The Focussed Information Criterion

Gerda Claeskens and Nils Lid Hjort

Texas A&M University and University of Oslo

ABSTRACT. A variety of model selection criteria have been developed, of general and specific types. Most of these aim at selecting a single model with good overall properties, e.g. formulated via average prediction quality or shortest estimated overall distance to the in some sense true model. The Akaike, the Bayesian and the deviance information criteria AIC, BIC, DIC, along with many suitable variations, are eminent examples of such methods, and are in frequent use. These methods are however not concerned with the actual use of the selected model, which varies with context and application.

The present paper takes the view that the model selector should instead focus on the parameter singled out for interest; in particular, a model which gives good precision for one estimand may be worse when used for inference for another estimand. We develop a method which for given focus parameter estimates the precision of any submodel-based estimator. The framework is that of large-sample likelihood inference. Using an unbiased estimate of limiting risk, we propose a focussed information criterion for model selection, the FIC. We investigate and discuss properties of the method, establish some connections to the AIC, and illustrate its use in a variety of situations.

KEY WORDS: *Akaike's information criterion, bias and variance balance, the focussed information criterion, logistic regression, moderate misspecification, variable selection*

1. Introduction and summary

Central to any statistical data analysis is the idea of selecting an adequate model. This selection process might involve mathematical deduction from assumptions about the physics of the underlying data generation mechanism, comparisons with models used on previous but similar occasions, and sometimes ad hoc arguments which perhaps have to do with tradition or convenience. The process is sometimes complex and difficult to formalise or specify a priori, as it might involve informal checks of residuals and other diagnostic plots. Among the more formalised techniques used by statisticians as ingredients of this endeavour are goodness-of-fit tests and model selection using established model information criteria. This article is concerned with the latter, but deviates from mainstream in that we allow different models to be selected for different parameters of interest. This reflects the view that one model might be best for inference about say the mean structure while a different one might be preferable for analysing the variance structure. In this section we first give a brief review of popular model information criteria, before presenting motivation and a guidemap for the rest of the article.

1.1. Model information criteria in popular use. One popular and well-studied method is Akaike's information criterion, the AIC (Akaike, 1973). As an estimated expected

Kullback–Leibler distance, it aims at finding, amongst the set of models under consideration, the best approximating model to the unknown true data generating process. Using a principle of parsimony, Akaike's method will select the model with the fewest parameters which fits the data well. For any model S, AIC is defined as twice the maximised log-likelihood for model S, penalised with twice the number of parameters in S. Its applications abound, ranging from multiple regression (Shibata, 1981, Nishii, 1984) to the well-studied selection of the order in autoregressive time series (e.g. Shibata, 1976) and neural networks (Murata, Yoshizawa and Amari, 1994), to name a few. For more information and worked out examples we refer to the monographs Linhart and Zucchini (1986) and Burnham and Anderson (2002). See also Sections 5.6 and 6.1 below.

A related model selection mechanism, the Bayesian information criterion BIC of Schwarz (1978) penalises instead with the logarithm of the sample size; see also Rissanen (1989) for additional arguments and results. Although this leads to a consistent model selector if the true data generating model belongs to the finite-parameter family under investigation, as shown by Haughton (1989) for exponential families, BIC selected models tend to underfit if this assumption does not hold. See also Wei (1992), who develops Rissanen's predictive least squares principles further and suggests a related criterion based on Fisher information. A recent Bayesian development is the deviance information criterion DIC proposed and discussed in Spiegelhalter, Best, Carlin and van der Linde (2002), based on adjusting the posterior mean deviance with a penalty term for complexity. It is well fitted to those models where analysis is being carried out via Markov chain Monte Carlo to assess the posterior distributions. Other recently developed criteria include covariance-based and adaptive penalties, see Ye (1998), Tishirani and Knight (1999), Shen and Ye (2002), and George and Foster (2000).

Several model choice criteria become equivalent when the sample size grows. Stone (1977) obtains equivalence between cross-validation and AIC. See Stone (1974) for a more detailed account on cross-validation model choice. Nishii (1984) shows that AIC, finite prediction error (Akaike, 1970), Mallows's (1973) C_p and the prediction sum of squares (Allen, 1971) are equivalent in having the same risk function in the limit. For a further overview of model selection methods, see Shao (1997) and McQuarrie and Tsai (1998).

Several other variations on the AIC theme exist. Takeuchi (1976) constructs an asymptotically unbiased estimator of the relative Kullback–Leibler distance by choosing the penalisation term equal to twice the estimated trace of a matrix product $\Omega(\theta)J(\theta)^{-1}$ where Ω represents the variance matrix of the first order derivatives and J minus the expected value of the matrix of second order derivatives of the log likelihood with respect to the parameter vector θ. The resulting model information criterion is sometimes called the TIC. Note that in case the true data generating model is a member of the model class under investigation, the matrices J and Ω coincide and the criterion simplifies to AIC. A similar construction is used for the network information criterion in neural networks (Murata, Yoshizawa and Amari, 1994, Ripley, 1996), giving the so-called NIC. Basu, Har-

ris, Hjort and Jones (1998) and Jones, Hjort, Harris and Basu (2001) develop a certain robustification of the maximum likelihood estimation method for general parametric families, and inside that framework supplement fitted models with a robustified information criterion, say the RIC, which has the AIC as a limit when a certain algorithmic parameter governing the degree of robustification is sent to its null value. Other adjustments include finite sample corrections and extensions to quasi-likelihood (Hurvich and Tsai, 1995) and semiparametric and additive model selection (Simonoff and Tsai, 1999). Issues of model selection in so-called data mining are addressed by Chatfield (1995) and Ye (1998).

Model selectors are not only used for mere model selection; regularly they are the core of formal lack of fit tests, see e.g. the AIC-based order selection tests of Eubank and Hart (1992) and Aerts, Claeskens and Hart (1999, 2000), the BIC-type test statistic of Ledwina (1994), and the general goodness-of-fit tests of Claeskens and Hjort (2003).

1.2. The FIC and the present paper. The idea of finding a single satisfactory statistical model for one's data, perhaps aided by model information criteria as above, is a central one in statistics, and carries with it considerable intellectual and conceptual appeal. The chosen model is fitted to data and is seen as the statistician's best approximation to the real data generating mechanism used by nature, and secures a coherent view of statistical analysis of a data set. In this article we carefully extricate ourselves from this classic point of view; that a single model should be used to explain all aspects of data or to predict all types of future data points seems to us a little too constrained. Our view is that such a 'best model' should depend on the parameter under focus.

In practice, model selection is often only a first step in statistical analysis. All of the above mentioned model selectors essentially sidestep this fact of statistical life and provide us with one single 'best' model, regardless of the purpose of the selection, irrespective of the inference to follow. This is our main motivation for constructing a more focussed information criterion, the FIC, tailored to the parameter singled out for interest. Such a parameter, say μ, must have a definition making it meaningful across competing models. In Section 2 we set up a broad framework for comparing competing parametric models, in particular encompassing the model choice problems associated with covariate subset selection in regression models. Our framework uses general parametric models and maximum likelihood as the estimation method of choice, and is amenable to analysis by general large-sample theory, as developed in Hjort and Claeskens (2003). In particular, an expression is derived for the limiting risk of any submodel-based estimator of the μ parameter. The focussed information criterion FIC emerges in Section 3 as the result of establishing an unbiased estimate of this limiting risk; the candidate model with the smallest value of FIC is chosen. The FIC values are easily obtained via standard statistical software. We illustrate their use in Section 4 in a list of general and specific applications. One of these concerns determining factors influencing the probability of a child being born with low birth weight. It is seen that for different natural estimands, different subsets of the regressors are singled

out as most important. This conflicts with the 'one single model is used to explain everything' tradition, but is not a paradox as such: two different estimands might simply be associated with two different influential subsets of covariates. There might be conflicting aims regarding interpretation and transparency on one hand versus prediction quality and estimator precision on the other, depending on the context and problem formulation, see e.g. the discussion in Breiman (2001), but here our methodology is geared by the logic of prediction and precision of estimators.

Section 5 provides several remarks and viewpoints pertaining to the FIC, including various connections to the AIC, which therefore is afforded additional insight. Precise large-sample results for the behaviour of the FIC and AIC selected estimators are reached and compared in Section 6. Further developments are then discussed in Section 7, including natural empirical Bayesian versions of the FIC. An assumption underlying the development of the FIC is that the true data generating mechanism is contained in the largest parametric model considered. Certain amendments are called for when this assumption is not deemed viable, as explained in Section 8. Finally proofs of two technical results are provided in Section 9.

2. Estimators in a model choice framework

Our aim here is to study model selection schemes based on behaviour of the resulting estimator-post-selection. Such estimators are special cases of the more general classes of compromise methods studied in our companion paper Hjort and Claeskens (2003), where a general machinery is developed. To make the present article self-contained we need a concise summary of the other paper's Sections 2 and 3, including basic notation.

2.1. The i.i.d. setup. The start assumption here is that independent data $Y_1, \ldots, Y_n$ come from a density of the form

$$f_{\rm true}(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}). \tag{2.1}$$

Here θ is p-dimensional and γ is q-dimensional, the idea being to study perturbations of the $f(y,\theta) = f(y,\theta,\gamma_0)$ model around $\gamma = \gamma_0$. Thus γ_0 is known, determined by the statistical problem of interest. Models are considered which include the full θ but potentially only some or none of the γ components. For a parameter μ of interest, a function of the underlying density, we may write $\mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. There are (at least) 2^q model subset estimators to consider, one for each subset S of $\{1, \ldots, q\}$. The maximum likelihood estimator corresponding to having selected S is $\widehat{\mu}_S = \mu(\widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})$, where $(\widehat{\theta}_S, \widehat{\gamma}_S)$ are maximum likelihood estimators in the model which contains θ and then only γ_js for $j \in S$. Let $J_{\rm full}$ be the $(p+q)\times(p+q)$ information matrix of the full model, where $S = \{1, \ldots, q\}$, assumed to be of full rank; this is the variance matrix of the score function, evaluated at the null point (θ_0, γ_0), with blocks $J_{00}, J_{01}, J_{10}, J_{11}$. We shall also need projection mappings π_S of size $|S| \times q$ which maps $v = (v_1, \ldots, v_q)^{\rm t}$ to v_S, those v_js which have $j \in S$; for

the full model, $\pi_{\rm full}$ is the identity matrix. Here $|S|$ denotes the number of elements in S. Let $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$ and $K_S = (\pi_S K^{-1}\pi_S^{\rm t})^{-1}$. Two further quantities of importance are

$$H_S = K^{-1/2}\pi_S^{\rm t}K_S\pi_S K^{-1/2} \quad \text{and} \quad \omega = J_{10}J_{00}^{-1}\tfrac{\partial\mu}{\partial\theta} - \tfrac{\partial\mu}{\partial\gamma}, \tag{2.2}$$

with the partial derivatives evaluated at the narrow model, where $S = \emptyset$, i.e. at (θ_0, γ_0). The H_S is a projection matrix (symmetric, idempotent, and orthogonal to $I - H_S$). For completeness we let $H_\emptyset$ be the null matrix of size $q \times q$. In order to state the result which will be the basis for our focussed criterion, we first mention

$$D_n = \widehat{\delta}_{\rm full} = \sqrt{n}(\widehat{\gamma}_{\rm full} - \gamma_0) \to_d D \sim {\rm N}_q(\delta, K), \tag{2.3}$$

see our companion paper for details and more discussion. We now have the following: The maximum likelihood estimator of μ in the S model has limiting distribution of the form

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_S = (\tfrac{\partial\mu}{\partial\theta})^{\rm t}J_{00}^{-1}M + \omega^{\rm t}(\delta - K^{1/2}H_S K^{-1/2}D), \tag{2.4}$$

where $M \sim {\rm N}_p(0, J_{00})$ is independent of D. See Lemma 3.3 of Hjort and Claeskens (2003).

Below we shall have occasion to use one more result from Hjort and Claeskens (2003, Section 3). Consider the Akaike score ${\rm AIC}_{n,S}$ for submodel S, which can be expressed as twice the maximised likelihood for model S minus two times $|S|$. Then, under (2.1) conditions,

$${\rm AIC}_{n,S} - {\rm AIC}_{n,\emptyset} \to_d {\rm AIC}_S = D^{\rm t}K^{-1/2}H_S K^{-1/2}D - 2|S|. \tag{2.5}$$

2.2. The regression framework. The above i.i.d. setup needs to be generalised to cover regression models, as explained and carried out in Hjort and Claeskens (2003). The point of departure is that independent observations $Y_1, \ldots, Y_n$ are available, where Y_i comes from a density of the form $f_{i,{\rm true}}(y \,|\, x_i) = f(y \,|\, x_i, \theta_0, \gamma_0 + \delta/\sqrt{n})$. Here θ_0 typically consists of a p-dimensional vector of regression coefficients β, most often but not always with an additional scale parameter σ, and the model allows up to q additional γ_j parameters. Let S be any subset of $\{1, \ldots, q\}$. Estimators of a focus parameter $\mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$, constructed in the submodel indexed by S by means of the maximum likelihood method, take the form $\widehat{\mu}_S = \mu(\widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^{\rm c}})$. An analogue of the lemma above is reached in Hjort and Claeskens (2003, Sections 2 and 3). To define the necessary quantities, introduce $U(y \,|\, x)$ and $V(y \,|\, x)$, the partial derivatives of $\log f(y \,|\, x, \theta, \gamma)$ w.r.t. θ and γ, evaluated at the null point (θ_0, γ_0). We need

$$J(x) = {\rm Var}_0\begin{pmatrix} U(Y \,|\, x) \\ V(Y \,|\, x)\end{pmatrix} = \int f(y \,|\, x, \theta_0, \gamma_0)\begin{pmatrix} U(Y \,|\, x) \\ V(Y \,|\, x)\end{pmatrix}\begin{pmatrix} U(Y \,|\, x) \\ V(Y \,|\, x)\end{pmatrix}^{\rm t} {\rm d}y,$$

and an important matrix is

$$J_{n,{\rm full}} = n^{-1}\sum_{i=1}^n J(x_i) = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11}\end{pmatrix}, \tag{2.6}$$

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 positive definite $J_{\rm full}$ as n increases. In many situations there would be some ergodic phenomenon making it natural to postulate a covariate distribution Q for the x_is, so that averages $n^{-1}\sum_{i=1}^n g(x_i)$ tend to expectations $\mathrm{E}_Q g(x)=\int g(x)\,Q(\mathrm{d}x)$; in this case, $J_{\rm full}=\int J(x)\,Q(\mathrm{d}x)$, and so on. The Q distribution would be the conceptual limit of the empirical covariate distribution $Q_n=n^{-1}\sum_{i=1}^n \delta(x_i)$ as n grows; here $\delta(x_i)$ is unit point mass at position x_i. There are natural analogues of methods of results summarised in Section 2.1, in particular of results (2.3)–(2.4). These extensions involve matrices $K_n, K_{n,S}, H_{n,S}$ constructed from $J_{n,\rm full}$ along with the vector ω of determining coefficients, found as with (2.2).

