})$ . There are at the outset  $2^q = 16$  different submodels to consider, corresponding to inclusion or not of the four  $\gamma_j$ s. In the simulation study we took  $\beta_0 = 1$  and  $\delta = (1, 1, 1, 1)^t$ , i.e.  $\gamma_j = 1/\sqrt{n}$ , and chose two focus points in the covariate space,  $u = (1, -0.9, -0.9, 1)^t$  and  $u = (1, 1, -0.6, -0.6)^t$ . The four covariates  $u_{i,1}, \dots, u_{i,4}$  were taken to be independent and standard normal. Estimators were then simulated 1,000 times for each setting using the empirical  $\widehat{J}$  matrix  $n^{-1} \sum_{i=1}^n \widehat{\xi}(u_i) u_i u_i^t$ , with its submatrices and corresponding  $K$  matrix, and the  $\omega$  vector calculated for each choice of  $u$  as  $\widehat{\omega} = \widehat{\xi}(u)(\widehat{J}_{10}/\widehat{J}_{00} - u)$ . While the main point here is to compare methods for finite sample sizes, we also include in the table the population quantities  $\tau_0^2 + R(\delta)$ , which by Theorem 4.1 are the limits of  $n$  times mse. To compute these we use the fact that  $J_{\text{full}} = \exp(\beta_0)I_5$ , from properties of the normal covariate distribution, in terms of the  $5 \times 5$  identity matrix, and which entails  $K = \exp(-\beta_0)I_4$ . We then evaluated  $R(\delta)$  by taking the average of a full million simulated versions of  $\{\omega^t \widehat{\delta}(D) - \omega^t \delta\}^2$ , with  $\widehat{\delta}(D)$  as in (4.2) and  $D \sim N_4(\delta, K)$ .

Table 7.1 shows simulated  $n$  mse values for the following estimators: post model selection using AIC and FIC; model averaged estimators using smoothed AIC and FIC weights as in (5.2) and (5.4), the latter with  $\kappa = 1$ ; the wide model estimator; and a testing approach where each variable is tested individually at a 5% level and only the significant variables are kept in the final model. Two sample sizes are used,  $n = 50$  and  $n = 200$ . From the table it is observed that the model averaged estimators have much lower mse values than the corresponding post-model selection estimators which select one single model. For these settings, the FIC yields significantly smaller mse values than the AIC. Model averaging also performs better in terms of mse than the wide model method, and outperforms the simple testing approach. For sample size 200, the simulated values are already close to the simulated theoretical mse values, confirming the theoretical derivations.

	setting (a)			setting (b)		
	$n = 50$	$n = 200$	limit	$n = 50$	$n = 200$	limit
post-AIC	20.26	16.65	17.85	12.13	17.74	14.22
smooth AIC	14.67	13.24	13.73	9.89	13.49	11.27
post-FIC	11.94	10.48	10.86	9.88	11.27	9.87
smooth FIC	8.53	8.19	8.20	7.41	8.22	7.62
wide model	15.47	12.11	12.56	10.01	12.03	10.11
testing	16.57	14.97	20.17	11.85	16.06	16.55

TABLE 7.1. Poisson regression with  $q = 4$  extra variables. Simulated  $n$  mse values for setting (a):  $u = (1, -0.9, -0.9, 1)^t$  and (b):  $u = (1, 1, -0.6, -0.6)^t$ .

Table 7.2 shows for the same settings simulated coverage probabilities for  $\xi(u)$  at a nominal 95% level, for sample sizes  $n$  equal to 50, 100 and 200, based on 10,000 simulation replicates of all estimators. The table includes first results of application of definitions (4.8) with  $D_n = \sqrt{n}\widehat{\gamma}_{\text{full}}$  for the first four methods shown: post-model selection using AIC

and FIC (with  $\kappa = 1$ ), and their smoothed model averaged versions. The approach by Buckland et al. (1997) uses the same smoothed AIC weights, but then uses the standard error estimate described in Section 4.3 above in conjunction with a simple non-biased normal approximation. The table furthermore displays results for the confidence interval coming from using the widest model estimator and the testing strategy explained above, both employing standard normal percentiles. It also shows the results of what happens to the coverage probability when ignoring the model selection step after AIC or BIC model selection.

With increasing sample size the corrected versions approach the nominal level of 95%, although the smoothed FIC values are a little larger than the nominal value, at least for this setting. By construction, the wide method is the safest method and will produce asymptotically correct confidence intervals. As a consequence of using an imperfect distributional approximation, the method by Buckland et al. does not reach nominal coverage in the performed simulations. The testing approach using normal percentiles produces confidence intervals with significantly lower than nominal coverage values. And as expected from theoretical considerations, see Section 4.3 above, ignoring model selection results in confidence intervals with too low coverage probabilities, as is illustrated by the last two lines in the table.

	setting (a)			setting (b)		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
post-AIC	0.935	0.946	0.948	0.941	0.948	0.946
smooth AIC	0.935	0.946	0.948	0.942	0.946	0.946
post-FIC	0.933	0.946	0.955	0.939	0.946	0.949
smooth FIC	0.957	0.967	0.974	0.957	0.968	0.971
Buckland et al.	0.925	0.926	0.929	0.928	0.897	0.916
wide model	0.936	0.947	0.948	0.944	0.947	0.946
testing	0.815	0.833	0.828	0.846	0.768	0.786
naive AIC	0.773	0.827	0.820	0.833	0.732	0.783
naive BIC	0.720	0.700	0.690	0.748	0.568	0.638

TABLE 7.2. Poisson regression with  $q = 4$  extra variables. Simulated coverage probabilities for  $\xi(u)$  for setting (a):  $u = (1, -0.9, -0.9, 1)^t$  and (b):  $u = (1, 1, -0.6, -0.6)^t$ .

## 8. Generalised ridging: Shrinking in parametric models

The development of Section 4 gave an instructive bridge from compromise estimators  $\hat{\mu}$  of type (4.1) to estimators of  $\psi = \omega^t \delta$  of type (4.6). For some purposes the class of (4.1) estimators is not quite large enough, however. For example, Theorem 4.1 does not cover the full class of natural  $\hat{a}_j(z) = \omega_j k_j^{1/2} W_j(Z) Z_j$  type estimators encountered as a consequence of exploiting this theorem; see Remark 4.3. This section expands the horizon by proposing and investigating certain generalised ridge estimators, which shrink the  $\hat{\gamma}_S$  estimators toward  $\gamma_0$ . Such shrinking may be particularly beneficial when the number  $q$  of

extra parameters is moderate or growing compared to a fixed number  $p$  of core parameters  $\theta$ . A quite general class of BMA estimators will in fact behave just in this way, as seen in Section 9.

