KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Minimising average risk in regression models

Gerda Claeskens and Nils Lid Hjort

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0610

Minimising Average Risk in Regression Models

Gerda Claeskens
ORSTAT and University Centre for Statistics
Katholieke Universiteit Leuven
Naamsestraat 69
B-3000 Leuven, Belgium
Gerda.Claeskens@econ.kuleuven.be

Nils Lid Hjort
Department of Mathematics
University of Oslo
N-0316 Oslo, Norway
nils@math.uio.no

December 2005

Abstract

Most model selection mechanisms work in an 'overall' modus, providing models without specific concern for how the selected model is going to be used afterwards. The focussed information criterion (FIC), on the other hand, is geared towards optimum model selection when inference is required for a given estimand. In this paper the FIC method is extended to weighted versions. This allows one to rank and select candidate models for the purpose of handling a range of similar tasks well, as opposed to being forced to focus on each task separately. Applications include selecting regression models that perform well for specified regions of covariate values. We derive these wFIC criteria, give asymptotic results, and apply the methods to real data. Formulae for easy implementation are provided for the class of generalised linear models.

Key-words: focussed information criterion, model selection, regression models.

1 Introduction

Model selection is most often the starting point of any data analysis. It is therefore of importance to carefully address this modelling step. Most model selection tools attach to each potential model a number, and then proceed by picking the model with the best (usually either smallest or largest) value of this number. Traditional model selection techniques work this way, such as Mallows's C_p (1973) for linear regression or the more generally applicable information criteria AIC (from Akaike, 1974) and BIC (from Schwarz, 1978). Once a model is selected, the actual estimation takes place in the selected model.

The focussed information criterion (Claeskens and Hjort, 2003) is also in this spirit, though distinguishes itself from the other information criteria in that it can be directed – focussed – towards a specific purpose. A model selected to estimate, for example, the mean return of an investment taken place in one month, should not necessarily be the same as the model to be used to estimate the mean return one year further in time. Or, in medical studies, the model which is best for estimating the survival probability should not be expected to be the same as the best model for estimating the median survival time. In the construction of the FIC we therefore start by specifying the focus parameter, the quantity we wish to estimate, and then use this information to obtain the actual FIC value via an estimator of the mean squared error of the focus parameter's estimator. When the focus parameter changes, also the value of the FIC might change, leading to a possibly different selected model.

The main issue addressed in this paper is that in many situations, the focussed information criterion demands too much focus of its user, so to speak. For example, for regression models the FIC can easily be used to select a model for the mean response value for a given single covariate position. This is sometimes of relevance, and one may take a 'median' or 'average' covariate position; but in other situations one wishes to construct a model that does well across many covariate positions.

To address this problem, we derive a weighted focussed information criterion, where we attach to each potential model a single wFIC value, valid over all or part of the covariate space. To explain matters clearly, we mainly concentrate on the class of generalised linear models, though our arguments and methods would extend without serious difficulties to more general regression models. More information on generalised linear models can be found in the books by McCullagh and Nelder (1989) and Dobson (2002).

Section 2 briefly reviews the FIC method, in a setting of generalised linear models. Working with weighted versions of mean squared errors, across all or a subset of covariates, leads in Section 3 to the weighted FIC method. The wFIC method relies on weights that are user- and context-specific. Applying a specific weighting scheme, with what we term glm weights, gives a procedure that turns out to be large-sample equivalent to the AIC. This is discussed in Section 4, along with some other weighting schemes of interest. Strategies for estimating (and then minimising) more generally formed averaged risks are then taken up in Section 5. Practical data examples, showing the applicability of the method, appear in Section 6. Section 7 offers some concluding comments, indicating the further scope of our work.

2 The FIC for generalised linear models

Let the observed data be denoted $Y_1, \ldots, Y_n$, together with observed covariate information $c_1, \ldots, c_n$. In a model selection situation the c_is are vector valued and we wish to build methods that somehow manage to select the components of most relevance. Often, we are sure about including some of these covariates in the model (an intercept is an example of this), so that the actual selection should take place over the other variables. Notationally we split the covariate vector c_i into two parts, the 'protected' x_i with say p covariates that are deemed necessary a priori and the 'open' z_i containing the say q remaining covariates, amongst which we intend to perform the selection.

2.1 The GLM class

For a generalised linear model (glm), there is a monotone differentiable link function $g(\cdot)$ such that

$$g(\mathrm{E}(Y_i \,|\, x_i, z_i)) = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma \quad \text{for } i = 1, \ldots, n, \tag{1}$$

mapping the mean response to a linear predictor defined in terms of regression coefficients $\beta_1, \ldots, \beta_p$ and $\gamma_1, \ldots, \gamma_q$. Thus selecting which covariate components $z_{i,j}$ to include amounts to determining which γ_js to keep in the model; excluding component $z_{i,j}$ corresponds to using $\gamma_j = 0$. For (classical) linear models, the link function is the identity function. For other examples, see Section 2.3.