3. The FIC

The (2.4) result yields expressions for the mean squared error of the limit distribution of $\widehat\mu_S$, for any S. Specifically, the limit distribution of $\sqrt n(\widehat\mu_S-\mu_{\rm true})$ has mean squared error

$$r(S)=\tau_0^2+\omega^{\rm t}(I-K^{1/2}H_SK^{-1/2})\delta\delta^{\rm t}(I-K^{-1/2}H_SK^{1/2})\omega+\omega^{\rm t}K^{1/2}H_SK^{1/2}\omega, \quad (3.1)$$

where $\tau_0^2=(\frac{\partial\mu}{\partial\theta})^{\rm t}J_{00}^{-1}\frac{\partial\mu}{\partial\theta}$. The programme now is to estimate this limiting risk quantity, for each S under consideration, thereby yielding a model choice criterion.

3.1. The FIC *for the limit experiment.* In (3.1) all quantities can be estimated with precision corresponding to familiar $\sqrt n$-rate consistency, with the crucial exception of δ. It is convenient to first argue directly in the limit experiment, where only D of (2.3) is random. We stress the simplicity and transparency of this limit experiment problem. In a sense all aspects of the model choice problem have been reduced to a single simple-structured problem: Which $\widehat\mu_S$ estimator should be used, when the risk $r(S)$ of (3.1) is associated with $\widehat\mu_S$, where all quantities are known except for δ, and when the single informative quantity to guide us is $D\sim\mathrm{N}_q(\delta,K)$? The various specifics of the parametric models being used and of the focus parameter, circumstances that will vary widely from application to application, have been reduced to simple and interpretable quantities like K, H_S and ω.

To estimate the limiting risk (3.1), note that $DD^{\rm t}$ has mean $\delta\delta^{\rm t}+K$, so we use $DD^{\rm t}-K$ as estimator for $\delta\delta^{\rm t}$. An unbiased estimator of limiting risk is therefore

$$\begin{aligned}\widehat r(S)&=\tau_0^2+\omega^{\rm t}(I-K^{1/2}H_SK^{-1/2})(DD^{\rm t}-K)(I-K^{-1/2}H_SK^{1/2})\omega+\omega^{\rm t}K^{1/2}H_SK^{1/2}\omega\\&=\tau_0^2+\{\omega^{\rm t}K^{1/2}(I-H_S)K^{-1/2}D\}^2+2\omega^{\rm t}K^{1/2}H_SK^{1/2}\omega-\omega^{\rm t}K\omega.\end{aligned}$$

Introducing $\widetilde\psi_{\rm full}=\omega^{\rm t}D$ for the full-model estimator of $\psi=\omega^{\rm t}\delta$, in the limit experiment, along with $\widetilde\psi_S=\omega^{\rm t}K^{1/2}H_SK^{-1/2}D$ for the S-subset estimator, we see that the above is the constant $\tau_0^2-\omega^{\rm t}K\omega$ away from the quantity

$$\begin{aligned}\mathrm{FIC}&=\omega^{\rm t}(I-K^{1/2}H_SK^{-1/2})DD^{\rm t}(I-K^{1/2}H_SK^{-1/2})^{\rm t}\omega+2\omega^{\rm t}K^{1/2}H_SK^{1/2}\omega\\&=(\widetilde\psi_{\rm full}-\widetilde\psi_S)^2+2\omega_S^{\rm t}K_S\omega_S.\end{aligned}\quad (3.2)$$

As before, $\omega_S = \pi_S \omega$. In other words, FIC is an unbiased estimator for $r(S)$ plus an additive constant not depending on S. The submodel with the smallest value of FIC is chosen.

3.2. The real FIC. The FIC above was derived for the limit experiment, where only D is random. For a real situation we must also estimate the ω and K_S, H_S quantities from data, and for D we must insert $D_n = \widehat{\delta}_{\rm full}$ of (2.3). From (2.2), this requires a suitable $\widehat{J}_{\rm full}$, from which we compute $\widehat{K}$, $\widehat{K}_S$ and $\widehat{H}_S$, along with consistent estimates of $\frac{\partial \mu}{\partial \theta}$ and $\frac{\partial \mu}{\partial \gamma}$. Such can be constructed by plugging in an estimate of θ in explicit formulae, if available, or via numerical approximations, say $\{\mu(\widehat{\theta}+\eta e_i, \gamma_0) - \mu(\widehat{\theta}, \gamma_0)\}/\eta$ for the components of $\frac{\partial \mu}{\partial \theta}$ and $\{\mu(\widehat{\theta}, \gamma_0 + \eta e_j) - \mu(\widehat{\theta}, \gamma_0)\}/\eta$ for the components of $\frac{\partial \mu}{\partial \gamma}$, for a small η value, where e_i is the ith unit vector.

Writing $\widehat{\psi}_{\rm full} = \widehat{\omega}^{\rm t}\widehat{\delta}_{\rm full}$ and $\widehat{\psi}_S = \widehat{\omega}^{\rm t}\widehat{K}^{1/2}\widehat{H}_S\widehat{K}^{-1/2}\widehat{\delta}_{\rm full}$, the result is the 'real' focussed information criterion

$$\begin{aligned}\widehat{\rm FIC} &= \widehat{\omega}^{\rm t}(I - \widehat{K}^{1/2}\widehat{H}_S\widehat{K}^{-1/2})\widehat{\delta}_{\rm full}\widehat{\delta}_{\rm full}^{\rm t}(I - \widehat{K}^{1/2}\widehat{H}_S\widehat{K}^{-1/2})^{\rm t}\widehat{\omega} + 2\widehat{\omega}_S^{\rm t}\widehat{K}_S\widehat{\omega}_S \\ &= (\widehat{\psi}_{\rm full} - \widehat{\psi}_S)^2 + 2\widehat{\omega}_S^{\rm t}\widehat{K}_S\widehat{\omega}_S.\end{aligned} \tag{3.3}$$

We note that for $\widehat{K}$ diagonal, matters simplify to

$$\widehat{\rm FIC} = \Big(\sum_{j\notin S} \widehat{\omega}_j\widehat{\delta}_{{\rm full},j}\Big)^2 + 2\sum_{j\in S}\widehat{\omega}_j^2\widehat{k}_j. \tag{3.4}$$

Also note that computing the (3.3) quantities for all submodels S of interest is an easy programming task, as long as estimates have been obtained for γ (in the fullest model), $J_{\rm full}$ (consistency under narrow model circumstances suffices), and ω.

There will typically be several estimation strategies for the key matrix $J_{\rm full}$ (with consequences for estimated K, K_S, H_S, ω). One might often enough find explicit formulae for the entries of the information matrix $J_{\rm full} = J(\theta_0, \gamma_0)$, see examples of Section 4, so that each entry may be estimated by plug-in, using either $\widehat{\theta}_{\rm narr}$ or $\widehat{\theta}_{\rm full}$. A simple and satisfactory alternative is by calculating the variance matrix of say 10,000 simulated score vectors at the estimated null model (this is sometimes much easier than from the estimated full model), if no formula or Hessian matrix is available. Note that application of the theory only requires estimators that are consistent under the narrow null model. We would often nevertheless wish to use an estimator of $J_{\rm full}$ constructed from the wide model $f(y, \theta, \gamma)$, thereby securing some model robustness in that one then does not have to rely on γ being close to γ_0.

4. Illustrations and applications

This section provides a list of illustrations of the FIC apparatus, partly listed as general recipes for models of interest, and partly in a specific logistic regression type application concerned with the probability of low birth weights.

4.1. A skewed regression model. For regression data (x_i, Y_i), let $Y_i \sim \mathrm{N}(\beta_0, \sigma^2)$ be the narrow model, around which we consider model departures in two directions, namely in the mean and in skewness. Specifically, the fullest model has $Y_i = \beta_0 + \beta_1 x_i + \sigma\varepsilon_i$, where ε_i comes from the skewed distribution with density $\lambda\Phi(u)^{\lambda-1}\phi(u)$. Here ϕ and Φ are the density and cumulative for the standard normal, and λ is the skewness parameter in question. Note that σ changes interpretation and value with β_1 and λ. In this situation we consider different parameters of interest when (β_1, λ) is in the vicinity of the narrow model's $(0,1)$, and are led to $\mu_{\rm true} = \mu(\beta_0, \sigma, \beta_1, \lambda) = \mu(\beta_0, \sigma, \delta_1/\sqrt{n}, 1+\delta_2/\sqrt{n})$. There are four model-selector estimator candidates, ranging from the simplest $\mu(\bar y, s, 0, 1)$ sticking in ordinary sample mean and standard deviation, to $\mu(\widehat\beta_{0,\rm full}, \widehat\sigma_{\rm full}, \widehat\beta_{1,\rm full}, \widehat\lambda_{\rm full})$ with estimates from full likelihood in the widest model.

The log-density for Y_i becomes

$$\log\lambda + (\lambda-1)\log\Phi\Big(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\Big) - \log\sigma - \tfrac12\Big(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\Big)^2 - \tfrac12\log(2\pi),$$

and after some algebra one finds that the score vector of partial derivatives, evaluated at the narrow model, has components ε_i/σ, $(\varepsilon_i^2-1)/\sigma$, $\varepsilon_i x_i/\sigma$, $1+\log\Phi(\varepsilon_i)$, with ε_i a standard normal. This leads with further analysis to

$$J_{n,\rm full} = \begin{pmatrix} 1/\sigma^2 & 0 & \bar x/\sigma^2 & c/\sigma \\ 0 & 2/\sigma^2 & 0 & d/\sigma \\ \bar x/\sigma^2 & 0 & (v_n^2+\bar x^2)/\sigma^2 & c\bar x/\sigma \\ c/\sigma & d/\sigma & c\bar x/\sigma & 1 \end{pmatrix}.$$

Here $\bar x$ and v_n^2 are the empirical mean and variance of the x_is, while $c = \mathrm{cov}\{\varepsilon_i, \log\Phi(\varepsilon_i)\} = 0.9032$ and $d = \mathrm{cov}\{\varepsilon_i^2, \log\Phi(\varepsilon_i)\} = -0.5956$. Inverting this matrix shows that $K_n = J_n^{11}$ is diagonal, with elements $k_{n,1} = \sigma^2/v_n^2$ and $k_{n,2} = 1/(1-c^2-\frac12 d^2) = 12.0877^2$. The general focussed information criterion becomes

$$\widehat{\rm FIC} = \Big(\sum_{j\notin S}\widehat\omega_j\widehat\delta_j\Big)^2 + 2\sum_{j\in S}\widehat\omega_j^2\widehat k_{n,j}, \tag{4.1}$$

where $\widehat\delta_1 = \sqrt{n}\widehat\beta_{1,\rm full}$ and $\widehat\delta_2 = \sqrt{n}(\widehat\lambda_{\rm full}-1)$, while $\widehat\omega_1$ and $\widehat\omega_2$ are estimates of

$$\omega_1 = \bar x\,\tfrac{\partial\mu}{\partial\beta_0} - \tfrac{\partial\mu}{\partial\beta_1} \quad\text{and}\quad \omega_2 = c\sigma\tfrac{\partial\mu}{\partial\beta_0} + \tfrac12 d\sigma\tfrac{\partial\mu}{\partial\sigma} - \tfrac{\partial\mu}{\partial\lambda}.$$

The partial derivatives are evaluated (and estimated, when necessary) in the narrow model where $(\beta_1,\lambda) = (0,1)$. The submodel is chosen with smallest value of

$$\widehat{\rm FIC} = \begin{cases} (\widehat\omega_1\widehat\delta_1 + \widehat\omega_2\widehat\delta_2)^2 & \text{for the narrow model,} \\ \widehat\omega_2^2\widehat\delta_2^2 + 2\widehat\omega_1^2 k_1 & \text{for including } \beta_1\text{, not } \lambda, \\ \widehat\omega_1^2\widehat\delta_1^2 + 2\widehat\omega_2^2 k_2 & \text{for including } \lambda\text{, not } \beta_1, \\ 2(\widehat\omega_1^2 k_1 + \widehat\omega_2^2 k_2) & \text{for the full model.} \end{cases} \tag{4.2}$$

We give four brief examples.

Example 1: Let μ be the mean of $Y = \beta_0 + \beta_1 x + \sigma\varepsilon$ for some given covariate value x, i.e. $\beta_0 + \beta_1 x + \sigma e(\lambda)$, where $e(\lambda) = \int u\lambda\Phi(u)^{\lambda-1}\phi(u)\,\mathrm{d}u$. Here one finds $\omega_1 = \bar{x} - x$ and $\omega_2 = 0$, using the derivative result $e'(1) = c$. For this mean estimand there is no award in involving the λ aspects of the data, as the added complexity does not alter the large-sample performance of estimators, hence the question is reduced to choosing between the narrow (β_0, σ) model or the broader $(\beta_0, \sigma, \beta_1)$ model. If $\sqrt{n}|\widehat{\beta}_{1,\mathrm{full}}|v_n/\widehat{\sigma} < \sqrt{2}$, then we leave out β_1, otherwise, we include it.

Example 2: Let μ instead be the median at covariate value x, i.e. $\mu = \beta_0 + \beta_1 x + \sigma\Phi^{-1}((\frac{1}{2})^{1/\lambda})$. Here $\omega_1 = \bar{x} - x$ and $\omega_2 = \sigma\{c - \frac{1}{2}(\log 2)/\phi(0)\} = 0.1313\,\sigma$. Model choice proceeds using (4.2).

Example 3: Consider the third central moment $\mu = \mathrm{E}(Y - \mathrm{E}Y)^3 = \sigma^3\mathrm{E}\{\varepsilon - e(\lambda)\}^3$. Here $\omega_1 = 0$, signalling that the inference is not touched by inclusion or exclusion of the β_1 parameter, while some work yields $\omega_2 = -0.2203\,\sigma^3$. If $\sqrt{n}|\widehat{\lambda}_{\mathrm{full}} - 1|/k_{n,2}^{1/2} < \sqrt{2}$ the narrow model suffices; otherwise we include λ. There is a similar conclusion when the focus parameter is the skewness $\mathrm{E}(Y - \mathrm{E}Y)^3/\{\mathrm{E}(Y - \mathrm{E}Y)^2\}^{3/2}$.