The intention is to stick to the narrow model as a form of basis, but to consider down-weighting aspects of the more risky  $\gamma$ -extensions, via estimators of the form  $\tilde{\gamma}_S$  that shrink the  $\hat{\gamma}_S$  towards  $\gamma_0$ , with an amount somehow dictated by  $D_n$ . The idea is to use  $(\hat{\theta}_S, \tilde{\gamma}_S, \gamma_{0,S^c})$  as estimators in the  $S$  subset model, and more specifically

$$\tilde{\mu}_S = \mu(\hat{\theta}_S, \tilde{\gamma}_S, \gamma_{0,S^c}), \quad \text{where } \tilde{\gamma}_S - \gamma_{0,S} = \{1 - \varepsilon_S(D_n)\}(\hat{\gamma}_S - \gamma_{0,S^c}),$$

for suitable functions  $\varepsilon_S(d)$ . The cases encountered earlier correspond to these functions being identically zero. For these estimators,

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\tilde{\gamma}_S - \gamma_{0,S}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} C_S \\ \{1 - \varepsilon_S(D)\}D_S \end{pmatrix},$$

which leads to

$$\sqrt{n}(\tilde{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = (\frac{\partial \mu}{\partial \theta})^t C_S + (\frac{\partial \mu}{\partial \gamma_S})^t \{1 - \varepsilon_S(D)\} D_S - (\frac{\partial \mu}{\partial \gamma})^t \delta.$$

With some algebraic work we find

$$\begin{aligned} \Lambda_S &= (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} M + \omega^t \{(I - K^{-1/2} H_S K^{1/2})^t \delta - K^{1/2} H_S K^{-1/2} W\} \\ &\quad - \varepsilon_S(D) (\frac{\partial \mu}{\partial \gamma})^t K^{1/2} H_S K^{-1/2} (\delta + W). \end{aligned}$$

To work with estimator-after-selection estimators we are again led to consider the class of estimators

$$\tilde{\mu} = \sum_S c(S | D_n) \tilde{\mu}_S = \sum_S c(S | D_n) \mu(\hat{\theta}_S, \tilde{\mu}_S, \gamma_{0,S^c}), \quad (8.1)$$

with coefficients summing to 1 and allowed to depend on  $D_n$  of (3.1). There is a limiting distribution for  $\sqrt{n}(\tilde{\mu} - \mu_{\text{true}})$  admitting the representation

$$\Lambda = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} M + \omega^t \{\delta - G(D)^t (\delta + W)\} - (\frac{\partial \mu}{\partial \gamma})^t G^*(D)^t (\delta + W),$$

where  $G(D) = \sum_S c(S | D) K^{-1/2} H_S K^{1/2}$  and  $G^*(D) = \sum_S c(S | D) \varepsilon_S(D) K^{-1/2} H_S K^{1/2}$ . A special case of interest is when each  $\varepsilon_S(D)$  the same, corresponding to equal ridging in all  $\gamma_j$  directions. Then  $G^*(D) = \varepsilon(D) G(D)$ .

These efforts lead to a limiting risk expression for estimators of the form (8.1), namely

$$E\Lambda^2 = \tau_0^2 + E[\omega^t \{\delta - G(D)^t D\} - (\frac{\partial \mu}{\partial \gamma})^t G^*(D)^t D]^2 = \tau_0^2 + E(\tilde{\psi} - \omega^t \delta)^2,$$

say, where the estimator of  $\psi = \omega^t \delta$  this time becomes

$$\{\omega^t G(D)^t + (\frac{\partial \mu}{\partial \gamma})^t G^*(D)^t\} D = \left[ (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} J_{01} G(D)^t - (\frac{\partial \mu}{\partial \gamma})^t \{G(D) - G^*(D)\}^t \right] D.$$

This would take the form  $\tilde{\psi} = [(\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} J_{01} - \{1 - \varepsilon(D)\} (\frac{\partial \mu}{\partial \gamma})^t] G(D)^t D$  in the case of the same  $\varepsilon(D)$  for all subsets under consideration. We see that this bridges from the situation of Theorem 4.1, for  $\varepsilon(D) = 0$ , to the result for the narrow procedure, for  $\varepsilon(D) = 1$ . It also shows that using  $\varepsilon(D)$  may shrink down the size of the  $\frac{\partial \mu}{\partial \gamma}$  component here, in its turn often lowering the variance level.

## 9. A frequentist view of BMA

Bayesian model averaging essentially amounts to putting down prior probabilities  $p(S)$  for all submodels, prior distributions  $\pi_S(\theta, \delta_S)$  for the parameters inside the  $S$  submodel, and then applying Bayes' theorem suitably. The basics of such machineries is covered in Draper (1995) and Hoeting et al. (1999). In our framework, the posterior of the parameters may be expressed as

$$\pi_n(\theta, \delta) = \sum_S p_n(S) \pi_{n,S}(\theta, \delta_S). \quad (9.1)$$

Here  $\pi_{n,S}(\theta, \delta_S)$  is the posterior calculated under the  $S$  model (in particular, then  $\delta_j = 0$  for  $j \notin S$ ) while  $p_n(S) = p(S) \lambda_n(S) / \sum_{S'} p(S') \lambda_n(S')$  is the probability of model  $S$  given data. Here

$$\lambda_n(S) = \int L_{n,S}(\theta, \gamma_0 + \delta_S/\sqrt{n}) \pi_S(\theta, \delta_S) d\theta d\delta_S \quad (9.2)$$

is the integrated likelihood of model  $S$ , involving the likelihood  $L_{n,S}$  for this model. The  $\lambda_n(S)$  is also the marginal distribution at the observed data. Below we derive approximations and precise limit distribution results for the quantities involved in (9.1) and (9.2), under our local alternatives framework. Limit behaviour of BMA schemes has apparently not been studied before.

*9.1. The posterior model probabilities.* We need to understand the behaviour of  $\lambda_n(S)$ . The familiar BIC statistic stems in fact from an approximation to this quantity. To review and comment on this approximation, let as before  $\hat{\theta}_S$  and  $\hat{\delta}_S = \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S})$  be the maximum likelihood estimators inside the  $S$  model. Then

$$\lambda_n(S) \doteq L_{n,S}(\hat{\theta}_S, \hat{\gamma}_S) n^{-(p+|S|)/2} (2\pi)^{(p+|S|)/2} |J_{n,S}|^{-1/2} \pi_S(\hat{\theta}_S, \hat{\delta}_S) \quad (9.3)$$

is one possible approximation. Here  $J_{n,S} = -n^{-1} \sum_{i=1}^n \partial^2 \log f(Y_i, \hat{\theta}_S, \hat{\gamma}_S) / \partial \alpha_S \partial \alpha_S^t$  is the observed information matrix of size  $(p+|S|) \times (p+|S|)$ , using  $\alpha_S$  to denote the parameter vector with  $\theta$  and  $\gamma_S$ . The consequent  $2 \log \lambda_n(S) \approx 2 \max \log L_{n,S} - (p+|S|) \log n$  is often called ‘the BIC approximation’; see e.g. Hoeting et al. (1999, equation (13), modulo an incorrect constant). Claim (9.3) may be proved using arguments similar to those needed to show Proposition 9.1 below.