We assume that $Y_1, \ldots, Y_n$ are independent with density function belonging to an exponential family of the form

$$f(y_i; \theta_i, \phi) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big\} \quad \text{for } y \in \mathcal{Y},$$

where the sample space $\mathcal{Y}$ is the same in each case and does not depend on the unknown parameters θ_i and ϕ. The functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are fully specified. The $b(\cdot)$ function plays a crucial role since its derivatives yield the mean and variance function, while $a(\phi)$ is a scale parameter. From the two first Bartlett identities, about moment properties of the first and second log-derivatives of the density, follow

$$\xi_i = \mathrm{E}(Y_i \,|\, x_i, z_i) = b'(\theta_i) \quad \text{and} \quad \mathrm{Var}(Y_i \,|\, x_i, z_i) = a(\phi)b''(\theta_i).$$

This expresses θ_i as a function of the mean response, given the covariates. When the so-called canonical link function is used, that is, $g(\cdot) = (b')^{-1}(\cdot)$, then $\theta_i = g(\xi_i) = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$.

2.2 The pointwise FIC for GLM

Here we derive explicit expressions for the FIC when applied to the linear predictor $\mu = x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma = g(\mathrm{E}(Y \,|\, x, z))$ associated with a given position (x, z) in the covariate space. The (pointwise) FIC is large-sample equivariant under smooth transformations, so the FIC for $\mu(x, z)$ will essentially yield the same model ranking as the FIC for $\xi(x, z) = \mathrm{E}(Y \,|\, x, z)$.

One of the main ingredients for computing the FIC is the (normalised) Fisher information matrix $J_{n,\mathrm{wide}}$, computed via the second order partial derivatives of the log-likelihood function with respect to (β, γ). We partition the matrix $J_{n,\mathrm{wide}}$ as

$$J_{n,\mathrm{wide}} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix},$$

such that the block $J_{n,11}$ in the lower right corner has dimension $q \times q$. In our calculations we will frequently need the the lower right submatrix of dimension $q \times q$ of $J_{n,\mathrm{wide}}^{-1}$, which we name K_n. This matrix may be found as $K_n = (J_{n,11} - J_{n,10}J_{n,00}^{-1}J_{n,01})^{-1}$.

For several members of the generalised linear model family, the scale parameter ϕ is either known or completely specified. Examples are the Poisson and binomial distribution, where $a(\phi) = 1$, or the normal distribution with known variance. In case ϕ is known, the information matrix takes a simple and general form, making it easy to compute in the full generality of generalised linear models. We first decompose the $n \times (p + q)$ design matrix into $X = (x_1^{\mathrm{t}}, \ldots, x_n^{\mathrm{t}})^{\mathrm{t}}$ of dimension $n \times p$ and $Z = (z_1^{\mathrm{t}}, \ldots, z_n^{\mathrm{t}})^{\mathrm{t}}$ of dimension $n \times q$. Then

$$J_{n,\mathrm{wide}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a(\phi)} v_i \begin{pmatrix} x_i \\ z_i \end{pmatrix} \begin{pmatrix} x_i \\ z_i \end{pmatrix}^{\mathrm{t}} = \frac{1}{n}\frac{1}{a(\phi)} \begin{pmatrix} X^{\mathrm{t}}VX & X^{\mathrm{t}}VZ \\ Z^{\mathrm{t}}VX & Z^{\mathrm{t}}VZ \end{pmatrix},$$

where V is the diagonal weight matrix $\mathrm{diag}(v_1, \ldots, v_n)$, with different formulae available for $v_i = v(x_i, z_i)$. One has

$$v_i = [b''(\theta_i)\{g'(\xi_i)\}^2]^{-1} = \frac{b''(\theta_i)}{a(\phi)}\Big(\frac{\partial\theta_i}{\partial\eta_i}\Big)^2, \tag{2}$$

where $\eta_i = g^{-1}(\xi_i) = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$ and $\xi_i = \mathrm{E}(Y_i \,|\, x_i, z_i) = b'(\theta_i)$. For the situation with ϕ known, therefore, the K_n matrix takes the form

$$K_n = a(\phi)\{n^{-1}Z^{\mathrm{t}}V(I - X(X^{\mathrm{t}}VX)^{-1}X^{\mathrm{t}}V)Z\}^{-1}. \tag{3}$$

For canonical link functions, where $\theta_i = \eta_i$, matters simplify to $v_i = b''(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)/a(\phi)$.

When the scale parameter ϕ is not known, such as with the normal distribution with unknown mean and variance, or the gamma distribution, we shall argue that formula (3) is still valid for K_n. This is because of an orthogonality property, namely that the mixed second derivatives of the log-likelihood function, with respect to ϕ and β or γ, are seen to

have mean zero. Thus the full information matrix takes the form

$$J_{n,\text{wide}} = \begin{pmatrix} J_{n,\text{scale}} & 0 & 0 \\ 0 & n^{-1}a(\phi)^{-1}X^{\mathrm{t}}VX & n^{-1}a(\phi)^{-1}X^{\mathrm{t}}VZ \\ 0 & n^{-1}a(\phi)^{-1}Z^{\mathrm{t}}VX & n^{-1}a(\phi)^{-1}Z^{\mathrm{t}}VZ \end{pmatrix},$$

where $J_{n,\text{scale}}$ is the required $-n^{-1}\sum_{i=1}^{n}\mathrm{E}\partial^2 \log f(Y_i;\theta_i,\phi)/\partial\phi^2$. This implies that the $q\times q$ lower right hand corner of the inverse information matrix remains as in (3). For the normal distribution with $\phi=\sigma$ and $a(\sigma)=\sigma^2$, for example, one finds $J_{n,\text{scale}}=2/\sigma^2$.