Example 4: Look at the cumulative distribution function $\Pr\{Y \le y\} = \Phi((y - \beta_0 - \beta_1 x)/\sigma)^\lambda$ associated with a given x value. Computing the derivatives w.r.t. $\beta_0, \sigma, \beta_1, \lambda$ one finds

$$\widehat{\omega}_1 = \frac{x - \bar{x}}{s}\phi\Big(\frac{y - \bar{y}}{s}\Big), \quad \widehat{\omega}_2 = -\Big(\tfrac{1}{2}d\frac{y - \bar{y}}{s} + c\Big)\phi\Big(\frac{y - \bar{y}}{s}\Big) - \Phi\Big(\frac{y - \bar{y}}{s}\Big)\log\Phi\Big(\frac{y - \bar{y}}{s}\Big),$$

in terms of average $\bar{y}$ and standard deviation s for the Y_i sample. Again model selection uses (4.2).

The situation considered here can be generalised to one with say $Y_i = \beta_0 + x_i^{\mathrm{t}}\beta_1 + u_i^{\mathrm{t}}\beta_2 + \sigma\varepsilon_i$, where the x_is are always to be included in the model whereas the u_is are extra candidates, along with the extra λ parameter for skewness of the noise part.

As mentioned at the end of Section 3 we would perhaps wish to use a wide model based estimator of J_{full}, for reasons of model robustness. The theory applies, but with less clear-cut formulae for FIC than in (4.1), in that the $\widehat{K}$ matrix involved would not be diagonal.

4.2. Variable selection in the linear normal model. Let $Y_1, \ldots, Y_n$ be independent and normal with the same variability, where it is not clear a priori which of several covariates to include in the model. For the $\mathrm{N}(x_i^{\mathrm{t}}\beta + u_i^{\mathrm{t}}\gamma, \sigma^2)$ model, where the intention is to always include x_i whereas components of u_i may or may not be included, one finds

$$J_{n,\mathrm{full}} = \frac{1}{\sigma^2}\begin{pmatrix} \Sigma_{n,00} & 0 & \Sigma_{n,01} \\ 0 & 2 & 0 \\ \Sigma_{n,10} & 0 & \Sigma_{n,11} \end{pmatrix},$$

where $\Sigma_{n,00} = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}}$, $\Sigma_{n,01} = n^{-1}\sum_{i=1}^n x_i u_i^{\mathrm{t}}$, and $\Sigma_{n,11} = n^{-1}\sum_{i=1}^n u_i u_i^{\mathrm{t}}$. This leads to $\omega = \Sigma_{n,10}\Sigma_{n,00}^{-1}\frac{\partial\mu}{\partial\beta} - \frac{\partial\mu}{\partial\gamma}$. For the parameter $\mu = x^{\mathrm{t}}\beta + u^{\mathrm{t}}\gamma$, for example, the mean of

Y at position (x,u), this gives $\omega = \Sigma_{n,10}\Sigma_{n,00}^{-1}x-u$. Ingredients in the general recipe include $\widehat{\delta} = \sqrt{n}\widehat{\gamma}$ in the full model, the matrix $K_n = \sigma^2 L_n$ where $L_n = (\Sigma_{n,11}-\Sigma_{n,10}\Sigma_{n,00}^{-1}\Sigma_{n,01})^{-1}$, the full estimator $\widehat{\psi}_{\rm full} = \omega^{\rm t}\widehat{\delta}$ along with the submodel estimators $\widehat{\psi}_S = \omega^{\rm t}L_n^{1/2}H_{n,S}L_n^{-1/2}\widehat{\delta}$, where $H_{n,S} = L_n^{-1/2}\pi_S^{\rm t}L_{n,S}\pi_S L_n^{-1/2}$ and $L_{n,S} = (\pi_S L_n^{-1}\pi_S^{\rm t})^{-1}$. This leads to

$$\widehat{\rm FIC} = (\widehat{\psi}_{\rm full} - \widehat{\psi}_S)^2 + 2\widehat{\sigma}^2\omega_S^{\rm t}L_{n,S}\omega_S.$$

Thus the FIC depends in this case on the covariate position (x,u) via $\omega = \omega(x,u)$, indicating that there could be different suggested covariate models in different covariate regions. This is not a paradox, and stems from the wish of estimating $\mathrm{E}(Y\,|\,x,u)$ with optimal precision, for each given (x,u). See also Section 5.5, where it is seen that the general large-sample risk approximations used here actually match exactly the appropriate mean squared errors.

Sometimes interest focusses on the impact of a particular covariate on the mean structure. This fits in with $\mu = \xi(x+e_k,u)-\xi(x,u) = \beta_k$, writing $\xi(x,u)$ for $\mathrm{E}(Y\,|\,x,u)$ and e_k for the kth unit vector. The FIC machinery can then be set to work, with $\omega = \Sigma_{10}\Sigma_{00}^{-1}e_k$ etc.

Let us illustrate the variable selection method in a situation where one considers augmenting a linear regression trend with a quadratic or cubic term. This fits the above with a $\mathrm{N}(\beta_0+\beta_1x_i+\beta_2x_i^2+\beta_3x_i^3,\sigma^2)$ model for data Y_i. Assume without loss of generality that the x_is have mean zero, and let $m_j = n^{-1}\sum_{i=1}^n x_i^j$ for $j = 2,3,4,5,6$. Let furthermore $\widehat{\delta}_2 = \sqrt{n}\widehat{\beta}_2$ and $\widehat{\delta}_3 = \sqrt{n}\widehat{\beta}_3$ in the full model. For estimating $\mu(x) = \mathrm{E}(Y\,|\,x)$ one finds that ω has the two components $\omega_2 = m_2 - x^2 + (m_3/m_2)x$ and $\omega_3 = m_3 - x^3 + (m_4/m_2)x$, along with

$$L_n = \begin{pmatrix} m_4 - m_2^2 - m_3^2/m_2, & m_5 - m_2m_3 - m_3m_4/m_2 \\ m_5 - m_2m_3 - m_3m_4/m_2, & m_6 - m_3^2 - m_4^2/m_2 \end{pmatrix}^{-1}.$$

The four FIC signals read

$$\begin{aligned}
\widehat{\rm FIC}_0(x) &= (\omega_2\widehat{\delta}_2 + \omega_3\widehat{\delta}_3)^2, \\
\widehat{\rm FIC}_2(x) &= \{(\omega_3 - \omega_2L_{n,1}L_n^{01})\widehat{\delta}_3\}^2 + 2\widehat{\sigma}^2\omega_2^2L_{n,1}, \\
\widehat{\rm FIC}_3(x) &= \{(\omega_2 - \omega_3L_{n,2}L_n^{10})\widehat{\delta}_2\}^2 + 2\widehat{\sigma}^2\omega_3^2L_{n,2}, \\
\widehat{\rm FIC}_{23}(x) &= 2\widehat{\sigma}^2\omega^{\rm t}L_n\omega,
\end{aligned}$$

representing respectively the narrow model, the model which includes β_2, the model which includes β_3, and the fullest model with both β_2,β_3, where we use L_n^{ij} to indicate the elements of L_n^{-1}. Also, $L_{n,1} = (m_4 - m_2^2 - m_3^2/m_2)^{-1}$ and $L_{n,2} = (m_6 - m_3^2 - m_4^2/m_2)^{-1}$. The method consists in choosing for each given x the submodel with smallest observed $\mathrm{FIC}(x)$ value, with the consequent estimate or predictor for $\mathrm{E}(Y\,|\,x)$. One may rule out the model which uses β_3 but not β_2, if one wishes; this corresponds to ignoring $\mathrm{FIC}_3(x)$. The method gives $\widehat{\mu}(x)$ estimators of different form over different intervals of x, according to which of the $\mathrm{FIC}(x)$ monitors is smallest.

The above method may be contrasted with e.g. the AIC strategy, which is to select the model with smallest value of say $C_0 = n \log \widehat{\sigma}_0$ using variance estimate for the narrow model, $C_2 = n \log \widehat{\sigma}_2 + 1$ with estimate from the model with β_2, $C_3 = n \log \widehat{\sigma}_3 + 1$ with estimate from the model with β_3, and finally $C_{23} = n \log \widehat{\sigma}_{23} + 2$ with estimate from the model including both β_2, β_3. Analysis along the lines of Section 6 will indicate that the FIC strategy often will lead to more accurate final precision than the AIC scheme.

4.3. Two model departures from linear regression. Around the traditional linear normal regression model, in which $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, we now build model extensions in two directions, the mean and the variance. Specifically, take $Y_i = \beta_0 + \beta_1 x_i + \beta_2 u_i + \sigma_i \varepsilon_i$, where the ε_is are independent and standard normal, and $\sigma_i = \sigma \exp(\zeta v_i)$. One might e.g. have $u_i = x_i^2$, if parabolic behaviour of the regression curve is anticipated, and $v_i = x_i - \bar{x}$, if there are indications that the variance might be log-linear in x.

Taking the derivatives of the log-density of Y_i w.r.t. $\beta_0, \beta_1, \sigma, \beta_2, \zeta$, one finds ε_i/σ_i, $\varepsilon_i x_i/\sigma_i$, $(\varepsilon_i^2 - 1)/\sigma$, $\varepsilon_i u_i/\sigma_i$, $(\varepsilon_i^2 - 1)v_i$. At the null model, where $\sigma_i = \sigma$, calculations give

$$J_{n,\mathrm{full}} = \begin{pmatrix} 1/\sigma^2 & \bar{x}/\sigma^2 & 0 & \bar{u}/\sigma^2 & 0 \\ \bar{x}/\sigma^2 & (\bar{x}^2 + s_x^2)/\sigma^2 & 0 & n^{-1}\sum_{i=1}^n x_i u_i/\sigma^2 & 0 \\ 0 & 0 & 2/\sigma^2 & 0 & 2\bar{v}/\sigma \\ \bar{u}/\sigma^2 & n^{-1}\sum_{i=1}^n x_i u_i/\sigma^2 & 0 & (\bar{u}^2 + s_u^2)/\sigma^2 & 0 \\ 0 & 0 & 2\bar{v}/\sigma & 0 & 2(\bar{v}^2 + s_v^2) \end{pmatrix}$$

for the full model information matrix. Here $\bar{x}, \bar{u}, \bar{v}$ are the means of x_i, u_i and v_i, while $s_x^2 = n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$ and similarly for s_u and s_v. To calculate the K_n matrix that drives the FIC and its properties, let $\det_{i,j}$ be the determinant of the submatrix where line i and row j are omitted. Some algebra yields $\det_{5,5} = (2/\sigma^8) s_x^2 s_u^2 (1 - \rho_n^2)$, $\det_{5,3} = -(2\bar{v}/\sigma^7) s_x^2 s_u^2 (1 - \rho_n^2)$ and $\det_{4,4} = (4/\sigma^8) s_x^2 s_v^2$, where $\rho_n = (n^{-1}\sum_{i=1}^n x_i u_i - \bar{x}\bar{u})/s_x s_u$ is the empirical correlation between x_i and u_i values. This gives $\det = (4/\sigma^2) s_x^2 s_u^2 s_v^2 (1 - \rho_n^2)$ for the full information matrix, along with

$$k_{n,1} = \frac{\det_{4,4}}{\det} = \frac{1}{s_u^2(1 - \rho_n^2)} \quad \text{and} \quad k_{n,2} = \frac{\det_{5,5}}{\det} = \frac{1}{2s_v^2}.$$

Also, $\det_{4,5} = 0$, which means that $K_n = \mathrm{diag}(k_{n,1}, k_{n,2})$. For any given estimand $\mu(\beta_0, \beta_1, \sigma, \beta_2, \zeta)$ of interest, we also find its determining coefficients, from $\omega = J_{n,01} J_{n,00}^{-1} (\frac{\partial \mu}{\partial \beta_0}, \frac{\partial \mu}{\partial \beta_1}, \frac{\partial \mu}{\partial \sigma})^{\mathrm{t}} - (\frac{\partial \mu}{\partial \beta_2}, \frac{\partial \mu}{\partial \zeta})^{\mathrm{t}}$, namely

$$\omega_1 = \bar{u}\Big\{\Big(\frac{\bar{x}^2}{s_x^2} + 1\Big)\frac{\partial \mu}{\partial \beta_0} - \frac{\bar{x}}{s_x^2}\frac{\partial \mu}{\partial \beta_1}\Big\} + (\bar{x}\bar{u} + \rho_n s_x s_u)\Big(-\frac{\bar{x}}{s_x^2}\frac{\partial \mu}{\partial \beta_0} + \frac{1}{s_x^2}\frac{\partial \mu}{\partial \beta_1}\Big) - \frac{\partial \mu}{\partial \beta_2},$$
$$\omega_2 = \sigma\bar{v}\tfrac{\partial \mu}{\partial \sigma} - \tfrac{\partial \mu}{\partial \zeta}.$$

The story continues, for each given μ in focus, as in (4.1)–(4.2) above, with $\widehat{\delta}_1 = \sqrt{n}\widehat{\beta}_{2,\mathrm{full}}$ and $\widehat{\delta}_2 = \sqrt{n}\widehat{\zeta}_{\mathrm{full}}$. For μ equal to a general quantile of the distribution of Y for given x, u, i.e. $\beta_0 + \beta_1 x + \beta_2 u + q\sigma \exp(\zeta v)$, some algebra leads to

$$\omega_1 = (x - \bar{x})\rho_n s_u/s_x - (u - \bar{u}) \quad \text{and} \quad \omega_2 = -q\sigma(v - \bar{v}).$$

4.4. Selection in logistic regression models. We consider the data set given in Appendix 1 of Hosmer and Lemeshow (1989), which concerns a study of $n = 189$ women with newborn babies and factors potentially associated with low infant birth weight, there taken to mean less than 2500 gram. Covariate information for the mothers were available as weight just prior to pregnancy (x_2, in pounds), age (x_3), indicator for race 'black' (x_4), and indicator for race 'other' (x_5). Women with $x_4 = 0$ and $x_5 = 0$ are of race 'white'. For the full logistic regression model, where also the constant $x_1 = 1$ is included, estimates are equal to 1.306, -0.014, -0.026, 1.003, 0.443, and the corresponding $\widehat{\beta}_j/\mathrm{se}(\widehat{\beta}_j)$ ratios are 1.226, -2.215, -0.770, 2.020, 1.233. For this illustration we take the view that

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

where $x = (1, x_2)^{\mathrm{t}}$ is always in the model while subsets of $u = (x_3, x_4, x_5)^{\mathrm{t}}$ are considered for possible inclusion. Label for simplicity the submodels in question '0', '3', '4', '5', '34', '35', '45', '345', corresponding to inclusion or not of x_3, x_4, x_5. An AIC analysis, calculating twice the maximal log-likelihoods penalised with twice the number of parameters, indicates that the best submodel is '4', followed by '45'; see Table 4.1.