It is important to note, though, that the asymptotic approximation (9.3), which underlies the BIC, is valid in the framework of fixed models  $f(y, \theta, \gamma)$  and a fixed  $f_{\text{true}}(y)$ , and where in particular also  $\delta = \sqrt{n}(\gamma - \gamma_0)$  grows with  $n$ . In such a framework the best

model will win in the end, that is, the candidate model  $S_0$  with smallest Kullback–Leibler distance to the true density will have  $p_n(S_0) \rightarrow 1$  as  $n$  grows. This follows since the dominant term of  $\max \log L_{n,S}$  will be  $n$  times  $\max \int f_{\text{true}}(y) \log f(y, \theta, \gamma) dy$ . In our framework of local alternative models the magnifying glass is focussed on the  $\sqrt{n}(\gamma - \gamma_0)$  scale, and different results apply. Maximised log-likelihoods are then not  $O_p(n)$  apart, as under the fixed models scenario, but have differences related to noncentral chi squared distributions. Secondly, the  $n^{-|S|/2}$  ingredient above, crucial to the BIC, disappears.

For the following result, which provides a more accurate approximation than the BIC-related (9.3) when the scale of model departures from the narrow model is that of  $\delta = \sqrt{n}(\gamma - \gamma_0)$ .

**PROPOSITION 9.1.** *Let the prior for the  $S$  subset model take the form  $\pi_0(\theta)\pi_S(\delta_S)$ , with  $\pi_0$  continuous in a neighbourhood around  $\theta_0$ . Then, under standard regularity conditions, when  $n$  grows,*

$$\lambda_n(S) \doteq L_{n,S}(\widehat{\theta}_S, \widehat{\gamma}_S) n^{-p/2} (2\pi)^{(p+|S|)/2} \pi_0(\widehat{\theta}_S) |J_{n,S}|^{-1/2} \kappa_n(S),$$

where  $\kappa_n(S) = \int \phi(\delta_S - \widehat{\delta}_S, J_{n,S}^{11}) \pi_S(\delta_S) d\delta_S$ . The approximation holds in the sense that  $\log \lambda_n(S)$  is equal to the logarithm of the right hand side plus a remainder term of size  $O_p(n^{-1/2})$ . Also,  $J_{n,S}^{11}$  is the lower right-hand  $|S| \times |S|$  submatrix of  $J_{n,S}^{-1}$ .

When  $n$  grows we also have  $J_{n,S} \rightarrow_p J_S$ , defined in Section 3.1, and the limit of  $J_{n,S}^{11}$  is  $K_S = (\pi_S K^{-1} \pi_S^t)^{-1}$ . Combining this with some previous results, reached in conjunction with (3.4), we find

$$\lambda_n(S) \doteq \text{const. } \exp\left(\frac{1}{2}\widehat{\delta}_S^t K_S^{-1} \widehat{\delta}_S\right) (2\pi)^{|S|/2} |J_S|^{-1/2} \int \phi(\delta_S - \widehat{\delta}_S, K_S) \pi_S(\delta_S) d\delta_S,$$

where the constant in question is  $n^{-p/2} (2\pi)^{p/2} \pi_0(\widehat{\theta})$ . This also leads to a precise description of posterior probabilities for the different models in the canonical limit experiment. This is the situation of large  $n$  where all quantities have been estimated with full precision except  $\delta$ , for which we must be content with the limit  $D \sim N(\delta, K)$  of  $D_n = \sqrt{n}(\widehat{\gamma}_{\text{full}} - \gamma_0)$ . Here  $p(S | D) \propto p(S)\lambda(S)$ , where

$$\begin{aligned} \lambda(S) &= \exp\left(\frac{1}{2}D_S^t K_S^{-1} D_S\right) (2\pi)^{|S|/2} |J_S|^{-1/2} \int \phi(\delta_S - D_S, K_S) \pi_S(\delta_S) d\delta_S \\ &= \exp\left(\frac{1}{2}\text{AIC}_S\right) \exp(|S|) (2\pi)^{|S|/2} |J_S|^{-1/2} \int \phi(\delta_S - D_S, K_S) \pi_S(\delta_S) d\delta_S \end{aligned}$$

and  $D_S = K_S \pi_S K^{-1} D$ . We use here  $\text{AIC}_S = D_S^t K_S^{-1} D_S - 2|S|$  from Section 3.3.

**9.2. Bayesian model choice with the canonical normal priors.** The primary special case is when  $\delta_S$  has the prior  $N(0, \tau_S^2 K_S)$ . This corresponds to independent and equally spread-out priors around zero for the transformed parameters  $a_S = \pi_S a$ , where  $a = K^{-1/2} \delta$  on the canonical scale, and where again  $Z \sim N_q(a, I)$ . Then

$$\lambda(S) = \exp\left(\frac{1}{2} \frac{\tau_S^2}{1 + \tau_S^2} D_S^t K_S^{-1} D_S\right) (1 + \tau_S^2)^{-|S|/2} |J_{00}|^{-1/2}. \quad (9.4)$$

The last determinant is independent of  $|S|$ , and emerges via  $|J_S|^{-1/2}|K_S|^{-1/2}$ , since  $|J_S| = |J_{00}||K_S|^{-1}$ . Result (9.4) is also valid for  $S = \emptyset$ , corresponding to the narrow model, for which  $\lambda(\emptyset) = |J_{00}|^{-1/2}$ .

This also gives rise to a new Bayesian information criterion, which we may term the BLIC, with L for ‘local’, reminding us of the local model extension framework (2.2). This criterion is reached by following the original BIC path, but using a different statistical magnifying glass, focussing on  $\gamma_0 + \delta/\sqrt{n}$  type neighbouring models. From (9.4), our criterion reads

$$\text{BLIC} = \frac{\tau_S^2}{1 + \tau_S^2} D_S^t K_S^{-1} D_S - |S| \log(1 + \tau_S^2) + 2 \log p(S),$$

since the posterior model probability is close to being proportional to  $p(S)\lambda(S)$ . Here  $\tau_S$  is meant to be a spread measure for  $\delta_S$  in submodel  $S$ , and for the narrow model  $\text{BLIC} = 2 \log p(\emptyset)$ . The candidate model with largest BLIC is the most probable one, given data, in the Bayesian formulation, and is selected.