We may now describe the FIC procedure for selecting the tentatively best model for estimating the focus parameter $\mu=\mu(x,z)=x^{\mathrm{t}}\beta+z^{\mathrm{t}}\gamma$. The criterion works specifically for the given (x,z) position in covariate space; if required, the procedure can be repeated for several positions. Different submodels are indexed by the various subsets S of covariate components $1,\ldots,q$, ranging from the empty set to the full list. For each submodel S one can evaluate the estimate

$$\widehat{\mu}_S(x,z)=x^{\mathrm{t}}\widehat{\beta}_S+(\pi_S z)^{\mathrm{t}}\widehat{\gamma}_S=x^{\mathrm{t}}\widehat{\beta}_S+z_S^{\mathrm{t}}\widehat{\gamma}_S \tag{4}$$

that uses maximum likelihood estimates $(\widehat{\beta}_S,\widehat{\gamma}_S)$ in the model that employs all of $\beta_1,\ldots,\beta_p$ but only those γ_j for which $j\in S$. Here π_S is the $|S|\times q$ projection matrix that sends z to $\pi_S z=z_S$, the vector of only those z_j for which $j\in S$, and $|S|$ is the number of components in S. The essence of the FIC is to estimate the mean squared error for each candidate estimator and then to pick the one with lowest possible mean squared error estimate.

We need the vector $\omega=J_{n,10}J_{n,00}^{-1}\frac{\partial\mu}{\partial\beta}-\frac{\partial\mu}{\partial\gamma}$, which here reads

$$\omega=Z^{\mathrm{t}}VX(X^{\mathrm{t}}VX)^{-1}x-z,$$

no matter whether ϕ is known or not. Note the dependence of ω on (x,z). Note furthermore that premultiplying $v_S=\pi_S v$ with π_S^{t} produces a vector $\pi_S^{\mathrm{t}}\pi_S v$ of full length q, with zeroes inserted for those components j with $j\notin S$. Define next $K_{n,S}=(\pi_S K_n^{-1}\pi_S^{\mathrm{t}})^{-1}$, which is the $|S|\times|S|$ lower right block of the inverse of $J_{n,S}$, the information matrix for the S model, and let finally $G_{n,S}=\pi_S^{\mathrm{t}}K_{n,S}\pi_S K_n^{-1}$. The value of the focussed information criterion is now obtained as

$$\mathrm{FIC}(S;x,z)=n\omega^{\mathrm{t}}(I_q-G_{n,S})\widehat{\gamma}_{\text{wide}}\widehat{\gamma}_{\text{wide}}^{\mathrm{t}}(I_q-G_{n,S})^{\mathrm{t}}\omega+2\omega^{\mathrm{t}}\pi_S^{\mathrm{t}}K_{n,S}\pi_S\omega, \tag{5}$$

with $\widehat{\gamma}_{\text{wide}}$ being the maximum likelihood estimator of γ in the largest of the considered models, i.e. the one containing all of $\gamma_1,\ldots,\gamma_q$. For given values of (x,z), the best model according to the FIC is that model, indexed by S, for which $\mathrm{FIC}(S;x,z)$ is the smallest.

The FIC stems from estimating and then adding a squared bias term and a variance term; see Claeskens and Hjort (2003) for further discussion and applications. There is a natural modification of (5) for the case of the squared bias being estimated with a negative number; in such cases we truncate that term to zero. For further discussion of this point, see also relevant comments in Hjort and Claeskens (2006).

2.3 Examples

Here we present a short list of examples that fit in with the general framework above. Each model may be fitted using statistical software packages, and the J_n and K_n matrices are easily computed via the appropriate form of the v_i weights. For each situation one may use (5) to determine the best submodel for estimating $x^{\rm t}\beta + z^{\rm t}\gamma$, or for any smooth function thereof, like $\mathrm{E}(Y \,|\, x, z)$ or the median response $\mathrm{median}(Y \,|\, x, z)$.

1. Consider a non-linear normal regression setup where Y_i is normal (θ_i, σ^2), with $\theta_i = r(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)$ for some specified function $r(\eta)$. Then $b(\theta) = \frac{1}{2}\theta^2$, and the glm weights of (2) become $v_i = r'(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)^2/\sigma^2$. The ordinary linear normal model corresponds to $r(\eta) = \eta$ with weights $v_i = 1/\sigma^2$.

2. Assume the Y_is are Poisson with parameters $\xi_i = \exp(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)$. This is Poisson regression with canonical link function. Then $b(\theta) = \exp(\theta)$, and $v_i = \exp(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)$.

3. Then let Y_i be binomial (m_i, p_i), with $p_i = H(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)$, for a suitable distribution function H. This is again a generalised linear model with $b(\theta_i) = m_i \log\{1 + \exp(\theta_i)\}$, and one finds

$$v_i = m_i H'(\eta_i)^2/[H(\eta_i)\{1 - H(\eta_i)\}],$$

with $\eta_i = x_i^{\rm t}\beta + z_i^{\rm t}\gamma$. This can be used for probit regression, for example, where $p_i = \Phi(\eta_i)$ with the cumulative standard normal. For logistic regression matters simplify to $v_i = m_i p_i(1 - p_i)$.

4. Suppose positive observations Y_i are modelled with Gamma distributions (c, d_i), where c is fixed but $d_i = \exp(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)$. We use the parametrisation where the mean of Y_i is $\xi_i = c/d_i$. Here one finds $v_i = c$, simply. Thus the $J_{n,\rm wide}$ matrix is proportional to the sample variance matrix of the covariate vectors.