In this situation

$$J_{n,\mathrm{full}} = n^{-1}\sum_{i=1}^{n} p_i(1-p_i)\begin{pmatrix} x_i x_i^{\mathrm{t}} & x_i u_i^{\mathrm{t}} \\ u_i x_i^{\mathrm{t}} & u_i u_i^{\mathrm{t}} \end{pmatrix},$$

with $p_i = L(x_i^{\mathrm{t}}\beta + u_i^{\mathrm{t}}\gamma)$ and $L(u) = \exp(u)/\{1+\exp(u)\}$. We estimate this matrix and the consequent K, K_S and H_S matrices using the estimates for the five parameters given above, that is, in the full model; we could also have estimated $J_{n,\mathrm{full}}$ using estimators from the narrow model, where $\gamma = 0$. It is also instructive to perform a test for $\gamma = 0$ inside the extended model, where the natural approximative χ^2_3 test statistic is $\widehat{\delta}^{\mathrm{t}}\widehat{K}^{-1}\widehat{\delta} = 5.927$, in terms of the departure indicators $\widehat{\delta} = \sqrt{n}\widehat{\gamma}$, which here are equal to $-0.351, 13.799, 6.096$. Thus these, which are also needed with each FIC application, do not as such indicate any strong evidence against simply sticking to the x_1, x_2 model. We shall see that the FIC nevertheless advocates including further covariate information, for some natural estimands.

A natural focus parameter is $p(x,u)$ itself, for different (x,u) corresponding to different strata of mothers. For each representative woman we may search through the eight submodels for u and compute FIC values, using the estimated version of $\omega = p(x,u)\{1-p(x,u)\}(J_{n,10}J_{n,00}^{-1}x - u)$. We first consider women of race 'white' and let $x = (1, 132.05)^{\mathrm{t}}$ and $u = (24.29, 0, 0)^{\mathrm{t}}$, corresponding to average weight and age in that group. Here the estimated low birth rate probability varies from 0.230 (model '345') to 0.298 (model '0'), and $\widehat{\omega} = (-0.245, 0.032, 0.065)^{\mathrm{t}}$. The FIC is found to recommend model '34' for making the best prediction, followed by model '4'; see Table 4.1. The final estimate is 0.2638.

model	$-$AIC	white	w-FIC	black	b-FIC	other	o-FIC	ratio	r-FIC
0	232.691	0.298	0.860	0.256	5.099	0.334	0.158	0.861	291.806
3	233.123	0.288	0.654	0.272	4.171	0.337*	0.140*	0.945	231.353
4	231.075*	0.269	0.375	0.412*	2.813*	0.310	0.694	1.533	110.376
5	234.101	0.279	0.695	0.242	6.481	0.369	0.797	0.868	272.466
34	232.175	0.264*	0.315*	0.413	2.813*	0.314	0.625	1.564*	106.519*
35	234.677	0.272	0.573	0.259	5.373	0.368	0.796	0.950	218.330
45	231.259	0.231	0.383	0.414	2.813*	0.368	0.795	1.794	110.938
345	232.661	0.230	0.385	0.414	2.813*	0.367	0.796	1.801	111.016

TABLE 4.1. *For submodels corresponding to inclusion or not of covariates x_3, x_4, x_5, the table lists the minus* AIC*, along with estimates and* FIC *values for four estimands of interest. These are the low birth weight probabilities p*(white)*, p*(black)*, p*(other) *and the ratio p*(black)*/p*(white)*. The asterisk indicates the selected model and the consequent final estimates.*

Next consider women of race 'black', letting $x = (1, 146.81)^{\mathrm{t}}$ and $u = (21.54, 1, 0)^{\mathrm{t}}$; the average woman here is younger but a bit heavier than in the previous group considered. Here estimates of $p(x, u)$ range from 0.242 (model '5') to 0.414 (model '345'), and $\widehat{\omega} = (0.429, -0.185, 0.073)^{\mathrm{t}}$. The FIC is nearly undecided between models '34', '4', '45' and '345'. It is comforting to see that the four estimates in question are close. Similarly, for women of race 'other', we let $x = (1, 120.01)^{\mathrm{t}}$ and $u = (22.39, 0, 1)^{\mathrm{t}}$, corresponding again to average weight and age in that group. Here probability estimates range from 0.310 (model '4') to 0.368 (model '45'), and $\widehat{\omega} = (0.045, 0.032, -0.135)^{\mathrm{t}}$. The FIC recommends model '3' ahead of model '0'. Again, these two estimates, between which FIC finds it difficult to make up its mind, are very close, 0.337 and 0.334, respectively.

To demonstrate the versatility of the FIC we include a final example of a different nature. It appears from the estimates above that black mothers have a chance perhaps 1.5 times higher than white mothers of having a low birth weight for their child. To examine this, focus on $\mu = p(x', u')/p(x, u)$ for suitable (x, u) and (x', u'), for which we find

$$\omega = \frac{p(x', u')}{p(x, u)}\big[\{1 - p(x', u')\}(J_{10}J_{00}^{-1}x' - u') - \{1 - p(x, u)\}(J_{10}J_{00}^{-1}x - u)\big]$$

from (2.2). For (x', u') corresponding to the average black and (x, u) to the average white mother, one finds $\widehat{\omega} = (3.783, -1.058, -0.190)^{\mathrm{t}}$, and the FIC recommends '34' followed by '4' and '45' for making the best ratio estimation.

The low birth weight data application is further discussed in Section 7.1.

4.5. An extended binary regression model. In applications like that above, one is often geared a little too strongly by tradition towards using the exact logistic transform, which however has no a priori reason to be close to the 'true' probability function. It is therefore of interest to consider classes of models which generalise the logistic one. One relatively simple such is to use

$$p(x, u) = \Pr\{Y = 1 \mid x, u\} = \Big\{\frac{\exp(x^{\mathrm{t}}\beta + u^{\mathrm{t}}\gamma)}{1 + \exp(x^{\mathrm{t}}\beta + u^{\mathrm{t}}\gamma)}\Big\}^{\kappa}.$$

This model, which adds asymmetry to the logistic transform, is briefly discussed in Hosmer and Hjort (2002). In general terms, $p(x,u) = H_\kappa(L(x^{\rm t}\beta + u^{\rm t}\gamma))$, where H_κ is a function which for a κ_0 value is equal to the identity function, and L is as above. In the low birth weight application, this would mean working with a 6×6 information matrix

$$J_{n,{\rm full}} = n^{-1}\sum_{i=1}^n \begin{pmatrix} p_i(1-p_i)x_ix_i^{\rm t} & p_i(1-p_i)x_iu_i^{\rm t} & \bar H(p_i)x_i \\ p_i(1-p_i)u_ix_i^{\rm t} & p_i(1-p_i)u_iu_i^{\rm t} & \bar H(p_i)u_i \\ \bar H(p_i)x_i^{\rm t} & \bar H(p_i)u_i^{\rm t} & \bar H(p_i)/(p_i(1-p_i)) \end{pmatrix},$$

evaluated under the null model $\kappa = \kappa_0$, where $p_i = L(x_i^{\rm t}\beta + u_i^{\rm t}\gamma)$, and where $\bar H(v) = \partial H_\kappa(v)/\partial\kappa$ evaluated at the null point κ_0. For the departure function $H_\kappa(v) = v^\kappa$ suggested above, one has $\bar H(v) = v\log v$.

In this situation one needs maximum likelihood estimation of all six parameters, and the four departure indicators $\widehat\delta_j = \sqrt{n}\widehat\gamma_j$ for $j = 1,2,3$ and $\widehat\delta_4 = \sqrt{n}(\widehat\kappa - 1)$. For the focus parameter $p(x,u)$, the coefficients of the crucial parameter $\psi = \omega^{\rm t}\delta$ are

$$\omega = p(1-p)J_{n,10}J_{n,00}^{-1}x - \begin{pmatrix} p(1-p)u \\ \bar H(p) \end{pmatrix} \quad \text{where } p = p(x,u).$$

For the low birth rate application mentioned above we carried out such analysis, employing a Newton type iteration algorithm to maximise the extended likelihood. The log-likelihood did not climb with a sufficient amount for inclusion of κ in the model to be advantageous, however, in this particular example. Such one-parameter extensions of the logistic regression model have a better chance of being effective in situations with fewer covariates, say one or two.

4.6. Generalised linear models. Our methodology finds easy applications in most of the traditional regression models where variable selection is among the problems considered. It is for instance easy to apply the machinery for direct variable selection in generalised linear models such as Poisson regression. We choose to illustrate the methods here in a situation where the extra variables that might be included reflect quadraticity. More general departures are also handled with similar efforts.

Consider a model framework for independent response variables Y_i in terms of regressors $x_i = (x_{i,1}, x_{i,2})^{\rm t}$ for $i = 1,\ldots,n$. For a known link function g, let the narrow model be that of $\xi(x_i) = g^{-1}(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})$, and consider as wide extension that of

$$\xi(x_i) = g^{-1}\Big(\beta_0 + \sum_{j=1}^2 \beta_j x_{i,j} + \sum_{j=1}^2 \tfrac12\gamma_{j,j}x_{i,j}^2 + \gamma_{1,2}x_{i,1}x_{i,2}\Big).$$

This amounts to anticipating perhaps modest degrees of quadraticity and interaction for covariate influence. Thus there are three extra parameters in the extended model, and at the outset eight different submodels to consider. Context information might reduce this number by disallowing certain subsets, for example the interaction effect. It is now

not difficult to implement the FIC strategy, for each focus parameter of interest. In this example there is a 6×6 information matrix

$$J_{n,\text{full}} = n^{-1} \sum_{i=1}^{n} \xi(x_i) \begin{pmatrix} z_i z_i^{\mathrm{t}} & z_i u_i^{\mathrm{t}} \\ u_i z_i^{\mathrm{t}} & u_i u_i^{\mathrm{t}} \end{pmatrix}$$

which is easily estimated, where $z_i = (1, x_{i,1}, x_{i,2})^{\mathrm{t}}$ contains the covariates which are always present while u_i has the two squared terms and the product term. This leads to all necessary $K_{n,S}$ and $H_{n,S}$ matrices. In the specific example of Poisson regression, the AIC chooses the submodel with largest value of $2\sum_{i=1}^{n}\{Y_i v_i^{\mathrm{t}}\widehat{\alpha} - \exp(v_i^{\mathrm{t}}\widehat{\alpha})\} - 2\#(\alpha)$, writing v_i for any appropriate extension of z_i with some of the u_i elements and α for the corresponding parameter vector, while the FIC has different intentions, which depend on the focus parameter. For $\xi(x) = \exp(\beta_0 + x^{\mathrm{t}}\beta + \frac{1}{2}x^{\mathrm{t}}\gamma x)$, for example, one has $\omega = \xi(x)\{J_{n,10}J_{n,00}^{-1}z - q(x)\}$, where $q(x)$ has two $\frac{1}{2}x_j^2$ elements and one $x_j x_k$ element. This would work as for the logistic regression example, or other examples of generalised linear models, with different submodels being recommended for different areas in the covariate space.

5. Discussion of the FIC

Here is a list of relevant comments pertaining to aspects of the FIC, some of which shed additional light on the AIC. See also Section 7 for further remarks.

5.1. Variance-bias balance. We see from (3.2) and (3.3) that the FIC balances modelling bias versus estimation variability. With a small S the reward is a small variance contribution $2\omega_S^{\mathrm{t}}K_S\omega_S$ but the penalty a bigger $(\widetilde{\psi}_{\text{full}} - \widetilde{\psi}_S)^2$ coming from modelling bias, while the situation is reversed for a bigger S subset. This trade-off game has a particularly transparent structure for the case of diagonal K, when the parameter-focussed information criterion takes the form $\text{FIC} = (\sum_{j\notin S}\omega_j D_j)^2 + 2\sum_{j\in S}\omega_j^2 k_j$, as with (3.4); including more components means more variance and lower bias, and vice versa.

It is important to realise that the FIC is sample-size dependent. From the (2.3) representation of $\widehat{\delta}_{\text{full}}$,

$$\widehat{\text{FIC}} = n\{\widehat{\omega}^{\mathrm{t}}(I - \widehat{K}^{1/2}\widehat{H}_S\widehat{K}^{-1/2})(\widehat{\gamma}_{\text{full}} - \gamma_0)\}^2 + 2\widehat{\omega}_S^{\mathrm{t}}\widehat{K}_S\widehat{\omega}_S.$$

Thus, for $\widehat{\gamma}_{\text{full}}$ being bounded with an increasing n, the first term would eventually dominate, giving large values of FIC, unless S is the full set, where $\widehat{H}_S = I$. In other words, for all large n the data would sensibly select the widest model.

5.2. The one-dimensional case. When there is only one model departure direction to consider, so that $q = 1$, then $\text{FIC} = \omega^2 D^2$ for the narrow estimator while $\text{FIC} = 2\omega^2 K$ for the wide estimator. For the trivial case of a μ parameter with $\omega = 0$, the two estimators are large-sample equivalent. For the more interesting case $\omega \neq 0$, the FIC chooses the

narrow model when $|D|/K^{1/2} < \sqrt{2}$ and the wide model when $|D|/K^{1/2} \geq \sqrt{2}$. In this one-dimensional case this is also equivalent to the AIC, as is seen from (2.5). We see that the estimator used in the end is of the pre-test kind, with $\widehat{\mu}_{\rm full}$ if the test statistic $D^2/K \geq 2$ and $\widehat{\mu}_{\rm narr}$ if $D^2/K < 2$. The significance level indirectly preferred here by the pre-test, by both the AIC and the FIC, is $\Pr\{\chi^2_1 \geq 2\} = 0.157$.

5.3. Model averaging. In cases where two or models score similarly with the FIC, so that there is no clear winner, one may consider compromising between models. This topic of frequentist model averaging is dealt with in Hjort and Claeskens (2003). The theory developed there is actually necessary to understand the behaviour of all inference-post-selection estimators, including those aided by FIC or AIC; see Section 6 below.

5.4. A modified estimator for small D. The basis for the FIC construction is result (3.1) for the limiting risk of arbitrary submodels, coupled with the unbiased estimate $DD^{\rm t} - K$ for the $\delta\delta^{\rm t}$ quantity appearing there. One may also employ alternative estimators, which would lead to modified versions of the focussed model information criterion. Empirical Bayes constructions might be considered, as might the estimator $c_1 DD^{\rm t} - c_2 K$ with c_1 and c_2 chosen to minimise a suitable risk estimate. We do not pursue such alternatives here, as we consider our FIC the canonical version, based on the natural unbiased estimate of risk and having close connections to the AIC. A simple modification worth mentioning, however, is to use 0 to estimate $\delta\delta^{\rm t}$ in the case where $DD^{\rm t} - K$ is negative definite. This is equivalent to $D^{\rm t}K^{-1}D \leq 1$. In such cases the smallest estimated version of (3.1) corresponds to using the narrow model, which is sensible and does not conflict with using (3.2).