The formula above is valid for the limit experiment. For real data we use  $\widehat{\delta}_S$  for  $D_S$ , leading to

$$\widehat{\text{BLIC}} = \frac{\tau_S^2}{1 + \tau_S^2} n(\widehat{\gamma}_S - \gamma_{0,S})^t \widehat{K}_S^{-1} (\widehat{\gamma}_S - \gamma_{0,S}) - |S| \log(1 + \tau_S^2) + 2 \log p(S).$$

Furthermore, we may estimate the spread. First,  $D_S^t K_S^{-1} D_S$  given  $\delta$  is a noncentral chi squared with parameter  $\delta_S^t K_S^{-1} \delta_S$ . Taking the mean of  $|S| + \delta_S^t K_S^{-1} \delta_S$  again gives  $|S|(1 + \tau_S^2)$ . We may thus suggest  $1 + \tau_S^2$  estimated, in this empirical Bayes fashion, by  $D_S^t K_S^{-1} D_S / |S|$ . This gives say

$$\text{BLIC}^* = |S|\{\widehat{\tau}_S^2 - \log(1 + \widehat{\tau}_S^2)\} + 2 \log p(S), \quad \text{with } \widehat{\tau}_S^2 = \max\{D_S^t K_S^{-1} D_S / |S| - 1, 0\}.$$

Various alternatives may also be considered.

*9.3. Posteriors in submodels.* We need to investigate the behaviour of the posterior distributions  $\pi_{n,S}(\theta, \delta_S)$ , conditional on model  $S$ , and in particular their means  $\tilde{\mu}_S = E_S(\mu | \text{data})$ . It will become clear that for large  $n$ , the distribution of  $\theta$  will be tightly concentrated around  $\widehat{\theta}_S$ , while the part of the prior related to  $\delta_S$  will not be ‘washed away’ by the data. This is since the chimeric parameter  $\delta$  will not be consistently estimated as data accumulate; the best we may do is via  $\widehat{\delta}_{\text{full}} \rightarrow_d N_q(\delta, K)$ . The posterior for  $\delta_S$  will in fact go to

$$\begin{aligned} \pi_S(\delta_S | \widehat{\delta}_S) &= \text{const. } \pi_S(\delta_S) \exp\left\{-\frac{1}{2}(\delta_S - \widehat{\delta}_S)^t K_S^{-1} (\delta_S - \widehat{\delta}_S)\right\} \\ &= \text{const. } \pi_S(\delta_S) \phi(\delta_S - \widehat{\delta}_S, K_S), \end{aligned} \tag{9.5}$$

as shall be seen below. Thus  $E(\delta_S | \text{data})$  is for large  $n$  essentially a function of  $\widehat{\delta}_S$ , which again is a function of  $Z_n$  of (3.3), per Section 3.3. In the limit experiment, where  $\widehat{\delta}_S \rightarrow_d$

$D_S = K_S \pi_S K^{-1}(\delta + W)$  with mean  $K_S \pi_S K^{-1}\delta$  and variance matrix  $K_S$ , write

$$E_L(\delta_S | D_S) = \frac{\int \delta_S \pi_S(\delta_S) \exp\{-\frac{1}{2}(\delta_S - D_S)^t K_S^{-1}(\delta_S - D_S)\} d\delta_S}{\int \pi_S(\delta_S) \exp\{-\frac{1}{2}(\delta_S - D_S)^t K_S^{-1}(\delta_S - D_S)\} d\delta_S}.$$

We then have the following extension of Lemma 3.3.

**PROPOSITION 9.2.** *Under the conditions of the previous proposition, the Bayesian submodel estimator  $\tilde{\mu}_S = E_S(\mu | \text{data})$  is asymptotically equivalent to the simpler estimator  $\bar{\mu}_S = E\{\mu(\hat{\theta}_S, \gamma_{0,S} + \delta_S/\sqrt{n}) | \hat{\delta}_S\}$ , where the distribution in question is that of (9.5). Also,*

$$\sqrt{n}(\tilde{\mu}_S - \mu_{\text{true}}) \rightarrow_d \tilde{\Lambda}_S = (\frac{\partial \mu}{\partial \theta})^t C_S + (\frac{\partial \mu}{\partial \gamma_S})^t E_L(\delta_S | D_S) - (\frac{\partial \mu}{\partial \gamma})^t \delta.$$

**9.4. BMA approximations.** The approximations to  $\tilde{\mu}_S$  indirectly touched on here are of separate value. The simplest of these, from the second half of the proof, is  $\hat{\mu}_S - (\frac{\partial \mu}{\partial \gamma_S})^t \{\hat{\delta}_S - E(\delta_S | \hat{\delta}_S)\}$ . It is also useful to record an approximation to the conditional variance  $\tilde{\sigma}_S^2 = \text{Var}_S(\mu | \text{data})$ . One first may show that  $\tilde{\sigma}_S^2 = E\{\mu(\hat{\theta}_S, \gamma_{0,S} + \delta_S/\sqrt{n})^2 | \hat{\delta}_S\} - \tilde{\mu}^2 + o(n^{-1})$ , and is then via (9.5) and renewed Taylor expansion led to  $\tilde{\sigma}_S^2 \doteq n^{-1}(\frac{\partial \mu}{\partial \gamma_S})^t \text{Var}(\delta_S | \hat{\delta}_S) \frac{\partial \mu}{\partial \gamma_S}$ .

The primary special case here is again the normal priors for  $\delta_S$  studied in Section 8.2. Then the posterior is normal with mean  $\rho_S \hat{\delta}_S$  and variance  $\rho_S K_S$ , where  $\rho_S = \tau_S^2 / (1 + \tau_S^2)$ . Thus

$$\tilde{\mu}_S \doteq \hat{\mu}_S - (\frac{\partial \mu}{\partial \gamma_S})^t \hat{\delta}_S (1 - \rho_S) \quad \text{and} \quad \tilde{\sigma}_S^2 \doteq n^{-1}(\frac{\partial \mu}{\partial \gamma_S})^t K_S \frac{\partial \mu}{\partial \gamma_S} \rho_S.$$

Also, from the proposition,  $\tilde{\Lambda}_S = (\frac{\partial \mu}{\partial \theta})^t C_S + \rho_S (\frac{\partial \mu}{\partial \gamma_S})^t D_S - (\frac{\partial \mu}{\partial \gamma})^t \delta$ . But this is exactly as in Section 6, with shrinking factor  $\varepsilon_S(D) = 1 - \rho_S = 1/(1 + \tau_S^2)$  independent of  $D$ . When  $\tau_S$  is small, the prior is informative and tight, the shrinkage high, and the Bayes estimator is in the  $\tau_S \rightarrow 0$  limit the same as the narrow estimator  $\hat{\mu}_{\text{narr}}$ . If on the other hand  $\tau_S$  becomes big, then the prior is diffuse and the shrinkage small; in the limit case  $\tau_S \rightarrow \infty$ , the Bayes estimator is the same as the maximum likelihood estimator  $\hat{\mu}_S$ .