Suppose on the other hand that Y_i is taken to be Gamma with parameters (c_i, d), this time with d fixed and flexible $c_i = \exp(\eta_i)$, with again $\eta_i = x_i^{\rm t}\beta + z_i^{\rm t}\gamma$. This actually corresponds to a generalised linear model in terms of the $\log Y_i$, and one finds $v_i = \exp(2\eta_i)\psi'(\exp(\eta_i))$, where ψ' is the trigamma function, the derivative of $\psi = \Gamma'/\Gamma$.

3 The weighted FIC for generalised linear models

Due to the dependence of the focus on the covariate values (x, z), also the FIC takes different values, and will produce different rankings of candidate models, for different locations in the covariate space. Sometimes we choose an average or median value for the regression variables (x, z), to represent some 'average' or 'median' subject, sometimes inside a stratum. Often, such a detailed focus point is not wanted, one might rather wish to find a good model which works well over a major part of, or over all of, the covariate space. We wish to select a good

model valid for a range of individuals simultaneously, say for a subgroup of the population.

We continue to use the generalised linear model setting of the previous section. To reach precise results and concise arguments for our weighted-focussed model selection schemes we shall employ a local misspecification framework where the true parameter is of the form $(\beta,\gamma)=(\beta_0,\delta/\sqrt{n})$. Here $\delta=(\delta_1,\ldots,\delta_q)^{\rm t}$ is fixed and unknown. This is as in Hjort and Claeskens (2003a) and Claeskens and Hjort (2003); see also the rejoinder Hjort and Claeskens (2003b) to the discussion of these papers. The idea is to prove results about model selectors and estimators in terms of the local model misspecification parameter δ, and to use such for developing appropriate model information criteria.

3.1 Some preliminary results

Suppose in general terms that a parameter of interest $\mu(\beta,\gamma;u)$ depends on some quantity u that varies in the population being studied. Under the framework outlined above, the true parameter value is $\mu(\beta_0,\delta/\sqrt{n};u)$. We shall now use results developed in Hjort and Claeskens (2003) and Claeskens and Hjort (2003), with suitable modifications. These rely on certain regularity conditions, detailed in these papers. These conditions are mild, and are not repeated in detail here. One such condition that we need to mention here, in order to adequately identify the appropriate limits below, is the existence of a limit matrix J to which $J_{n,\rm wide}$ converges, with blocks J_{00}, etc. Similarly this defines limit versions K of K_n, and a fortiori J_S and K_S, limit versions of $J_{n,S}$ and $K_{n,S}$. Also, $K_S=(\pi_S K^{-1}\pi_S^{\rm t})^{-1}$, the lower right block $(J_S^{-1})_{11}$ of J_S.

For each u the theory as developed in Claeskens & Hjort (2003) applies to the subset-model-based maximum likelihood estimators $\widehat{\mu}_S(u)$, for which we have

$$\sqrt{n}\{\widehat{\mu}_S(u)-\mu_{\rm true}(u)\}\xrightarrow{d}\Lambda_S(u)=(\tfrac{\partial\mu(\beta,\gamma;u)}{\partial\beta})^{\rm t}J_{00}^{-1}M+\omega(u)^{\rm t}(\delta-\pi_S^{\rm t}K_S\pi_SK^{-1}D),$$

where $\omega(u)=J_{10}J_{00}^{-1}\partial\mu(\beta,\gamma;u)/\partial\beta-\partial\mu(\beta,\gamma;u)/\partial\gamma$, with partial derivatives evaluated at the null point $(\beta_0,0)$. Furthermore, $M\sim{\rm N}_p(0,J_{00})$ and $D\sim{\rm N}_q(\delta,K)$, and these random vectors are independent. The (M,D) variables are needed in Section 5, and furthermore the variables (C_S,D_S) that appear now are linear functions of (M,D).

For a fixed set S and a fixed focus point, Lemma 3.2 of Hjort and Claeskens (2003) applies, and yields

$$\begin{pmatrix}\sqrt{n}(\widehat{\beta}_S-\beta_0)\\ \sqrt{n}\widehat{\gamma}_S\end{pmatrix}\xrightarrow{d}\begin{pmatrix}C_S\\ D_S\end{pmatrix}\sim{\rm N}_{p+|S|}(\xi_S,J_S^{-1}),$$

where

$$\xi_S=\begin{pmatrix}J_{00}^{-1}J_{01}(I_q-\pi_S^{\rm t}K_S\pi_SK^{-1})\delta\\ K_S\pi_SK^{-1}\delta\end{pmatrix}.$$

A special case of importance is that of S being the full set $\{1,\ldots,q\}$, for which

$$\widehat{\delta}_{\rm wide} = \sqrt{n}\widehat{\gamma}_{\rm wide} \xrightarrow{d} D \sim {\rm N}_q(\delta, K). \tag{6}$$

For convenience of notation we adopt the notion that estimators without subset-subscript correspond to the full model, with all $\gamma_1,\ldots,\gamma_q$ parameters; thus $\widehat{\delta} = \widehat{\delta}_{\rm wide}$, etc.