5.5. Exact mean squared error in linear regression. Study the linear regression model with $p+q$ covariables, with independent observations Y_i having mean $x_i^{\rm t}\beta + u_i^{\rm t}\gamma$ and the same standard deviation σ for $i = 1, \ldots, n$. For estimating $\mu = x^{\rm t}\beta + u^{\rm t}\gamma$, the mean at some fixed position (x, u), we consider submodel estimators of the form $\widehat{\mu}_S = x^{\rm t}\widehat{\beta}_S + u_S^{\rm t}\widehat{\gamma}_S$, where $u_S = \pi_S u$ and employing least squares estimators in the model which includes $u_{i,j}$ for j in the set S. The structure of this problem is sufficiently clear to allow an exact mean squared error expression to be derived, with some algebra and patience. Also note that we do not need to assume normal distributions. Let

$$\Sigma_n = n^{-1}\sum_{i=1}^n \begin{pmatrix} x_i \\ u_i \end{pmatrix}\begin{pmatrix} x_i \\ u_i \end{pmatrix}^{\rm t} \quad \text{with sub-matrices} \quad \Sigma_{n,S} = n^{-1}\sum_{i=1}^n \begin{pmatrix} x_i \\ u_{i,S} \end{pmatrix}\begin{pmatrix} x_i \\ u_{i,S} \end{pmatrix}^{\rm t}.$$

Also, write Σ_{ij} and $\Sigma_{ij,S}$ for the appropriate blocks of the Σ_n and $\Sigma_{n,S}$ matrices, and Σ^{ij} and $\Sigma^{ij,S}$ for blocks of their inverses, for $i, j = 0, 1$. Finally, partly suppressing 'n' in the notation, let $L = \Sigma^{11}$, $L_S = (\pi_S L^{-1}\pi_S^{\rm t})^{-1}$ and $H_S = L^{-1/2}\pi_S^{\rm t}L_S\pi_S L^{-1/2}$, like in Section 2. We avoid 'J and K' notation here, since we operate without specifying a parametric model, working only with the mean and variance structure. If we in addition postulate normality, we would get a J_n involving $(1/\sigma^2)\Sigma_n$, then $K_n = \sigma^2 L$, and so on, in the notation of previous sections.

The variance matrix of $(\widehat{\beta}_S^{\rm t}, \widehat{\gamma}_S^{\rm t})^{\rm t}$ is $\sigma^2 \Sigma_{n,S}^{-1}$, which implies that n times the variance of $\widehat{\mu}_S$ can be written

$$\sigma^2 \begin{pmatrix} x \\ u_S \end{pmatrix}^{\rm t} \Sigma_{n,S}^{-1} \begin{pmatrix} x \\ u_S \end{pmatrix} = \sigma^2 (x^{\rm t} \Sigma_{00}^{-1} x + V),$$

where the V term after some efforts is seen to be identical to $\omega^{\rm t} L^{1/2} H_S L^{1/2} \omega$, where $\omega = \Sigma_{10} \Sigma_{00}^{-1} x - u$. Thus the variance part matches perfectly that of (3.1). To calculate the bias part, we start from

$$\mathrm{E} \begin{pmatrix} \widehat{\beta}_S \\ \widehat{\gamma}_S \end{pmatrix} = \Sigma_{n,S}^{-1} n^{-1} \sum_{i=1}^n \begin{pmatrix} x_i \\ u_{i,S} \end{pmatrix} (x_i^{\rm t} \beta + u_i^{\rm t} \gamma) = \Sigma_{n,S}^{-1} \begin{pmatrix} \Sigma_{00} \beta + \Sigma_{01} \gamma \\ \Sigma_{10,S} \beta + \pi_S \Sigma_{11} \gamma \end{pmatrix}$$

and use this to derive an exact expression for the mean of $\widehat{\mu}_S$. With some stamina one finds that the bias is $\omega^{\rm t}(I - L^{1/2} H_S L^{-1/2})\gamma$. This again matches result (3.1), and shows that n times the exact mean squared error of $\widehat{\mu}_S$ is

$$\sigma^2 (x^{\rm t} \Sigma_{00}^{-1} x + \omega^{\rm t} L^{1/2} H_S L^{1/2} \omega^{\rm t}) + n \omega^{\rm t} (I - L^{1/2} H_S L^{-1/2}) \gamma \gamma^{\rm t} (I - L^{-1/2} H_S L^{1/2}) \omega,$$

which is an exact match of the general large-sample result (3.1).

It is quite encouraging to see that the general large-sample apparatus we have developed gives recipes which for the case of mean parameters in the linear-normal model give exactly correct results, with no further finite-sample modifications being necessary. Such might be called for in other situations.

5.6. Understanding AIC *from the* FIC *perspective.* The classic AIC model selector, in the transparent context of the limit experiment, is to choose the S for which AIC_S of (2.5) is largest. We shall see that the FIC development and viewpoint are in harmony with AIC for a certain specialisation.

Consider first the estimand $\mu(y) = \log f(y, \theta, \gamma)$. Then $\omega = J_{10} J_{00}^{-1} U(y) - V(y)$, where U and V are the partial derivatives w.r.t. θ and γ, evaluated at the null point (θ_0, γ_0). Thus, by (3.1), the limiting risk of the S-submodel estimator $\log f(y, \widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})$ is

$$\begin{aligned} \mathrm{E} \Lambda_S(y)^2 &= U(y)^{\rm t} J_{00}^{-1} U(y) \\ &\quad + \omega(y)^{\rm t} \{ (I - K^{1/2} H_S K^{-1/2}) \delta \delta^{\rm t} (I - K^{1/2} H_S K^{-1/2})^{\rm t} + K^{1/2} H_S K^{1/2} \} \omega(y). \end{aligned}$$

One may compute the $\mathrm{FIC} = \mathrm{FIC}(y)$ to decide on the best model candidate, for each given y, based on this quality measure. Consider instead the expected quality, when the y comes from $f_0(y) = f(y, \theta_0, \gamma_0)$; in other words, the average mean squared error $\mathrm{risk}_S = \int f_0(y) \mathrm{E} \Lambda_S(y)^2 \, \mathrm{d}y$ associated with submodel S. This is the limit version of $n \int f_{\rm true}(y) \mathrm{mse}(y) \, \mathrm{d}y$, where $\mathrm{mse}(y)$ is the mean squared error of $\log f(y, \widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})$ for estimating $\log f_{\rm true}(y)$. The following is proved in Section 9.

RESULT. *The average mean squared error in the limit experiment can be expressed as*

$$\text{risk}_S = p + \delta^{\mathrm{t}} K^{-1/2}(I - H_S)K^{-1/2}\delta + |S|, \tag{5.1}$$

and an unbiased estimator thereof is $p - q + D^{\mathrm{t}}K^{-1}D - D^{\mathrm{t}}K^{-1/2}H_S K^{-1/2}D + 2|S|$.

The important consequence is that minimising estimated limiting risk is equivalent to the AIC method, as is now clear via (2.5).

6. Performance analysis

We have given arguments for preferring the focussed *johoryo-tokeigaku* to other information criteria, like the AIC, on the grounds of unbiased assessment of limiting risk. Since the proof of the pudding is in the eating, we also ought to investigate the performance on the resulting estimator, taking into account the variability involved in the model selection step.

6.1. Limiting risk. Generally, a large class of selection-estimators are of the form $\widehat{\mu} = \sum \bar{c}(S \mid Z_n)\widehat{\mu}_S$, where $c(S \mid z)$ is an indicator for z falling the region R_S where submodel S is selected. Here $Z_n = \widehat{K}^{-1/2}\widehat{\delta}_{\text{full}}$, see (2.3), with limiting form $Z = K^{-1/2}D \sim \mathrm{N}_q(a, I)$, where $a = K^{-1/2}\delta$. In Hjort and Claeskens (2003) it is shown that a general compromise estimator of the type $\sum_S \bar{c}(S \mid Z_n)\widehat{\mu}_S$ has a limiting distribution Λ for $\sqrt{n}(\widehat{\mu} - \mu_{\text{true}})$ with risk $\mathrm{E}\Lambda^2 = \tau_0^2 + \bar{R}(a)$, where

$$\bar{R}(a) = \mathrm{E}(\omega^{\mathrm{t}}\widehat{\delta} - \omega^{\mathrm{t}}\delta)^2 = \mathrm{E}\{\omega^{\mathrm{t}}K^{1/2}\widehat{a}(Z) - \omega^{\mathrm{t}}K^{1/2}a\}^2,$$

in which $\widehat{a}(Z) = \sum_S \bar{c}(S \mid Z)H_S Z$ is the corresponding estimator of a based on $Z \sim \mathrm{N}_q(a, I)$. Of course $\bar{R}(a)$ can also be represented as a function of $\delta = K^{1/2}a$.

Consider the AIC method first. It chooses $\widehat{S}_{\text{aic}}$ to maximise $Z^{\mathrm{t}}H_S Z - 2|S|$, and the limiting risk of $\widehat{\mu}_{\text{aic}} = \widehat{\mu}_{\widehat{S}_{\text{aic}}}$ takes the form $\tau_0^2 + \bar{R}_{\text{aic}}(a)$, where the latter function can be evaluated via simulation or numerical integration for any K and ω. For the case of K diagonal we may find an explicit formula. Then S is chosen to maximise $\sum_{j\in S}(Z_j^2 - 2)$, and is seen to contain exactly those j for which $|Z_j| \geq \sqrt{2}$. We find

$$\bar{R}_{\text{aic}}(a) = \sum_{j=1}^{q} \omega_j^2 k_j \mathrm{E}(\widehat{a}_j - a_j)^2 + \sum_{j\neq l} \omega_j\omega_l k_j^{1/2}k_l^{1/2}\mathrm{E}(\widehat{a}_j - a_j)(\widehat{a}_l - a_l),$$

where formulae for the moments of $\widehat{a}_j = Z_j I\{|Z_j| \geq \sqrt{2}\}$ are found via the functions $Q_m(a,b) = \int_a^b x^m\phi(x)\,\mathrm{d}x$ for $m = 0, 1, 2$ and their cousins $\bar{Q}_m(a) = Q_m(-\sqrt{2}-a, \sqrt{2}-a)$. In fact, $Q_0(a,b) = \Phi(b) - \Phi(a)$, $Q_1(a,b) = \phi(a) - \phi(b)$ and $Q_2(a,b) = a\phi(a) - b\phi(b) + Q_0(a,b)$. This leads to

$$\mathrm{E}\widehat{a}_j = a_j - \int_{-\sqrt{2}}^{\sqrt{2}} z_j\phi(z_j - a_j)\,\mathrm{d}z_j = a_j - a_j\bar{Q}_0(a_j) - \bar{Q}_1(a_j),$$

$$\mathrm{E}\widehat{a}_j^2 = 1 + a_j^2 - \int_{-\sqrt{2}}^{\sqrt{2}} z_j^2\phi(z_j - a_j)\,\mathrm{d}z_j = 1 + a_j^2 - a_j^2\bar{Q}_0(a_j) - 2a_j\bar{Q}_1(a_j) - \bar{Q}_2(a_j).$$

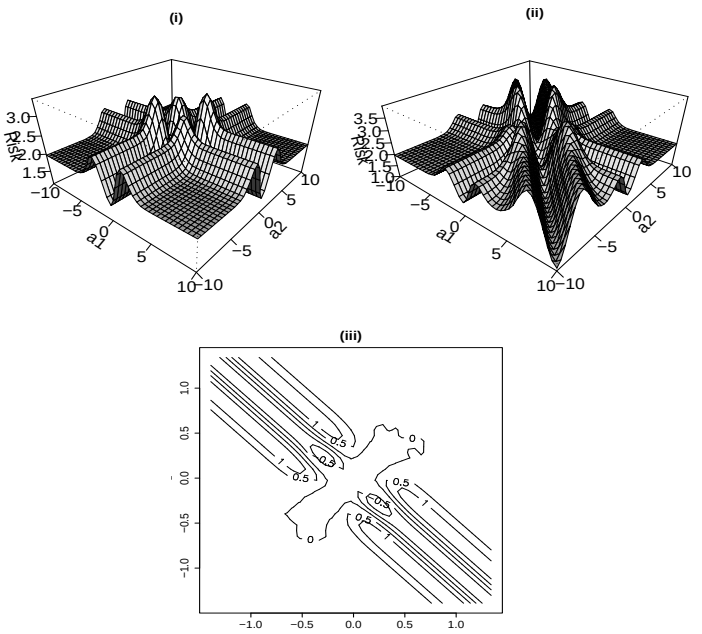

FIGURE 6.1. *Limiting mean squared error risk surface* $\bar{R}(a)$ *using (i) AIC and (ii) FIC for* $q = 2$, $\omega_1 = \omega_2 = 1$ *and* $K = \mathrm{diag}(1,1)$. *(iii) Risk difference* $\bar{R}_{\mathrm{fic}}(a) - \bar{R}_{\mathrm{aic}}(a)$. *The smallest risk for FIC is in the centre stripe and in the two side-lobes.*

In particular, $\bar{R}_{\mathrm{aic}}(0) = \sum_{j=1}^{q} \omega_j^2 k_j \{1 - \bar{Q}_2(0)\}$, which is 0.572 times the constant risk $\omega^{\mathrm{t}} K \omega$ of the full model procedure (corresponding to using $c(S \mid z) = 1$ for S equal to the full set). On the other hand the risk function, while bounded, has maximum value clearly exceeding the $\omega^{\mathrm{t}} K \omega$ level; the factor involved depends on ω and K, but is for example 1.714 when $\omega_1 = \omega_2$ and $K = \mathrm{diag}(k,k)$.

Next consider the FIC, which leads to $\widehat{\mu}_{\mathrm{fic}}$ with $S = \widehat{S}_{\mathrm{fic}}$ chosen to minimise FIC_S. The limiting risk of $\widehat{\mu}_{\mathrm{fic}} = \widehat{\mu}_{\widehat{S}_{\mathrm{fic}}}$ is $\tau_0^2 + \bar{R}_{\mathrm{fic}}(a)$ where the latter can be evaluated from the above, via simulation or numerical integration, using the appropriate version of $\widehat{a}_{\mathrm{fic}}(Z)$. For the diagonal K case, the region R_S where $z \in R_S$ determines that S is chosen, is that where FIC_S is smaller than the others, where $\mathrm{FIC}_S = (\sum_{j \notin S} \omega_j k_j^{1/2} Z_j)^2 + 2 \sum_{j \in S} \omega_j^2 k_j$.