For BMA estimators the limiting risk function to study is  $R(\delta) = E(\tilde{\psi} - \omega^t \delta)^2$ , where

$$\tilde{\psi} = \omega^t G(D)^t D + (\frac{\partial \mu}{\partial \gamma})^t G^*(D)^t D,$$

$G(D) = \sum_S c(S | D) K^{-1/2} H_S K^{1/2}$  and  $G^*(D) = \sum_S c(S | D) (1 + \tau_S^2)^{-1} K^{-1/2} H_S K^{1/2}$ . Furthermore,  $c(S | D)$  is proportional to  $p(S) \lambda(S)$ , with  $\lambda(S)$  as in (9.4).

A result analogous to (9.1) holds for the posterior distribution of  $\mu = \mu(\theta, \gamma_0 + \delta/\sqrt{n})$ , which we write as  $\pi_n(\mu) = \sum_S p_n(S) \pi_{n,S}(\mu)$ . The Bayes estimator (under quadratic loss) becomes  $\tilde{\mu} = E(\mu | \text{data}) = \sum_S p_n(S) \tilde{\mu}_S$  while  $\text{Var}(\mu | \text{data})$ , the natural Bayesian measure of spread, becomes  $\sum_S p_n(S) \{\tilde{\sigma}_S^2 + (\tilde{\mu}_S - \tilde{\mu})^2\}$ . These formulae allows one to carry out approximate BMA analysis with simple computations, without e.g. MCMC computations.

## 10. Concluding remarks

*10.1. Amendments when the largest model does not hold.* Our machinery has been developed under the key assumption (2.2), which says that the true data generating mechanism should be inside the largest of the parametric models considered. Such an assumption may be checked via goodness-of-fit methods, but can never be established with full certainty. Here we investigate briefly what happens when assumption (2.2) is not required to hold, relying on extended theory developed in Claeskens and Hjort (2003, Section 8).

Assume that the true density for data takes the form

$$f_{\text{true}}(y) = f(y, \theta_0, \gamma_0) \{1 + r(y)/\sqrt{n}\} + o(1/\sqrt{n}) \quad (10.1)$$

for a suitable  $r(y)$  function, with  $\int f_0|r| dy$  finite and  $\int f_0 r dy = 0$ , where  $f_0(y) = f(y, \theta_0, \gamma_0)$ . Condition (2.2) corresponds to the special case  $r(y) = V(y)^t \delta$ , with  $V(y)$  as in Section 3.1. Since there are no ‘true parameters’ now, consider instead the least false parameter, say  $\mu_{\text{lf}} = \mu(\theta_n, \gamma_n)$ . Here  $(\theta_n, \gamma_n)$  are the least false parameters inside the  $f(y, \theta, \gamma)$  family, i.e. those minimising which is the Kullback–Leibler distance  $\int f_{\text{true}}(y) \log\{f_{\text{true}}(y)/f(y, \theta, \gamma)\} dy$ . It is shown in Claeskens and Hjort (2003, Section 8) that  $\theta_n = \theta_0 + \eta_0/\sqrt{n}$  and  $\gamma_n = \gamma_0 + \delta_0/\sqrt{n}$ , apart from terms of lower order, for constants  $\eta_0, \delta_0$  depending on  $\int f_0 Ur dy$  and  $\int f_0 Vr dy$ , as explained there. It is also shown that

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{lf}}) \rightarrow_d \tilde{\Lambda}_S = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} M + \omega^t \{\delta_0 - K^{1/2} H_S K^{-1/2}(\delta_0 + W)\}. \quad (10.2)$$

This is actually close to the result derived in Lemma 3.3, but now under wider start assumptions. The first point to note is that  $\eta_0$  has dropped out, the second is that the agnostic parameter  $\delta_0$  takes the place of our earlier  $\delta$ . Also,  $D_n = \hat{\delta}_{\text{full}} \rightarrow_d D = \delta_0 + W \sim N_q(\delta_0, K)$ , in generalisation of (3.1). This means that the theory of Sections 4–6, about compromise, post-selection and shrinkage estimators essentially goes through, with small amendments, and the methods developed are still in force. The difference is mostly related to interpretation, not to algorithms, so to speak; precision of estimators is interpreted and assessed in terms of closeness of the agnostic  $\mu_{\text{lf}}$ , rather than to the ‘true’ focus parameter.

*10.2. Breadth of applications.* It should be clear from our unified framework and application examples that there is a wide range of potential applications of our methods. Subset selection and model averaging can in particular be implemented and studied for quite general regression models, like generalised linear models. Versions of our methods and results would also hold for models with dependence and for various stochastic process models. The essential requirement is that ordinary likelihood analysis should be valid, with limiting normality of the maximum likelihood estimators and so on. Our study indicates that it would be useful to carry out more extensive risk comparisons in the limit experiment, as touched on in Section 7.1, in that conclusions reached there will have implications for a fair range of situations.

**10.3. Tolerance radii.** Sometimes ignorance is strength, and it may be better to stick to a simple model than going for a more complex one. This is captured well by our results of Sections 3 and 4. These may be used to characterise situations where a given  $S$  subset model gives better results than a competing  $S'$ . We find in particular that inference using the narrow model is better than using the fullest model provided  $|\omega^t \delta| \leq (\omega^t K \omega)^{1/2}$ . For a given estimand this describes a band of infinite length for  $\delta$ . On the other hand, inside the ellipsoid where  $\delta^t K^{-1} \delta \leq 1$  narrow model inference is better than full model inference, for all estimands.

**10.4. Two uses of models.** Sometimes statistical modelling strives for coming close to a superior scientific explanation of the phenomenon being studied, e.g. in physics or biology. In this article we are employing models differently, as pragmatic approximations to reality, with the aim of generating estimates and predictions with good precision. See the engaging discussion of Breiman (2001) and Section 1 of Claeskens and Hjort (2003) for further comments.

**10.5. Optimal methods.** In addition to methods proposed in Section 5 one may try to develop FMA schemes with suitable optimality properties. The Bayes methods we have discussed are optimal w.r.t. the criterion of minimising prior-weighted risk. The full-model estimator is the unique minimax estimator, under the  $(\omega^t \delta - \omega^t \hat{\delta})^2$  loss function, with constant risk  $\omega^t K \omega$ . Other criteria might in one way or another involve ideas of restricting max-risk under the constraint of doing well at or near  $\delta = 0$ . Methods developed, for other purposes, in Bickel (1981, 1983, 1984) and Berger (1982) are of relevance here, but cannot be applied directly in that we restrict attention to FMA regimes.