The S-indexed model works with estimates $\widehat{\gamma}_S$ for γ_j with $j \in S$ but uses 0 for $j \notin S$. To deal efficiently with different functions of these it will be convenient to introduce the extended projection matrix of dimension $(p+|S|)\times(p+q)$,

$$\widetilde{\pi}_S = \begin{pmatrix} I_p & 0_{p,q} \\ 0_{|S|,p} & \pi_S \end{pmatrix}.$$

Allow also the introduction of the $q\times q$ matrix $G_S = \pi_S^{\rm t} K_S \pi_S K^{-1}$. Then

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_S - \beta_0 \\ \pi_S^{\rm t}\widehat{\gamma}_S - \delta/\sqrt{n}\end{pmatrix} \xrightarrow{d} \begin{pmatrix} C_S \\ \pi_S^{\rm t} D_S - \delta\end{pmatrix},$$

which is seen to have mean vector and variance matrix

$$F_S\delta = \begin{pmatrix} J_{00}^{-1}J_{01}(I_q - G_S) \\ -(I_q - G_S)\end{pmatrix}\delta \quad \text{and} \quad \Gamma_S = \widetilde{\pi}_S^{\rm t} J_S^{-1}\widetilde{\pi}_S.$$

3.2 The wFIC for GLM

Consider again the linear predictor $\mu(x,z) = x^{\rm t}\beta + z^{\rm t}\gamma$, and the collection of submodel-based estimators (4). The vector $\widehat{\beta}_S$ is always of length p, but takes on different values for different index sets S. In contrast, the estimator $\widehat{\gamma}_S$ has length $|S|$. The construction of the weighted FIC proceeds as follows. We start with the weighted average quadratic loss function on the scale of the linear predictor, of the form

$$\begin{aligned} L_n(S) &= \sum_{i=1}^n w(x_i,z_i)\{\widehat{\mu}_S(x_i,z_i) - \mu_{\rm true}(x_i,z_i)\}^2 \\ &= \sum_{i=1}^n w(x_i,z_i)(x_i^{\rm t}\widehat{\beta}_S + z_{i,S}^{\rm t}\widehat{\gamma}_S - x_i^{\rm t}\beta_0 - z_i^{\rm t}\delta/\sqrt{n})^2. \end{aligned} \tag{7}$$

The weights $w(x_i,z_i)$ are user-specified and in general different from the weights $v(x_i,z_i)$ in the glm weight matrix V. We shall show that this random loss has a limit distribution, under mild regularity conditions. Let

$$\Omega_{n,w} = \frac{1}{n}\sum_{i=1}^n w(x_i,z_i)\begin{pmatrix} x_i \\ z_i\end{pmatrix}\begin{pmatrix} x_i \\ z_i\end{pmatrix}^{\rm t}, \tag{8}$$

and assume this matrix converges in probability to a nonnegative definite Ω_w, depending of course on the weight function $w(x,z)$.

To properly study the random loss function it is convenient to express it as

$$L_n(S) = n \begin{pmatrix} \widehat{\beta}_S - \beta_{\rm true} \\ \pi_S^{\rm t}\widehat{\gamma}_S - \gamma_{\rm true} \end{pmatrix}^{\rm t} \Omega_{n,w} \begin{pmatrix} \widehat{\beta}_S - \beta_{\rm true} \\ \pi_S^{\rm t}\widehat{\gamma}_S - \gamma_{\rm true} \end{pmatrix}.$$

From conditions and results noted above,

$$L_n(S) \xrightarrow{d} L(S) = \begin{pmatrix} C_S \\ \pi_S^{\rm t} D_S - \delta \end{pmatrix}^{\rm t} \Omega_w \begin{pmatrix} C_S \\ \pi_S^{\rm t} D_S - \delta \end{pmatrix}.$$

Under mild conditions, the expected loss $w\text{-risk}_n(S) = \mathrm{E}\, L_n(S)$ will converge to $w\text{-risk}(S) = \mathrm{E}\, L(S)$. The limit loss $L(S)$ is a quadratic form in normal variables, and has mean value

$$\begin{aligned} w\text{-risk}(S) = \mathrm{E}\, L(S) &= \delta^{\rm t} F_S^{\rm t} \Omega_w F_S \delta + \mathrm{trace}(\Omega_w \Gamma_S) \\ &= \mathrm{trace}(\Omega_w F_S \delta\delta^{\rm t} F_S^{\rm t}) + \mathrm{trace}(\Omega_w \Gamma_S) \\ &= \mathrm{I}(S) + \mathrm{II}(S), \end{aligned}$$

say, involving matrices F_S and Γ_S defined above. We see that the $\mathrm{I}(S)$ term corresponds to weighted squared bias whereas the second term $\mathrm{II}(S)$ is related to weighted variance. By earlier efforts, this second term can be expressed as $\mathrm{II}(S) = \mathrm{trace}(\Omega_w \widetilde{\pi}_S^{\rm t} J_S^{-1} \widetilde{\pi}_S)$.

This leads upon estimating unknown quantities to a weighted-focussed information criterion. The second term is not problematic, and we use

$$\widehat{\mathrm{II}}(S) = \mathrm{trace}(\Omega_{n,w} \widetilde{\pi}_S^{\rm t} \widehat{J}_{n,S}^{-1} \widetilde{\pi}_S),$$

where $\widehat{J}_{n,S}$ is the appropriate sub-matrix of

$$\widehat{J}_n = \frac{1}{n}\frac{1}{a(\widehat{\phi})} \sum_{i=1}^n \widehat{v}_i \begin{pmatrix} x_i \\ z_i \end{pmatrix} \begin{pmatrix} x_i \\ z_i \end{pmatrix}^{\rm t} = \frac{1}{n}\frac{1}{a(\widehat{\phi})} \begin{pmatrix} X^{\rm t}\widehat{V}X & X^{\rm t}\widehat{V}Z \\ Z^{\rm t}\widehat{V}X & Z^{\rm t}\widehat{V}Z \end{pmatrix},$$