Figure 6.1 depicts the risk surface $\bar{R}_{\mathrm{aic}}(a)$ in panel (i) along with $\bar{R}_{\mathrm{fic}}(a)$ in (ii), for the case of $q = 2$ γ parameters, where all submodels are under consideration. In this setting $\omega_1 = \omega_2 = 1$ and $K = \mathrm{diag}(1,1)$. For AIC the risk values range from 1.116 to 3.421 with the minimum value reached at $a = 0$. The range of $\bar{R}_{\mathrm{fic}}$ values goes from 0.989 to 3.955, also with minimum at the narrow model corresponding to $a = 0$. To facilitate the comparison of these surfaces, Panel (iii) shows contours of the risk difference $\bar{R}_{\mathrm{fic}} - \bar{R}_{\mathrm{aic}}$. FIC has smaller risk than AIC in the centre of the NW-to-SE oriented area, as well as in the smaller side-lobes. AIC has smaller risk in the narrow areas right above and right below the FIC-favourable regions, except for the centre part, see Figure 6.1(iii). Values of the difference range from -1.390 to 1.341.

Qualitatively similar results can be found for $q \geq 3$.

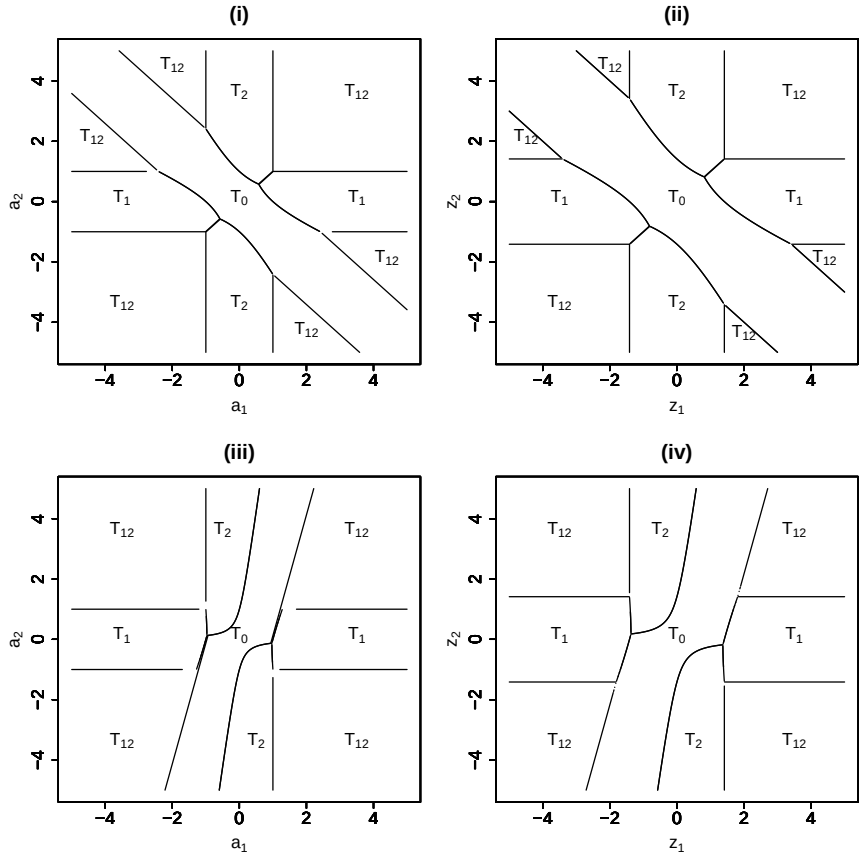

FIGURE 6.2. *For the case of* $K = \mathrm{diag}(1,1)$, *these figures relate to how successful the* FIC *and the* AIC *strategies are for locating the optimal regions* T_0, T_1, T_2, T_{12} *in the parameter space of* $a = (a_1, a_2)$. *Figures (i) and (iii) show these ideal regions for the cases* $\omega = (1,1)^{\mathrm{t}}$ *and* $\omega = (1, -0.25)^{\mathrm{t}}$, *respectively. These are to be compared with deciding regions of* $z = (z_1, z_2)$ *for the* FIC, *shown in figures (ii) and (iv) respectively. The corresponding regions for the* AIC, *not shown, remain the same for each* ω.

6.2. Is the right submodel being chosen? One way to appreciate the difference in perspective and performance between the AIC and the FIC is to study the chances involved of deciding on the 'right' submodel, defined as the one where the accompanying risk function is smallest. For simplicity we limit this brief discussion to the case of a diagonal K. Here the limiting risk for submodel S is a fixed value τ_0^2 plus the quantity $\rho(S,a) = (\sum_{j \notin S} \omega_j k_j^{1/2} a_j)^2 + \sum_{j \in S} k_j \omega_j^2$. This defines optimal or ideal regions $T_S = \{a\colon \rho(S,a) \text{ is smallest}\}$; $T_\emptyset$ is the parameter region where one ideally should have used the narrow model, and so on.

With this perspective, model selection strategies like the AIC and the FIC may be viewed as attempting to come close to the ideal T_S regions. The AIC does this via $\sum_{j \in S}(z_j^2 - 2)$, and in effect uses $\widetilde{T}_S$ equal to the set of z where $|z_j| \geq \sqrt{2}$ for $j \in S$ and $|z_j| < \sqrt{2}$ for $j \notin S$; when Z lands in $\widetilde{T}_S$, estimator $\widehat{\mu}_S$ is being used. The AIC stands by this decision regardless of which estimand is under consideration. The FIC is on the other hand observant to the particular aspects of the μ under foci, via the ω coefficients,

and uses $\widehat{T}_S$ equal to the set of a where $\mathrm{FIC}_S = (\sum_{j\notin S}\omega_j k_j^{1/2}a_j)^2 + 2\sum_{j\in S}\omega_j^2 k_j$ is smaller than all other $\mathrm{FIC}_{S'}$.

These matters are illustrated in Figure 6.2, which relates to two different estimands in the situation where $q = 2$ and $K = \mathrm{diag}(1,1)$. For given estimand, ideal parameter regions T_0, T_1, T_2, T_{12} are found, corresponding to subsets $\emptyset, \{1\}, \{2\}, \{1,2\}$. One sees that the FIC succeeds in emulating these underlying regions, in contrast to the AIC scheme, which uses fixed sets in z-space. This also helps explain why the FIC estimator-post-selection scheme often will work better than the AIC one, as measured also by resulting limiting risk. When studying Figure 6.2, note that the decision regions for the AIC, not shown, remain the same for each ω, and are $\widetilde{T}_0 = \{z\colon |z_j| \le \sqrt{2} \text{ for } j = 1,2\}$, $\widetilde{T}_1 = \{z\colon |z_1| > \sqrt{2}, |z_2| \le \sqrt{2}\}$, $\widetilde{T}_2 = \{z\colon |z_1| \le \sqrt{2}, |z_2| > \sqrt{2}\}$, $\widetilde{T}_{12} = \{z\colon |z_j| > \sqrt{2} \text{ for } j = 1,2\}$.

6.3. Risk comparison in a simulation experiment. We here compare simulated risk values of the FIC to some other model selection practices. Next to the FIC, in the comparison we have used the criteria AIC and BIC, the latter with $|S|\log n$ penalty for model S. Furthermore, we included the adaptive model selection procedure by Shen and Ye (2002). Instead of using the fixed value 2 for AIC, or $\log n$ for BIC, the adaptive selector estimates a penalisation value from the data. Since this approach is computationally intensive, as opposed to the other criteria in our comparison, we investigated a full comparison of the mse results for two situations only.

Data are generated from a linear regression model $Y_i = \theta + \gamma^t u_i + \varepsilon_i$ where for $i = 2, \ldots, n$ the errors ε_i are independent with a standard normal distribution, and are also independent of the 3-dimensional covariate vector $u_i = (u_{i,1}, u_{i,2}, u_{i,3})^{\mathrm{t}}$. These three covariates are also taken independent and standard normal. In the simulation study the intercept value $\theta = 1$, while $\gamma = \delta/\sqrt{n}$ with $\delta = (1,1,1)^{\mathrm{t}}$. The focus parameter for the FIC equals $\mu = \mathrm{E}(Y \mid u)$ where $u = (0.3, -0.1, 0.3)^{\mathrm{t}}$. In the comparison we considered all eight possible models.

For sample size $n = 50$ we obtained the following simulated values for n mse for the criteria FIC, BIC and AIC: 1.407, 1.457, 1.461 and for the adaptive selector the value 1.466. For the bigger sample size $n = 100$, BIC gives a worse performance as compared to the smaller sample size, as now the penalisation constant is bigger. The values ordered from smallest to largest are 1.179, 1.194, 1.205 and 1.258 for FIC, AIC, the adaptive method and BIC, respectively. Similar results are obtained for different focus parameters by specifying different choices of u.

7. Further developments

The article has focussed on the motivation for and development of the FIC, along with its application to a variety of situations and a brief investigation into its actual performance. Below we offer further comments of relevance, some pointing to connections to the AIC and some to competing selection strategies which also emerge from our general framework and results.

7.1. Bayesian and empirical Bayesian model selection. To touch on problems and solutions related to what may be thought of as 'likely' or 'more important' areas of δ values, we choose for clarity of presentation to discuss this inside the framework of the limit experiment, where quantities are known except for δ, for which there is an estimator $D \sim \mathrm{N}_q(\delta, K)$. See the first paragraph of Section 3.1. Methods given and conclusions drawn after working with the limit experiment may then be modified for real finite-sample applications, in the way we went from FIC of Section 3.1 to its estimated version $\widehat{\mathrm{FIC}}$ of Section 3.2.

We have determined that the limiting risk of using submodel S is the fixed amount τ_0^2 plus the quantity

$$\rho(S, \delta\delta^{\mathrm{t}}) = \omega^{\mathrm{t}}\{(I - K^{1/2}H_S K^{-1/2})\delta\delta^{\mathrm{t}}(I - K^{-1/2}H_S K^{1/2}) + K^{1/2}H_S K^{1/2}\}\omega,$$

see (3.1). One wishes to pick the S with smallest value of $\rho(S, \delta\delta^{\mathrm{t}})$. Our FIC solution has been to estimate this unbiasedly, inserting $DD^{\mathrm{t}} - K$ for $\delta\delta^{\mathrm{t}}$, and then pick the minimiser. Other options might involve weighting the risk across the space of δ values, in a suitable fashion, which may or may not come from Bayesian considerations. We outline three such solutions.

(i) One may weight the risk difference w.r.t. a suitable distribution $\mathrm{d}\pi(\delta)$, and then minimise the resulting $\rho(S, \pi) = \int \rho(S, \delta\delta^{\mathrm{t}})\,\mathrm{d}\pi(\delta)$ over submodels S. This is readily done as soon as $\int \delta\delta^{\mathrm{t}}\,\mathrm{d}\pi(\delta)$ is specified. A natural choice is a $\mathrm{N}_q(0, \tau^2 K)$ distribution for δ, which corresponds to an isotropic $\mathrm{N}_q(0, \tau^2 I)$ distribution on the canonical transformed scale $a = K^{-1/2}\delta$, with values closer to the null model more important than values further away. In this case, one chooses the submodel S with smallest value of

$$\rho(S, \delta\delta^{\mathrm{t}}) = \omega^{\mathrm{t}}K^{1/2}\{\tau^2(I - H_S) + H_S\}K^{1/2}\omega.$$

For τ small the dominating term is the second, which hinges on estimation variability, and one chooses the narrow model. For τ large the first term dominates, coming from modelling bias, and one selects the fullest model.

(ii) The above used the 'likely average' of all risk values to decide on S. This sidesteps the perhaps more principled idea that what is at stake are the sizes of $\rho(S, \delta\delta^{\mathrm{t}})$ for the actual δ; it is for the underlying but unknown δ that one wishes to find the best S. This may be placed in a decision-theoretic framework using as loss function

$$L(\delta, S) = \begin{cases} 0 & \text{if } S = \operatorname{argmin} \rho(., \delta\delta^{\mathrm{t}}), \\ 1 & \text{otherwise} \end{cases}$$

to represent the loss involved if choosing S when δ is the true value. One may indeed study performances of model selectors in terms of the expected loss as a function of δ, in other words comparing the probabilities that e.g. the AIC and the FIC select the correct S. Here we show how a natural Bayesian type strategy can be implemented. From a distribution

π for δ, one wishes to construct a model selector $\widehat{S}$ such that $\mathrm{E}L(\delta,\widehat{S})$ is minimised. From Bayesian theory, one should compute $\mathrm{pr}(S)=\Pr\{\mathrm{argmin}\,\rho(.,\delta\delta^{\mathrm{t}})=S\,|\,D\}$ for each S, and select the subset with largest such probability. This is readily done by simulation, when δ can be simulated given D. For the natural $\mathrm{N}_q(0,\tau^2K)$ distribution for δ also used above, one has $\delta\,|\,D\sim\mathrm{N}_q(\kappa D,\kappa K)$, where $\kappa=\tau^2/(\tau^2+1)$. For a large number of simulated δ_j from this distribution, one computes vectors $\{\rho(S,\delta_j\delta_j^{\mathrm{t}})\colon S\ \text{subset}\}$. From these one may read off the required $\mathrm{pr}(S)$ probabilities.

(iii) Instead of the sharp 0-1 loss above, which penalises all non-optimal subsets with the same Draconian sword, one may use

$$\bar{L}(\delta,S)=\rho(S,\delta\delta^{\mathrm{t}})-\min_{S'}\rho(S',\delta\delta^{\mathrm{t}})$$

to better reflect the real loss in risk involved. Again one may, from a start distribution π over the parameter space, determine the optimal strategy, which is to let $\widehat{S}$ minimise $\lambda(S)=\mathrm{E}\{\bar{L}(\delta,S)\,|\,D\}$. This latter quantity may again be evaluated via simulations.