**10.6. Bootstrapping does not work.** To explain why bootstrapping cannot be relied upon in our model choice framework, consider the following situation. It is simple but representative of our general local model choice context. There are independent observations  $Y_i \sim N(\mu, 1)$ , where the narrow model holds that  $\mu = 0$  and the wider model takes  $\mu$  unknown; thus  $\hat{\mu}_{\text{full}} = \bar{Y}_n$ . In the framework of local alternatives  $\mu = \delta/\sqrt{n}$ , where  $Z_n = \sqrt{n}\bar{Y}_n$  is the natural test statistic, consider a model average estimator  $\hat{\mu}$  which gives weight  $1 - W(Z_n)$  to  $\hat{\mu}_{\text{narr}}$  and weight  $W(Z_n)$  to  $\hat{\mu}_{\text{full}}$ , that is,  $\hat{\mu} = W(\sqrt{n}\bar{Y}_n)\bar{Y}_n$ . First study

$$\Lambda_n = \sqrt{n}(\hat{\mu} - \mu_{\text{true}}) = W(\sqrt{n}\bar{Y}_n)\sqrt{n}\bar{Y}_n - \delta =_d W(\delta + N)(\delta + N) - \delta,$$

where  $N$  represents a standard normal. Then study bootstrapped data  $Y_i^*$  from the estimated full model  $N(\hat{\mu}_{\text{full}}, 1)$ , with resulting bootstrap estimator  $\hat{\mu}^* = W(\sqrt{n}\bar{Y}_n^*)\bar{Y}_n^*$ . Here we find

$$\Lambda_n^* = \sqrt{n}(\hat{\mu}^* - \hat{\mu}) = W(\sqrt{n}\bar{Y}_n^*)\sqrt{n}\bar{Y}_n^* - \hat{\delta} =_d W(\hat{\delta} + N')(\hat{\delta} + N') - \hat{\delta},$$

where  $N'$  represents another standard normal, independent of  $N$  above. Thus the distributions of  $\Lambda_n$  and  $\Lambda_n^*$  are *not* close (excluding now the special case  $W = 1$ , which corresponds to using the wide estimator), since  $\hat{\delta} = \sqrt{n}\hat{\mu}$  does not go to  $\delta$  in probability.

**10.7. Finite-sample correction and non-ML estimators.** We have made extensive use of the first-order asymptotic theory for maximum likelihood estimation, yielding clear and concise descriptions of limit distributions etc. While this is already quite satisfactory it is clear that suitable finite-sample corrections could be developed, perhaps for particular classes of models, in order to improve approximations. Work by Hurvich and Tsai (1989), extensively discussed in Burnham and Anderson (2002), is of relevance here. Another direction for future research is that of using robust estimators, perhaps of the M-estimator variety, instead of maximum likelihood estimators. It may also be important to use more robust weighting schemes.

## 11. Proofs of lemmas and theorems

For the set-up of Section 3.1, it is assumed that the log-density has two continuous partial derivatives around  $(\theta_0, \gamma_0)$ , so that

$$\log \frac{f(y, \theta_0 + s, \gamma_0 + t)}{f(y, \theta_0, \gamma_0)} = \begin{pmatrix} U(y) \\ V(y) \end{pmatrix}^t \begin{pmatrix} s \\ t \end{pmatrix} + \frac{1}{2} \begin{pmatrix} s \\ t \end{pmatrix}^t W(y) \begin{pmatrix} s \\ t \end{pmatrix} + R(y, s, t) \quad (11.1)$$

for  $(s, t)$  small in  $\mathcal{R}^{p+q}$ , involving the matrix  $W(y)$  of second log-density derivatives at the null point and a remainder term  $R(y, s, t)$ . It is also required that the variance matrix  $J_{\text{full}}$  of  $(U(Y), V(Y))$  under  $f_0(y) = f(y, \theta_0, \gamma_0)$ , which is also the negative mean of  $W(Y)$  under  $f_0$ , is finite and of full rank. This also gives rise to the representation  $f(y, \theta_0, \gamma_0 + t) = f_0(y)\{1 + V(y)^t t + R_2(y, t)\}$ , where  $R_2(y, t)$  is typically small enough to make  $f_0(y)R_2(y, t)$  of order  $o(\|t\|^2)$  uniformly in  $y$ , and to

$$f_{\text{true}}(y) = f_0(y)\{1 + V(y)^t \delta/\sqrt{n} + R_2(y, \delta/\sqrt{n})\}. \quad (11.2)$$

Various sets of regularity conditions may now be put up to reach the desired conclusions, working either with (11.1) or (11.2) as convenient. Consider in fact the following assumptions.

- (C1): The two integrals  $\int f_0(y)U(y)R_2(y, t) dy$  and  $\int f_0(y)V(y)R_2(y, t) dy$  are both  $o(\|t\|)$ .
- (C2): The variables  $|U_i^2V_j|$  and  $|V_i^2V_j|$  have finite mean under  $f_0$ , for each  $i, j$ .
- (C3): The two integrals  $\int f_0(y)\|U(y)\|^2R_2(y, t) dy$  and  $\int f_0(y)\|V(y)\|^2R_2(y, t) dy$  are both  $o(1)$ .
- (C4): The log density has three continuous derivatives w.r.t. all  $p + q$  parameters in a neighbourhood around  $(\theta_0, \gamma_0)$ , and are there dominated by functions with finite means under  $f_0$ . Conditions (C1) and (C3) are quite weak, and are implied by the stronger condition (C4). Then the integrals of (C1) and (C3) are in fact  $o(\|t\|^2)$ . Condition (C4) will hold for most models, as will (C2).

**PROOF OF LEMMA 3.1,** under conditions (C1), (C2), (C3): This is accomplished via the multivariate Lindeberg theorem (which is the univariate Lindeberg theorem in conjunction with the Cramér–Wold device; see e.g. Serfling (1980, Section 1.9)). (C1) implies  $EU(Y_i) = J_{01}\delta/\sqrt{n} + o(1/\sqrt{n})$  and  $EV(Y_i) = J_{11}\delta/\sqrt{n} + o(1/\sqrt{n})$ , while (C2)

and (C3) see to it that the variance matrix of  $\sqrt{n}(\bar{U}_n, \bar{V}_n)$  goes to  $J_{\text{full}}$ . The Lindeberg requirement for asymptotic normality here demands integrals of  $U_i^2 I\{\|U\| \geq \sqrt{n}\varepsilon\}$  and  $V_i^2 I\{\|V\| \geq \sqrt{n}\varepsilon\}$  w.r.t. the (11.2) density to go to zero, and this is secured, for each positive  $\varepsilon$ , under (C2) and (C3). ■

**PROOF OF LEMMA 3.2,** under condition (C4): This follows by suitable extension of traditional arguments given for proving asymptotic normality of maximum likelihood estimators in fixed parametric models, see e.g. Lehmann (1983, Ch. 6). The essence is that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} \doteq J_S^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_{n,S} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} J^{00,S} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{pmatrix} \begin{pmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{pmatrix},$$

with  $N_S$  denoting the vector of  $N_j$ s with  $j \in S$ . ■

We mention that Lemma 3.2 often may be proved to hold under weaker conditions than (C4), in cases where the log density is concave in the parameters. This is important for the somewhat more difficult statements and proofs required when extending Lemmas 3.1–3.3 to regression models. Space does not allow us giving such in suitable detail here, but transparent proofs, under minimal conditions of the type  $n^{-1/2} \max_{i \leq n} \|x_i\| \rightarrow 0$ , may be given for log-concave models using convexity arguments of Hjort and Pollard (1994).