and $\widehat{v}_i$ is the estimated version of (2), inserting $x_i^{\rm t}\widehat{\beta}_{\rm wide} + z_i^{\rm t}\widehat{\gamma}_{\rm wide}$ for the linear predictor $x_i^{\rm t}\beta + z_i^{\rm t}\gamma$. See the examples in Section 2.3. For the first term, we note that

$$\widehat{\delta}\widehat{\delta}^{\rm t} = n\widehat{\gamma}\widehat{\gamma}^{\rm t} \xrightarrow{d} DD^{\rm t},$$

a variable with mean $\delta\delta^{\rm t} + K$; recall the convention noted around (6) that $\widehat{\gamma}$ means $\widehat{\gamma}_{\rm wide}$, etc. We therefore use

$$\begin{aligned} \widehat{\mathrm{I}}(S) &= \mathrm{trace}\{\Omega_{n,w}\widehat{F}_S(\widehat{\delta}\widehat{\delta}^{\rm t} - \widehat{K}_n)\widehat{F}_S^{\rm t}\} \text{ if this is positive,} \\ &= 0 \text{ if otherwise.} \end{aligned}$$

The *weighted focussed information criterion* consists in evaluating

$$w\mathrm{FIC}(S) = \widehat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S) \tag{9}$$

for each candidate model S, and in the end selecting the model with smallest value of this estimated average risk. Note that all components of the above expressions are easily obtained via standard output of any software fitting generalised linear models.

Remark 1. When more components are put into S, the I(S) term becomes smaller; for S equal to the full $\{1,\ldots,q\}$ we find $F_S = 0$, making $I(\text{wide}) = 0$. On the other hand, with more components in S, the bigger is the variance II(S). Thus the wFIC method reflects the squared modelling bias against variance balance.

Remark 2. In linear models the expression for the pointwise (unweighted) FIC has been shown to be exact, see Claeskens and Hjort (2003), Section 5.5. There it is shown that, without assuming normality and when using least squares estimators, the exact expression for the mean squared error matches that obtained by FIC. This is a favourable property which helps appreciating the obtained expression of the FIC. More precisely, in a linear model with $Y_i = x_i^{\rm t}\beta + z_i^{\rm t}\gamma + \varepsilon_i$, where $\mu = x^{\rm t}\beta + z^{\rm t}\gamma$ is to be estimated, we use least squares estimators in a submodel indexed by S, leading to the estimator $\widehat{\mu}_S = x^{\rm t}\widehat{\beta}_S + (\pi_S z)^{\rm t}\widehat{\gamma}_S$. Computing the exact bias and variance of this estimator leads to the following expression for the mean squared error of $\widehat{\mu}_S$:

$$n^{-1}(x^{\rm t}J_{n,00}^{-1}x + \omega^{\rm t}\pi_S^{\rm t}K_{n,S}\pi_S\omega) + \omega^{\rm t}(I_q - \pi_S^{\rm t}K_{n,S}\pi_S K_n^{-1})\gamma\gamma^{\rm t}(I_q - K_n^{-1}\pi_S^{\rm t}K_{n,S}\pi_S)\omega. \quad (10)$$

This is, up to a constant not depending on the subset S, identical to the limiting expression on which the FIC is based, compare with (5).

The same property holds when a (non-random) weight function w is included in the loss function L_n in (8). For linear models, computing the exact mean squared error of $L_n(S)$, with least squares estimators inserted for $\widehat{\beta}_S$ and $\widehat{\gamma}_S$, leads to an expression which is equal to that one which wFIC as in (9) is based upon. Note that the error variance σ^2 appears in the denominator of the matrix J_n as the scaling factor $a(\sigma)$. There is accordingly a σ^2 factor implicitly featuring in the first two terms of (10), but not in the final two terms.

4 GLM weights and the AIC

The wFIC method developed above provides a quite general and versatile model selection scheme, in that the weights $w_i = w(x_i, z_i)$ are fully user-specified, meant to reflect what aspects are deemed more important than others for the use of the finally selected model. The only caveat is that the weights should not be seriously unstable; the mathematical requirement for our asymptotics to go through is that the $\Omega_{n,w}$ matrix converges in probability with increasing sample size. This section discusses various types of weights.

4.1 GLM weights

The log-likelihood structure of a generalised linear model itself suggests a natural type of weights for use in the wFIC method. If one chooses $w(x_i, z_i) = v_i/a(\phi)$, then $\Omega_{n,w} = J_{n,\text{wide}}$. This leads to simplifications in the w-risk(S) and wFIC(S) expressions, as we shall see now. For simplicity of notation and presentation we work directly in the limit experiment, where J and K etc. replace $\widehat{J}_n$ and $\widehat{K}_n$ etc.; we also write $G_S = \pi_S^{\mathrm{t}} K_S \pi_S K^{-1}$. First look at $\mathrm{I}(S) = \delta^{\mathrm{t}} F_S^{\mathrm{t}} J F_S \delta$, where some manipulations give

$$JF_S = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \begin{pmatrix} J_{00}^{-1} J_{01}(I_q - G_S) \\ -I_q + G_S \end{pmatrix} = \begin{pmatrix} 0 \\ -K^{-1}(I_q - G_S) \end{pmatrix}.$$