These three solutions rely of course on giving a distribution π for δ, thought either to reflect genuine prior knowledge about which δs are more likely than others, or to correspond to 'degrees of importance' in the parameter space regarding performance of model selectors. One strategy is to use the $\mathrm{N}_q(0,\tau^2K)$ distribution mentioned above, with a value of τ either picked from arguments of plausibility or importance, or from empirical Bayes considerations. The variable $Z^{\mathrm{t}}Z=D^{\mathrm{t}}K^{-1}D$ has mean value $q+a^{\mathrm{t}}a=q+\delta^{\mathrm{t}}K^{-1}\delta$ for given δ, and under the described prior its marginal mean value is $q(1+\tau^2)$. This invites specifying $\widehat{\tau}^2$ as $\max(D^{\mathrm{t}}K^{-1}D/q-1)$; this also corresponds to using maximum likelihood in the marginal model for D. In particular, when $D^{\mathrm{t}}K^{-1}D\le q$, one is content with the narrow model.

model	p(white)		p(black)		p(other)		ratio	
	$\mathrm{pr}(S)$	$\lambda(S)$	$\mathrm{pr}(S)$	$\lambda(S)$	$\mathrm{pr}(S)$	$\lambda(S)$	$\mathrm{pr}(S)$	$\lambda(S)$
0	0.175	0.203	0.193	1.097	0.263	0.092	0.121	63.867
3	0.141	0.153	0.295*	0.872	0.399*	0.087*	0.117	49.205
4	0.170	0.089	0.132	0.561*	0.125	0.225	0.170*	19.919
5	0.058	0.168	0.047	1.437	0.057	0.259	0.096	59.241
34	0.210*	0.075*	0.146	0.561*	0.056	0.208	0.159	19.021*
35	0.078	0.138	0.070	1.166	0.017	0.259	0.161	46.080
45	0.145	0.097	0.101	0.561*	0.072	0.258	0.150	20.253
345	0.024	0.097	0.017	0.561*	0.009	0.258	0.027	20.273

TABLE 7.1. *For submodels 0, 3, 4, 5, 34, 35, 45, 345 the table lists model information probabilities* $\mathrm{pr}(S)$ *and model information scores* $\lambda(S)$*, corresponding to loss functions discussed as (ii) and (iii) above, for the four estimands* p(white)*,* p(black)*,* p(other)*, and the ratio* p(white)$/p$(black)*. 10,000 simulations were used to find the* $\mathrm{pr}(S)$ *and* $\lambda(S)$ *numbers. Asterisks indicate the selected models in question. Results are in quite close agreement with those using the direct* FIC*, see the corresponding Table 4.1, which also lists the ensuing final estimates.*

We carried out such analysis for the low birth weight application studied in Section 4.4 above. There $\widehat{\delta}^{\rm t}\widehat{K}^{-1}\widehat{\delta} = 5.927$, leading to $\widehat{\tau} = 0.9877$. We could therefore simulate 10,000 versions of δ from $\mathrm{N}_3(0.4938\,\widehat{\delta}, 0.4938\widehat{K})$ and estimate the $\mathrm{pr}(S)$ and $\lambda(S)$ quantities needed for (ii)–(iii) above. Note that these again depend on the estimand under focus via the $\widehat{\omega}$ coefficients, but that the same computer programme otherwise may be used for all estimands. Results are given in Table 7.1, which should be studied together with Table 4.1, in that the final parameter estimates are found there after using Table 7.1 to determine the most appropriate submodels. We note that interpreting $\mathrm{pr}(S)$ numbers must be done with some care, in that these sometimes may spread themselves across several equally promising submodels. It may therefore be unwise to focus too quickly on the 'winner' with highest $\mathrm{pr}(S)$ score.

7.2. Good regression models for given covariate regions. We saw in Section 5.6 that the FIC is related to average quality of estimation of log-densities, which again is related to average quality of predictions. This theme is even more important in regression contexts, where the issues also become less clear-cut, in that prediction quality might differ from one covariate region to another.

For the regression framework of Section 2.2, consider the estimand $\mu = \log f(y \,|\, x, \theta, \gamma)$. Then $\omega = J_{10}J_{00}^{-1}U(y \,|\, x) - V(y \,|\, x)$, and the limiting risk of the S-submodel estimator $\log f(y \,|\, x, \widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})$ is

$$\begin{aligned}\mathrm{E}\Lambda_S(x,y)^2 &= U(y \,|\, x)^{\rm t} J_{00}^{-1} U(y \,|\, x) \\ &\quad + \omega(x,y)^{\rm t}\{(I - K^{1/2}H_S K^{-1/2})\delta\delta^{\rm t}(I - K^{-1/2}H_S K^{1/2}) \\ &\quad + K^{1/2}H_S K^{1/2}\}\omega(x,y),\end{aligned}$$

by a parallel to (3.1). It follows that for a fixed x, the average estimation quality $\int f(y \,|\, x)\, \mathrm{E}\Lambda_S(x,y)^2\, \mathrm{d}y$ for estimating the log density is

$$\begin{aligned}\mathrm{risk}_S(x) &= \mathrm{Tr}\{J_{00}^{-1}J_{00}(x)\} \\ &\quad + \mathrm{Tr}\Big[\{(I - K^{1/2}H_S K^{-1/2})\delta\delta^{\rm t}(I - K^{-1/2}H_S K^{1/2}) + K^{1/2}H_S K^{1/2}\} \\ &\qquad \{J_{10}J_{00}^{-1}J_{00}(x)J_{00}^{-1}J_{01} + J_{11}(x) - J_{10}J_{00}^{-1}J_{01}(x) - J_{10}(x)J_{00}^{-1}J_{01}\}\Big].\end{aligned}$$

Furthermore, if this is averaged over some covariate distribution R of interest, to reach a global performance criterion, the result is

$$\begin{aligned}\mathrm{risk}_S &= \int \mathrm{risk}_S(x)\, R(\mathrm{d}x) \\ &= \mathrm{Tr}\{J_{00}^{-1}\bar{J}_{00}\} + \mathrm{Tr}\big[\{(I - K^{1/2}H_S K^{-1/2})\delta\delta^{\rm t}(I - K^{-1/2}H_S K^{1/2}) + K^{1/2}H_S K^{1/2}\} \\ &\qquad \times \{J_{10}J_{00}^{-1}\bar{J}_{00}J_{00}^{-1}J_{01} + \bar{J}_{11} - J_{10}J_{00}^{-1}\bar{J}_{01} - \bar{J}_{10}J_{00}^{-1}J_{01}\}\big],\end{aligned} \tag{7.1}$$

in which $\bar{J}_{00} = \int J_{00}(x)\,R(\mathrm{d}x)$ and so on. For the special case where the $\mathrm{risk}_S(x)$ is averaged w.r.t. the real covariate distribution Q, see Section 2.2, then risk_S, which is also to be thought of as the limit of $n^{-1}\sum_{i=1}^n \mathrm{risk}_S(x_i)$, simplifies to

$$\mathrm{risk}_S = \dim(\theta) + \mathrm{Tr}\{(I - K^{1/2}H_S K^{-1/2})^{\mathrm{t}} K^{-1}(I - K^{1/2}H_S K^{-1/2})\delta\delta^{\mathrm{t}}\} + |S|,$$

just as in Section 5.6 above. Again, this leads upon estimating $\delta\delta^{\mathrm{t}}$ with the unbiased $DD^{\mathrm{t}} - K$ to a criterion asymptotically equivalent to the AIC. Note, however, that for some prediction situations it would be more natural to specify a different R distribution than the full covariate distribution Q; one might wish the best submodel for predicting Y outcomes in a subregion of covariates, for example. This is easily accomplished with an appropriate R distribution, which by insertion of $DD^{\mathrm{t}} - K$ for $\delta\delta^{\mathrm{t}}$ in (7.1) leads to a tailor-made model selection criterion different from the AIC. An interesting special construction, when parameters of the type $\mu(x)$ are considered, would be to employ a gliding window for R around x values of interest; this would lead to a gliding estimate of $\mu(x)$ which for each x involves an appropriate model selection choice.

7.3. Minimising expected Kullback–Leibler distance. Arguably, a selected submodel S is good if the distance from the true density to the estimated density $f(y, \widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})$ is small. A sensible exercise is to attempt to select the submodel with the smallest Kullback–Leibler distance

$$\mathrm{KL}_{n,S} = \int f_{\mathrm{true}}(y) \log\{f_{\mathrm{true}}(y)/f(y, \widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c})\}\,\mathrm{d}y,$$

where f_{true} is as in (2.1). The following is proved in Section 9. The consequence, following the Result of Section 5.6, is that minimising estimated Kullback–Leibler distance is yet another strategy which becomes equivalent to the AIC scheme.

RESULT. *Under standard regularity conditions,* $2n\,\mathrm{KL}_{n,S} \to_d \mathrm{KL}_S$, *a variable with mean value equal to* risk_S *of (5.1).*

7.4. Minimising expected weighted ISE. Consider the weighted integrated squared error quantity $\mathrm{ISE}_{n,S} = \int\{f(y, \widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c}) - f_{\mathrm{true}}(y)\}^2/f_{\mathrm{true}}(y)\,\mathrm{d}y$. Efforts similar to those exuded in Section 9.2, to prove the result of Section 7.3, lead to

$$n\,\mathrm{ISE}_{n,S} \to_d \mathrm{ISE}_S = \int f(y, \theta_0, \gamma_0)\Big\{U(y)^{\mathrm{t}} C_S + V(y)^{\mathrm{t}}\Big(\begin{pmatrix} D_S \\ 0 \end{pmatrix} - \delta\Big)\Big\}^2\,\mathrm{d}y,$$

with (C_S, D_S) the limit in distribution of $\sqrt{n}(\widehat{\theta}_S - \theta_0, \widehat{\gamma}_S - \gamma_{0,S})$, see Hjort and Claeskens (2003, Section 3). But from the proof in 9.2 this is seen to be the same variable as KL_S, with the same consequences for model choice.

7.5. Non-likelihood estimators and Cox regression. Our story or stories have been told for the case of maximum likelihood estimators when comparing different submodels.

In some situations there are reasons for choosing alternative ingredients when forming say $\widetilde{\mu}_S = \mu(\widetilde{\theta}_S, \widetilde{\gamma}_S)$, with perhaps more robust estimators for the parameters. One may generalise our results to the case of robust M-estimators, and to the minimum density divergence estimators of the Basu, Harris, Hjort and Jones (1998, 2001) variety. Results will generally be less elegant and less concise than for maximum likelihood methods, and will involve additional matrices and linear algebra. Our methods may also be generalised to the semiparametric Cox regression model, with added efforts. This will be reported on in forthcoming work.

7.6. Finite-sample corrections. Our theory has been developed exploiting the first-order asymptotics properties of maximum likelihood estimators, leading to a precise description of the limit experiment and so on. For some classes of models there might be a need for fine-tuning the FIC via appropriate sample-size dependent corrections. For this one might draw on work pertaining to the AIC by Hurvich and Tsai (1989, 1995), McQuarrie and Tsai (1998), Burnham and Anderson (2002) and others.

7.7. Generalised ridging when q is big. Our framework has been the classic one for large-sample likelihood analysis, where the number of data points grows and the number of parameters, at most $p+q$, stays bounded. It is more challenging to develop safe methods for model comparison and e.g. regressor subset selection when either p or q is allowed to become bigger with n. Some model choice methods are specifically constructed to do well in such situations, like Breiman's (1992) little bootstrap; see further references in his paper and the recent paper of Efron, Hastie, Johnstone and Tibshirani (2003). A general idea for coping with non-small q is to shrink estimators of the γ part towards the γ_0 position. Several of the methods and results of this article may actually be generalised to encompass such shrinking type estimators; see Hjort and Claeskens (2003, Sections 8–9).

8. The focussed robust information criterion

Our model compromise and model selection apparatus has been built under the key operating assumption (2.1), which in particular demands that the full $p+q$-parameter model is correct, for a γ parameter not too far from γ_0. It appears important to investigate what might happen if this assumption does not hold. Assume, therefore, that the true data generating mechanism takes the form

$$f_{\rm true}(y) = f(y, \theta_0, \gamma_0)\{1 + r(y)/\sqrt{n}\} + o(1/\sqrt{n}) \tag{8.1}$$

for a suitable $r(y)$ function, with $\int f_0|r|\,{\rm d}y$ finite and $\int f_0 r\,{\rm d}y = 0$, where $f_0(y) = f(y, \theta_0, \gamma_0)$. Results derived earlier in our article have used (2.1), which corresponds to the special case $r(y) = V(y)^{\rm t}\delta$, with $V(y) = \partial \log f(y, \theta_0, \gamma_0)/\partial\gamma$, cf. Hjort and Claeskens (2003, Section 3). In this framework there are no 'true parameters' (θ, γ). Instead we consider the least false parameter $\mu_{\rm lf} = \mu(\theta_n, \gamma_n)$, where (θ_n, γ_n) are the least false parameters inside the $f(y, \theta, \gamma)$ family, defined as those minimising the Kullback–Leibler

distance $\int f_{\rm true}(y)\log\{f_{\rm true}(y)/f(y,\theta,\gamma)\}\,{\rm d}y$. Those are the parameter values aimed at by the maximum likelihood estimators in the full model.

Some analysis shows that, apart from remainder terms of size $o(1/\sqrt{n})$, $\theta_n = \theta_0 + \eta_0/\sqrt{n}$ and $\gamma_n = \gamma_0 + \delta_0/\sqrt{n}$, where

$$\begin{pmatrix}\eta_0\\ \delta_0\end{pmatrix} = J_{\rm full}^{-1}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = J_{\rm full}^{-1}\begin{pmatrix}\int f_0 U r\,{\rm d}y\\ \int f_0 V r\,{\rm d}y\end{pmatrix},$$

with notation as in Section 3 of Hjort and Claeskens (2003). The case (2.1) corresponds to $\eta_0 = 0$ and $\delta_0 = \delta$. We may apply and generalise arguments used to reach Lemmas 3.1–3.3 in Hjort and Claeskens (2003) to show that

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm lf}) \to_d \widetilde{\Lambda}_S = (\tfrac{\partial\mu}{\partial\theta})^{\rm t} C_S + (\tfrac{\partial\mu}{\partial\gamma_S})^{\rm t} D_S - (\tfrac{\partial\mu}{\partial\theta})^{\rm t}\eta_0 - (\tfrac{\partial\mu}{\partial\gamma})^{\rm t}\delta_0, \tag{8.2}$$

where

$$\begin{pmatrix}\sqrt{n}(\widehat{\theta}_S - \theta_0)\\ \sqrt{n}(\widehat{\gamma}_S - \gamma_{0,S})\end{pmatrix} \to_d \begin{pmatrix}C_S\\ D_S\end{pmatrix} = J_S^{-1}\begin{pmatrix}\alpha + M\\ \beta + N_S\end{pmatrix}.$$

The vector with components M and N is as in Lemma 3.1 of Hjort and Claeskens (2003), basically since the covariance matrix of $(U(Y_i), V(Y_i))$ converges to $J_{\rm full}$ also under (8.1) circumstances. With $W = K(N - J_{10}J_{00}^{-1}M)$ again, algebraic efforts lead to

$$\begin{aligned} C_S &= J_{00}^{-1}M - J_{00}^{-1}J_{01}\pi_S^{\rm t}K_S\pi_S K^{-1}W + \eta_0 + J_{00}^{-1}J_{01}(I - \pi_S^{\rm t}K_S\pi_S K^{-1})\delta_0,\\ D_S &= K_S\pi_S K^{-1}W + K_S\pi_S K^{-1}\delta_0. \end{aligned}$$

All this combines to yield

$$\widetilde{\Lambda}_S = (\tfrac{\partial\mu}{\partial\theta})^{\rm t}J_{00}^{-1}M + \omega^{\rm t}\{\delta_0 - K^{1/2}H_S K^{-1/2}(\delta_0 + W)\}. \tag{8.3}$$

This is, amazingly, very nearly the same result as in Lemma 3.3 of Hjort and Claeskens (2003). The point is that η_0 drops out and that the least false related parameter δ_0 takes the place of our earlier δ. We also have $D_n = \widehat{\delta}_{\rm full} \to_d D = \delta_0 + W \sim {\rm N}_q(\delta_0, K)$, in generalisation of (2.3). The consequence is that also the theory of compromise and post-selection estimators goes through, with methods of Sections 3 and 4 still applicable. The necessary modification is only that in (8.2), precision of $\widehat{\mu}_S$ is assessed and interpreted in terms of nearness to the least false $\mu_{\rm lf}$ rather than to the 'true' focus parameter.