**PROOF OF LEMMA 3.3.** Using a delta method type Taylor expansion for  $\mu(\hat{\theta}_S, \hat{\gamma}_S) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ , in conjunction with Lemma 3.2, we easily establish that there is a limit distribution, which can be represented as  $\Lambda_S = (\frac{\partial \mu}{\partial \theta})^t C_S + (\frac{\partial \mu}{\partial \gamma_S})^t D_S - (\frac{\partial \mu}{\partial \gamma})^t \delta$ . (For the delta method, see e.g. Barndorff-Nielsen and Cox (1989, Ch. 2.) It is furthermore clear that  $\Lambda_S$  in this form is normal, and it is not difficult to work out valid expressions for mean and variance, and hence the limiting mean squared error. We will take the trouble to first derive certain simplified expressions for  $\Lambda_S$ , however, since these will be fruitful also for other purposes.

Using Lemma 3.2 in connection with expressions for the blocks of  $J^{-1}$  one finds after some algebraic manipulations that

$$\begin{aligned} C_S &= (J^{00,S} J_{01} + J^{01,S} \pi_S J_{11})\delta + J^{00,S} M + J^{01,S} N_S \\ &= J_{00}^{-1} J_{01} (I - K^{1/2} H_S K^{-1/2})\delta + J_{00}^{-1} M - J_{00}^{-1} J_{01} \pi_S^t K_S \pi_S (N - J_{10} J_{00}^{-1} M) \\ &= J_{00}^{-1} J_{01} (I - K^{1/2} H_S K^{-1/2})\delta + J_{00}^{-1} M - J_{00}^{-1} J_{01} K^{1/2} H_S K^{-1/2} W, \end{aligned}$$

while similarly

$$D_S = (J^{10,S} J_{01} + J^{11,S} \pi_S J_{11})\delta + J^{10,S} M + J^{11,S} N_S = K_S \pi_S K^{-1} (\delta + W).$$

This leads to a fruitful expression for  $\Lambda_S$  in terms of bias part and a zero-mean normal. Its mean is  $b_S^t \delta$ , where

$$b_S = (I - K^{-1} \pi_S^t K_S \pi_S) J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} + (K^{-1} \pi_S^t K_S \pi_S - I) \frac{\partial \mu}{\partial \gamma} = (I - K^{-1/2} H_S K^{1/2}) \omega.$$

Working similarly with the random part we find the expression for  $\Lambda_S$  given in Lemma 3.3. Using the independence between  $M$  and  $W$ , which was noted before stating this lemma, we easily obtain the variance formula stated in the lemma. ■

PROOF OF THEOREM 4.1. There is simultaneous convergence in distribution of all the  $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$  jointly with  $D_n$  to the corresponding collection of  $\Lambda_S$  and  $D$ . This follows via arguments used to prove Lemma 3.3, and the fact that all limit variables can be expressed in terms of  $(M^t, N^t)^t$ . Thus there is also joint convergence in distribution of all  $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$  with  $c(S|D_n)$  to corresponding  $\Lambda_S$  with  $c(S|D)$ , in that  $c(S|d)$  is almost continuous in  $d$ . Consequently,

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) = \sum_S c(S|D_n) \sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \rightarrow_d \sum_S c(S|D) \Lambda_S.$$

The second expression for  $\Lambda$  follows with some efforts, using that  $c(S|D)$  sum to 1 for fixed  $D$ , in conjunction with the representation featured in Lemma 3.3. As commented on earlier,  $M$  and  $W = D - \delta$  turn out to be stochastically independent. Hence  $\Lambda$  given  $D = d$  is a normal distribution, with  $\text{Var}(\Lambda|d) = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta} = \tau_0^2$ , the minimal possible limit distribution variance for estimators under consideration, and  $E(\Lambda|d) = \omega^t \{\delta - G(d)^t d\}$ . It follows that the limiting mean squared error of an arbitrary estimator in the class under study can be expressed as  $E\Lambda^2 = \tau_0^2 + E[\omega^t \{\delta - G(D)^t D\}]^2$ . ■

PROOF OF PROPOSITION 9.1. We choose to work with the case of the full model, i.e.  $S = \{1, \dots, q\}$ , where we also are content to write  $\hat{\theta}$  and  $\hat{\delta}$  for  $\hat{\theta}_{\text{full}}$  and  $\hat{\delta}_{\text{full}}$ , and so on. The general case can be handled quite similarly. Introduce

$$Q_n(s, t) = \frac{L_n(\hat{\theta} + s/\sqrt{n}, \gamma_0 + (\hat{\delta} + t)/\sqrt{n})}{L_n(\hat{\theta}, \gamma_0 + \hat{\delta}/\sqrt{n})} = \frac{L_n(\hat{\theta} + s/\sqrt{n}, \hat{\gamma} + t/\sqrt{n})}{L_n(\hat{\theta}, \hat{\gamma})}.$$

Then, with Taylor expansion analysis, one sees that

$$\log Q_n(s, t) = -\frac{1}{2} \binom{s}{t}^t J_n \binom{s}{t} + O_p(n^{-1/2} \left\| \binom{s}{t} \right\|^3).$$

For a calculation needed in a moment, note that for a symmetric positive definite  $(p+q) \times (p+q)$  matrix  $A$ ,

$$\int \exp\left\{-\frac{1}{2} \binom{s}{t}^t A \binom{s}{t}\right\} ds = (2\pi)^{p/2} |A|^{-1/2} |A^{11}|^{-1/2} \exp\left\{-\frac{1}{2} t^t (A^{11})^{-1} t\right\},$$

where  $A^{11}$  is the  $q \times q$  lower right-hand submatrix of  $A^{-1}$ ; this follows from properties of the multinormal density. Substituting  $\theta = \hat{\theta} + s/\sqrt{n}$  and  $\delta = \hat{\delta} + t$  in the  $\lambda_n$  integral now leads to

$$\begin{aligned} \lambda_n &= L_n(\hat{\theta}, \hat{\gamma}) n^{-p/2} \int Q_n(s, t) \pi_0(\hat{\theta} + s/\sqrt{n}) \pi(\hat{\delta} + t) ds dt \\ &\doteq L_n(\hat{\theta}, \hat{\gamma}) n^{-p/2} \pi_0(\hat{\theta}) (2\pi)^{p/2} |J_n|^{-1/2} |J_n^{11}|^{-1/2} \int \pi(\hat{\delta} + t) \exp\left\{-\frac{1}{2} t^t (J_n^{11})^{-1} t\right\} dt. \end{aligned}$$

This proves the claims made.