This leads with some further efforts to

$$\mathrm{I}(S) = \delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}} K^{-1}(I_q - G_S)\delta = \delta^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})\delta,$$

which is estimated by

$$\begin{aligned} \widehat{\mathrm{I}}(S) &= \mathrm{trace}\{(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})(DD^{\mathrm{t}} - K)\} \\ &= D^{\mathrm{t}} K^{-1} D - D^{\mathrm{t}} K^{-1} \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} D - q + |S|, \end{aligned}$$

as long as this expression is positive; it is otherwise truncated to zero. Next,

$$\mathrm{II}(S) = \mathrm{trace}(J\widetilde{\pi}_S^{\mathrm{t}} J_S^{-1} \widetilde{\pi}_S) = p + |S|.$$

This is easily obtained since the effect of pre- and post-multiplication by $\widetilde{\pi}_S$ is that only that submatrix of J^{-1} is kept for which the row (and column) numbers belong to the index set S, entries on all other rows and columns being replaced by zero. This implies that J is multiplied by part of its inverse matrix, leading to the simplification $p + |S|$. This is also true for the data-based version $\widehat{\mathrm{II}}(S)$.

To summarise this, with glm weights the limit risk takes the form

$$w\text{-risk}(S) = \delta^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})\delta + p + |S|,$$

and the canonical risk estimate (for the limit experiment) is

$$\widehat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S) = \begin{cases} D^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})D + 2|S| + p - q & \text{if } N(S) \text{ takes place,} \\ p + |S| & \text{otherwise.} \end{cases} \tag{11}$$

Here $N(S)$ is the event that the trace in $\widehat{\mathrm{I}}(S)$ is positive, i.e. that

$$D^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})D > q - |S|.$$

The $N(S)$ takes place with high probability if δ is some distance away from zero, but in situations where the underlying γ vector is close to zero, i.e. the narrow model is close to being correct, the probability that $N(S)$ does not take place is significant. The finite-sample version of the risk estimate uses $\widehat{\delta}$ for D and $\widehat{K}_n$ and $\widehat{K}_{n,S}$ for K and K_S.

4.2 The wFIC and the AIC

Now consider Akaike's information criterion, which in the present circumstances takes the form

$$\mathrm{AIC}(S) = 2\sum_{i=1}^{n} \log f(y_i; \widehat{\theta}_i, \widehat{\phi}) - 2(p + |S|),$$

with $\widehat{\theta}_i$ being the appropriate maximum likelihood estimate of $\theta_i = \theta(x_i, z_i)$, which again depends on $\widehat{\beta}_S$ and $\widehat{\gamma}_S$. The AIC scores may be computed for each submodel S, down to that of the most narrow model which corresponds to $S = \emptyset$ and which uses only $\beta_1, \ldots, \beta_p, \phi$ as model parameters. When subtracting the smallest model's AIC value from $\mathrm{AIC}(S)$, and performing a one-step Taylor expansion, we find that, for the limiting situation where $n \to \infty$,

$$\mathrm{AIC}(S) - \mathrm{AIC}(\emptyset) \xrightarrow{d} D^{\mathrm{t}} K^{-1} \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} D - 2|S|.$$

See Claeskens and Hjort (2003, eq. (2.5)). The best models have the highest AIC scores.

We see from this and (11) that the wFIC method, when using glm weights, is essentially large-sample equivalent to the AIC method. The word 'essentially' relates to the modification for truncating an estimate of a squared bias to zero, when relevant, spelled out in (11). Thus the wFIC provides a fresh perspective on the AIC, and our arguments even suggest a correction to the AIC scores in cases where the event $N(S)$ does not take place.

One may go through the list of examples in Section 2.3 to see the appropriate random loss functions that correspond to the AIC. For the logistic regression setup of Example 3 in that section, for example, model selection by estimating the mean of

$$L_n(S) = \sum_{i=1}^{n} m_i p(x_i, z_i)\{1 - p(z_i, z_i)\}\Big(\log \frac{\widehat{p}_{i,S}}{1 - \widehat{p}_{i,S}} - \log \frac{\widehat{p}_i}{1 - \widehat{p}_i}\Big)^2$$

is essentially the same as AIC. Here $\widehat{p}_{i,S}$ is the estimated probability under model S. Similarly, for Poisson regression, basing model selection on estimating

$$w\text{-risk}_n(S) = \mathrm{E} \sum_{i=1}^{n} \exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)(\log \widehat{\lambda}_{i,S} - \log \lambda_i)^2$$

will be large-sample equivalent to the AIC scheme, where $\widehat{\lambda}_{i,S}$ is the estimate of Poisson rate i inside the S model. Of course other weights and other transformations can be worked with, for logistic and Poisson regression, and such alternatives can in the perspective developed here be seen as cousins to the AIC method.

4.3 Other types of weights

In addition to the perhaps canonical glm weights choice discussed above, the following types of weights may be considered.

Uniform weights. A simple choice of weights could just assign mass $1/n$ to each contribution in the summand, with Ω_n matrix simply equal to the empirical variance matrix of the $p+q$-sized covariate vectors. Another choice is to let $w_i = 1$ precisely for individuals i belonging to a stratum of interest. This is exemplified in Section 6.

Gliding covariate window. A version of the above is to use $w_i = w(x_i, z_i; x_0)$ equal to 1 for individuals inside a certain neighbourhood of some fixed x_0. This gives a ranking of candidate models around each fixed x_0. Smoothed versions can also be contemplated, also in a context of model assessment where the x_0 is moved in its covariate space. A good final model, then, should rank highly in all these local competitions.