The above may be generalised one step further. Assume, instead of (8.1), that

$$f_{\rm true}(y) = g(y)\{1 + r(y)/\sqrt{n}\} + o(1/\sqrt{n}), \tag{8.4}$$

where g is simply a fixed density, not necessarily belonging to the $(p+q)$-parametric family. We need to assume that g has the property that if $f(y,\theta',\gamma')$ is the least false parametric approximation to $f_{\rm true}$, with the parameters selected to minimise the Kullback–Leibler distance from $f_{\rm true}$ to the parametric approximation, then $\gamma' = \gamma_0$. Without such an

assumption, there will not be convergence in distribution of $\sqrt{n}(\widehat{\gamma}_{\rm full}-\gamma_0)$, for example, and no fruitful local asymptotics theory can be worked out; indeed modelling bias will then dominate, for all large n, making the wide model the winner in the end. But under the assumption mentioned, one has not only $\int gU\,{\rm d}y=0$ but also $\int gV\,{\rm d}y=0$, with consequent generalisations of Lemmas 3.1–3.3 of Hjort and Claeskens (2003). We find that algebraic results associated with (8.2) and (8.3) continue to hold, involving the least false $\mu_{\rm lf}=\mu(\theta_n,\gamma_n)$, with one crucial difference. Let J be minus the matrix of expected second order derivatives of the log density, evaluated at density g, and let Ω be the variance matrix of $(U(Y),V(Y))$, also evaluated at g. Under earlier assumptions, these two matrices have been equal. Both may be estimated consistently from data. Under present conditions, (M,N) have covariance matrix Ω rather than J. In particular, M and W worked with above are not necessarily independent now.

Consider the limiting risk $r(S)={\rm E}\widetilde{\Lambda}_S^2$ under the present agnostic circumstances. It consists of a variance term $r_v(S)$ and a squared bias term $r_b(S)$. The first involves the variance of W and its covariance with M, and is found to be

$$r_v(S)=\tau_0^2+\omega^{\rm t}K^{1/2}H_SG_1H_SK^{1/2}\omega-2(\tfrac{\partial\mu}{\partial\theta})^{\rm t}J_{00}^{-1}G_2K^{1/2}H_SK^{1/2}\omega,$$

in terms of

$$\begin{aligned}G_1&=K^{1/2}(\Omega_{11}-\Omega_{10}J_{00}^{-1}J_{01}-J_{10}J_{00}^{-1}\Omega_{01}+J_{10}J_{00}^{-1}\Omega_{00}J_{00}^{-1}J_{01})K^{1/2},\\ G_2&=\Omega_{01}-\Omega_{00}J_{00}^{-1}J_{01}.\end{aligned}$$

When $J=\Omega$, as under (8.1) conditions, $G_1=I$ and $G_2=0$. The second contribution is $r_b(S)=\{\omega^{\rm t}(I-K^{1/2}H_SK^{-1/2})\delta_0\}^2$, which we estimate inserting $DD^{\rm t}-K$ for $\delta_0\delta_0^{\rm t}$, since $D_n\to_d D\sim{\rm N}_q(\delta_0,K)$ even under (8.4). This yields an unbiased estimator of limiting risk, in the limit experiment. A little work leads to

$$\widehat{r}(S)=r_v(S)+\widehat{r}_b(S)=r_v(S)+(\widetilde{\psi}_{\rm full}-\widetilde{\psi}_S)^2-\omega^{\rm t}(K-K^{1/2}H_SK^{1/2})\omega,$$

which is a constant away from the focussed model-robust information criterion

$${\rm FRIC}=(\widetilde{\psi}_{\rm full}-\widetilde{\psi}_S)^2+\omega^{\rm t}K^{1/2}H_S(G_1+I)H_SK^{1/2}\omega-2(\tfrac{\partial\mu}{\partial\theta})^{\rm t}J_{00}^{-1}G_2K^{1/2}H_SK^{1/2}\omega.$$

This generalises FIC of Section 3.1. For real data, estimation of the necessary matrices can be carried out via natural empirical versions of J and Ω. Inserting also estimates of partial derivatives, along with using $\widehat{\psi}=\widehat{\omega}^{\rm t}D_n$ and $\widehat{\psi}_S=\widehat{\omega}^{\rm t}\widehat{K}^{1/2}\widehat{H}_S\widehat{K}^{-1/2}D_n$, leads to a natural $\widehat{\rm FRIC}$, in generalisation of Section 3.2.

9. Proofs of two results

9.1. Proof of Section 5.6's Result. We find that $\omega(Y)$ has covariance matrix K^{-1}, and are led to

$$\begin{aligned}\text{risk}_S &= \text{Tr}\Big(J_{00}^{-1}\int f_0 UU^{\rm t}\,{\rm d}y\Big)\\ &\quad + \text{Tr}\Big[\{(I-K^{1/2}H_SK^{-1/2})\delta\delta^{\rm t}(I-K^{-1/2}H_SK^{1/2})+K^{1/2}H_SK^{1/2}\}\int f_0\omega\omega^{\rm t}\,{\rm d}y\Big]\\ &= p+\delta^{\rm t}(I-K^{-1/2}H_SK^{1/2})K^{-1}(I-K^{1/2}H_SK^{-1/2})\delta+m_S\\ &= p+\delta^{\rm t}K^{-1/2}(I-H_S)K^{-1/2}\delta+m_S,\end{aligned}$$

where $m_S=\text{Tr}(H_S)$. We have used the projection matrix property $H_S^2=H_S$ here. Now estimate the above quantity by inserting the unbiased $DD^{\rm t}-K$ for $\delta\delta^{\rm t}$, as with the development that led to FIC of (3.2). This leads after further algebra to

$$\begin{aligned}\widehat{\text{risk}}_S &= p+\text{Tr}\{K^{-1/2}(I-H_S)K^{-1/2}(DD^{\rm t}-K)\}+m_S\\ &= p+D^{\rm t}K^{-1/2}(I-H_S)K^{-1/2}D+m_S-q+m_S\\ &= p-q+D^{\rm t}K^{-1}D-D^{\rm t}K^{-1/2}H_SK^{-1/2}D+2m_S.\end{aligned}$$

It remains to show that $m_S=|S|$. Let for the convenience of presentation S be the first $|S|$ indexes of $\{1,\ldots,q\}$ and the complement set S^c the $q-|S|$ last ones. Then, upon decomposing K^{-1} into blocks K^{ij}, we find first that K_S defined before (2.2) is equal to $(K^{00})^{-1}$, and next that

$$H_S=K^{-1/2}\begin{pmatrix}(K^{00})^{-1} & 0\\ 0 & 0\end{pmatrix}K^{-1/2}$$

must have trace equal to $|S|$, as claimed. ■

9.2. Proof of Section 7.2's Result. By Taylor expanding $\log f(y,\widehat\theta_S,\widehat\gamma_S,\gamma_{0,S^c})-\log f(y,\theta_0,\gamma_0+\delta/\sqrt n)$, one finds that $\text{KL}_{n,S}$ can be expressed as

$$\begin{aligned}&U(y,\theta_0,\gamma_0+\delta/\sqrt n)^{\rm t}(\widehat\theta_S-\theta_0)+V(y,\theta_0,\gamma_0+\delta/\sqrt n)^{\rm t}\begin{pmatrix}\widehat\gamma_S-\gamma_{0,S}-\delta_S/\sqrt n\\ \gamma_{0,S^c}-\gamma_{0,S^c}-\delta_{S^c}/\sqrt n\end{pmatrix}\\ &\quad+\tfrac12\begin{pmatrix}\widehat\theta_S-\theta_0\\ \widehat\gamma_S-\gamma_{0,S}-\delta_S/\sqrt n\\ -\delta_{S^c}/\sqrt n\end{pmatrix}^{\rm t}W(y,\theta_n',\gamma_n')\begin{pmatrix}\widehat\theta_S-\theta_0\\ \widehat\gamma_S-\gamma_{0,S}-\delta_S/\sqrt n\\ -\delta_{S^c}/\sqrt n\end{pmatrix},\end{aligned}$$

where U and V are the partial derivatives of $\log f(y,\theta,\gamma)$ w.r.t. θ and γ, where $W(y,\theta_n',\gamma_n')$ is the $(p+q)\times(p+q)$ matrix of second order partial derivatives of the log-density, evaluated at a point sandwiched between $(\theta_0,\gamma_0+\delta/\sqrt n)$ and $(\widehat\theta_S,\widehat\gamma_S,\gamma_{0,S^c})$. Under suitable regularity conditions, $\int f_{\rm true}(y)W(y,\theta_n',\gamma_n')\,{\rm d}y\to_p -J_{\rm full}$. Next note that

$$\sqrt n\begin{pmatrix}\widehat\theta_S-\theta_0\\ \widehat\gamma_S-\gamma_{0,S}-\delta_S/\sqrt n\\ -\delta_{S^c}/\sqrt n\end{pmatrix}\to_d\begin{pmatrix}C_S\\ D_S-\delta_S\\ -\delta_{S^c}\end{pmatrix}$$

when n travels to infinity, using notation and results of Hjort and Claeskens (2003, Section 3.2). It follows that

$$2n\,\mathrm{KL}_{n,S} \to_d \mathrm{KL}_S = \left(\begin{pmatrix} C_S \\ \begin{pmatrix} D_S \\ 0 \end{pmatrix} - \delta \end{pmatrix}^{\mathrm{t}} J_{\mathrm{full}} \left(\begin{pmatrix} C_S \\ \begin{pmatrix} D_S \\ 0 \end{pmatrix} - \delta \end{pmatrix} \right.$$

To evaluate the mean of the limit distribution, introduce first $P_S = K^{1/2} H_S K^{-1/2}$. From the proof of Lemma 3.3 in Hjort and Claeskens (2003), one finds $\mathrm{E}C_S = J_{00}^{-1} J_{01}(I - P_S)\delta$ and

$$\mathrm{E}\Big(\begin{pmatrix} D_S \\ 0 \end{pmatrix} - \delta \Big) = \begin{pmatrix} K_S \pi_S K^{-1} \delta \\ 0 \end{pmatrix} - \delta = -(I - P_S')\delta, \quad \text{with } P_S' = \begin{pmatrix} K_S \pi_S K^{-1} \\ 0 \end{pmatrix},$$

while the variance matrix of (C_S, D_S) is J_S^{-1}. Following results at the end of 9.1 above one sees that $P_S' = P_S$. The mean of KL_S is the sum of a variance contribution

$$\mathrm{Tr}\Big\{ J_{\mathrm{full}}\, \mathrm{Var} \begin{pmatrix} C_S \\ D_S \\ 0 \end{pmatrix} \Big\} = p + |S|$$

and a squared bias contribution

$$\begin{aligned}
&\begin{pmatrix} J_{00}^{-1} J_{01}(I - P_S)\delta \\ -(I - P_S)\delta \end{pmatrix}^{\mathrm{t}} J_{\mathrm{full}} \begin{pmatrix} J_{00}^{-1} J_{01}(I - P_S)\delta \\ -(I - P_S)\delta \end{pmatrix} \\
&\quad = \delta^{\mathrm{t}} [(I - P_S)^{\mathrm{t}} J_{10} J_{00}^{-1} J_{01}(I - P_S) + (I - P_S)^{\mathrm{t}} J_{11}(I - P_S) \\
&\qquad\quad - (I - P_S)^{\mathrm{t}} J_{10} J_{00}^{-1} J_{01}(I - P_S) - (I - P_S)^{\mathrm{t}} J_{10} J_{00}^{-1} J_{01}(I - P_S)]\delta \\
&\quad = \delta^{\mathrm{t}} (I - P_S)^{\mathrm{t}} K^{-1} (I - P_S)\delta \\
&\quad = \delta^{\mathrm{t}} K^{-1/2} (I - H_S) K^{-1/2} \delta,
\end{aligned}$$

where we have used $J_{11} - J_{10} J_{00}^{-1} J_{01} = K^{-1}$. ■

Acknowledgements

The authors wish to thank the Editor, Associate Editor and referees for their careful reading and constructive comments, and Editor Frank Samaniego in particular for his encouragement and attention. The research of Claeskens is partly supported by NSF Grant DMS-02-03884.

References

Aerts, M., Claeskens, G. and Hart, J.D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association* **94**, 869–879.

Aerts, M., Claeskens, G. and Hart, J.D. (2000). Testing lack of fit in multiple regression. *Biometrika* **87**, 405–424.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (ed. B. Petrov and F. Csáki), 267–281. Akadémiai Kiadó, Budapest.

Allen, D.M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics* **13**, 469–475.

Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

Breiman, L. (2001). Statistical modeling: The two cultures [with discussion]. *Statistical Science* **16**, 199–231.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer, New York.

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). Goodness of fit via nonparametric likelihood ratios. Submitted for publication.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2003). Least angle regression. To appear.

Eubank, R.K. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* **20**, 1412–1425.

George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–727.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā, Series A* **51**, 45–58.

Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, to appear[!].

Hosmer, D.W. and Hjort, N.L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine* **21**, 2723–2738.

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. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51**, 1077–1084.

Jones, M.C., Hjort, N.L., Harris, I.R. Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**, 865–873.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association* **89**, 1000–1005.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley, New York.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

McQuarrie, A.D.R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection.* World Scientific Publishing, Singapore.

Murata, N., Yoshizawa, S. and Amara, S. (1994). Network information criterion – determining the number of hidden units for artificial natural network models. *IEEE Transactions on Neural Networks* **5**, 865–872.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* **12**, 758–765.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* Series in Computer Science **15**, World Scientific, Singapore.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection [with discussion]. *Statistica Sinica* **7**, 221–264.

Shen, X. and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association* **97**, 210–221.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–126.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45–53.

Simonoff, J. and Tsai, C.-L. (1999). Semiparametric and additive model selection using an improved AIC criterion. *Journal of Computational and Graphical Statistics* **8**, 22–40.

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.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions [with discussion]. *Journal of the Royal Statistical Society* **B 36**, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by dross-validation and Akaike's criterion. *Journal of the Royal Statistical Society* **B 39**, 44–47.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (Mathematical Sciences) **153**, 12–18. [In Japanese.]

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society* **B 61**, 529–546.

Wei, C.Z. (1992). On predictive least squares principles. *Annals of Statistics* **20**, 1–42.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93**, 120–131.