In this exposition we have glossed over certain technicalities that in a more careful proof need attention. These have to do with process convergence of  $\log Q_n(s, t)$  over the space of functions of  $(s, t)$  defined over a compact region, and secondly with limiting the size and influence of  $\log Q_n(s, t)$  outside such a compact region. We omit these details here, but refer to techniques and details provided in Hjort (1986), invented and developed there for a different but sufficiently similar problem. ■

PROOF OF PROPOSITION 9.2. We start re-expressing

$$\begin{aligned}\tilde{\mu}_S &= \frac{\int \mu(\theta, \gamma_{0,S} + \delta_S/\sqrt{n}) L_{n,S}(\theta, \gamma_{0,S} + \delta_S/\sqrt{n}) \pi_0(\theta) \pi_S(\delta_S) d\theta d\delta_S}{\int L_{n,S}(\theta, \gamma_{0,S} + \delta_S/\sqrt{n}) \pi_0(\theta) \pi_S(\delta_S) d\theta d\delta_S} \\ &= \frac{\int \mu(\hat{\theta}_S + s/\sqrt{n}, \gamma_{0,S} + (\hat{\delta}_S + t)/\sqrt{n}) Q_{n,S}(s, t) \pi_0(\hat{\theta}_S + s/\sqrt{n}) \pi_S(+\hat{\delta}_S + t) ds dt}{\int Q_{n,S}(s, t) \pi_0(\hat{\theta}_S + s/\sqrt{n}) \pi_S(\hat{\delta}_S + t) ds dt},\end{aligned}$$

in terms of

$$Q_{n,S}(s, t) = \frac{L_{n,S}(\hat{\theta}_S + s/\sqrt{n}, \gamma_{0,S} + (\hat{\delta}_S + t)/\sqrt{n})}{L_{n,S}(\hat{\theta}_S, \hat{\gamma}_S)} \doteq \exp\left\{-\frac{1}{2} \binom{s}{t}^t J_{n,S} \binom{s}{t}\right\},$$

in generalisation of the  $Q_n$  process used in the proof of the previous proposition. Taylor expanding the  $\mu$  term here w.r.t. the first parameter gives  $\mu(\hat{\theta}_S, \gamma_{0,S} + (\hat{\delta}_S + t)/\sqrt{n})$  plus  $(\partial\mu/\partial\theta)(\hat{\theta}_S, \gamma_{0,S} + (\hat{\delta}_S + t)/\sqrt{n})$  times  $s/\sqrt{n}$ , and then integrating over  $s$ , as with Proposition 7.1, shows indeed that  $\sqrt{n}(\tilde{\mu}_S - \bar{\mu}_S) \rightarrow_p 0$ . A fact used here is that the integral of  $s$  times the limit of  $Q_{n,S}(s, t)$ , over  $s$ , is zero.

For the rest of the proof, we use Taylor expansion w.r.t. the second parameter, and find

$$\bar{\mu}_S = E\{\hat{\mu}_S + (\frac{\partial\mu}{\partial\gamma_S})^t (\delta_S - \hat{\delta}_S)/\sqrt{n} | \hat{\delta}_S\} + o_p(n^{-1/2}),$$

which when compared to Lemma 3.3 gives the required result. ■

### Acknowledgements

The authors owe particular thanks to Editor Frank Samaniego for his constructive attention and consistent encouragement, and are also grateful to the Associate Editor and referees for their critical comments, which led to an improved presentation. The research of Claeskens is partly supported by NSF Grant DMS-02-03884.

### References

- Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.
- Berger, J.O. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related Topics III*, editors Berger and Gupta, 109–141. Academic Press, New York.

- Bickel, P.J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Annals of Statistics* **9**, 1301–1309.
- Bickel, P.J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. *Recent Advances in Statistics, Festschrift for Herman Chernoff* (eds. M.H. Rizvi, J. Rustagi and D. Siegmund), 511–528. Academic Press, New York.
- Bickel, P.J. (1984). Parametric robustness: small biases can be worthwhile. *Annals of Statistics* **12**, 864–879.
- Breiman, L. (2001). Statistical modeling: The two cultures [with discussion]. *Statistical Science* **16**, 199–231.
- Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603–618.
- Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer, New York.
- Bühlmann, P. (1999). Efficient and adaptive post-model-selection estimators. *Journal of Statistical Planning and Inference* **79**, 1–9.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* **158**, 419–466.
- Claeskens, G. and Hjort, N.L. (2003). The focussed information criterion. *Journal of the American Statistical Association*, to appear[!].
- Draper, D. (1995). Assessment and propagation of model uncertainty [with discussion]. *Journal of the Royal Statistical Society, Series B* **57**, 45–97.
- Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22**, 1947–1975.
- George, E.I. (1986a). Minimax multiple shrinkage estimation. *Annals of Statistics* **14**, 188–205.
- George, E.I. (1986b). Combining minimax shrinkage estimators. *Journal of the American Statistical Association* **81**, 437–445.
- Green, P.J. (2003). Trans-dimensional Markov chain Monte Carlo [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort and S. Richardson), 179–206. Oxford University Press, 2003.
- Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial [with discussion]. *Statistical Science* **19**, 382–417. [A version where the number of misprints has been significantly reduced is available at <http://www.stat.washington.edu/raftery/>.]
- Hjort, N.L. (1986). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics* **13**, 63–85.
- Hjort, N.L. and Pollard, D. (1994). Asymptotics for minimisers of convex processes. Statistical Research Report, Department of Mathematics, University of Oslo.
- Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.
- Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

- Hurvich, C.M. and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician* **44**, 214–217.
- Kabaila, P. (1995). The effect 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.
- Leeb, H. and Pötscher, B.M. (2000). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. Technical Report TR 2000-03, Institut für Statistik und Decision Support Systems, Universität Wien.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.
- Pötscher, B.M. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163–185.
- Rao, J.S. and Tibshirani, R. (1997). The out-of-bootstrap method for model averaging and selection. Technical Report, Department of Statistics, University of Toronto.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Sen, P.K. and Saleh, A.K.M.E. (1987). On preliminary test and shrinkage  $M$ -estimation in linear models. *The Annals of Statistics* **15**, 1580–1592.
- Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit [with discussion]. *Journal of the Royal Statistical Society B* **64**, 583–639.
- Yang, Y. (2001). Adaptive regression by mixing. *Journal of American Statistical Association* **96**, 574–588.