Averaging over z for fixed x. Another version with some appeal is to assess models in terms of how well they perform for a given covariate x, averaged across all likely z values. To indicate in one particular fashion how this may be handled, let

$$w_i \propto f(z_i - \xi_x, \Sigma_x),$$

in terms of the density f of the multinormal density $\mathrm{N}_q(0, \Sigma_x)$, where ξ_x and Σ_x are the estimated mean and variance matrix for z given x.

Robust weights. There is a whole area of research addressing issues related to robust model choice. Robustness is concerned with the downweighting (and sometimes identification) of data vectors that are 'outliers' or that may exert too strong an influence on maximum likelihood estimators. The methodology developed in this paper sticks to the maximum likelihood estimators, as such, but the weight function of the wFIC allows downweighting schemes that make model selection less dependent on extreme data vectors. One version of this would be

$$w(x_i, z_i) = h(\|(x_i, z_i) - (x_0, z_0)\|) \quad \text{for } i = 1, \ldots, n,$$

where the norm in question could measure a suitable distance from covariate vectors to some robustly identified centre location (x_0, z_0), and where $h(u)$could be taken as 1 over a broad interval, but made to go to zero beyond that interval. Another choice is $h(u) = \exp(-c|u|)$ for a perhaps small value of c. Such a scheme assures robustness with respect to extreme or overly influential covariate vectors.

Note before we come to the next point that the wFIC methodology also works with weights more general than $w_i = w(x_i, z_i)$, as long as the $\Omega_{n,w}$ matrix of (8) converges in probability. Thus weights of the form $w_i = h(\mathrm{res}_i)$, functions of suitably defined glm residuals, are allowed. One such version, among several, is

$$w_i = h(\mathrm{res}_i) = \begin{cases} 1 & \text{if } |\mathrm{res}_i| \le c, \\ c/|\mathrm{res}_i| & \text{if } |\mathrm{res}_i| > c, \end{cases}$$

where $\text{res}_i = (Y_i - \widehat{\xi}_i)/\widehat{\sigma}_i$, featuring (perhaps robustified) estimates of mean and standard deviation for Y_i. Weights of this type are dicussed, in a rather different context of robust linear regression, by Ronchetti and Staudte (1994), who also find that $c = 1.345$ is a reasonable default value for this weight function.

5 General risk averages

Above we developed a wFIC method for estimating naturally weighted averaged risks in generalised linear models. Sometimes different risk averages are called for, however, and this section extends the wFIC to handle such cases. We choose to stay inside the glm framework exposited at the start of Section 3, although more general situations can be considered. A brief motivating example is as follows: Suppose Y_i data are exponentially distributed with parameters $\theta_i = \exp(x_i^{\rm t}\beta + z_i^{\rm t}\gamma)$, and that one wishes to estimate the quantile distribution

$$\mu(u) = \mu(u \,|\, x, z) = \{-\log(1-u)\}/\exp(x_i^{\rm t}\beta + z_i^{\rm t}\gamma) \quad \text{for } u \in (0,1).$$

How can we select a model that provides good estimates $\widehat{\mu}_S(u)$ across all deciles $u = 0.1, \ldots, 0.9$, say?

Assume in general terms that a parameter $\mu(u)$ is to be estimated, defined in terms of the parameters (β, γ) of a generalised linear model, and depending on some parameter u that may or may not depend on the covariates. As in previous sections limit distributions will be established in the framework where $(\beta, \gamma) = (\beta_0, \delta/\sqrt{n})$, aiming at providing adequate finite-sample approximations for risks and averaged risks. We have

$$\sqrt{n}\{\widehat{\mu}_S(u) - \mu_{\rm true}(u)\} \xrightarrow{d} \Lambda_S(u) = (\tfrac{\partial \mu(u)}{\partial \beta})^{\rm t} J_{00}^{-1} M + \omega(u)^{\rm t}(\delta - G_S D),$$

where $\omega(u) = J_{10}J_{00}^{-1}\partial\mu/\partial\beta - \partial\mu/\partial\gamma$ and $G_S = \pi_S^{\rm t} K_S \pi_S K^{-1}$; also, M and D are independent, and $M \sim {\rm N}_p(0, J_{00}^{-1})$ and $D \sim {\rm N}_q(\delta, K)$. Consider the loss average function

$$L_n(S) = n \int \{\widehat{\mu}_S(u) - \mu_{\rm true}(u)\}^2 \, {\rm d}W_n(u),$$

where W_n represents some relevant distribution of u values, like the deciles in the quantile example above. Assuming W_n converging to a suitable weight distribution W (or that it simply stays fixed, independent of sample size), we have

$$L_n(S) \xrightarrow{d} L(S) = \int \Lambda_S(u)^2 \, {\rm d}W(u)$$

under mild conditions. We measure the total averaged risk via the expected value of L_n, which converges to

$$w\text{-risk}(S) = {\rm E}\, L(S) = \int {\rm E}\Lambda_S(u)^2 \, {\rm d}W(u),$$

again under mild conditions. Here

$$\begin{aligned}\mathrm{E}\Lambda_S(u)^2 &= \tau_0(u)^2 + \omega(u)^{\mathrm{t}}\mathrm{E}(\delta - G_S D)(\delta - G_S D)^{\mathrm{t}}\omega(u)\\ &= \tau_0(u)^2 + \omega(u)^{\mathrm{t}}\{(I_q - G_S)\delta\delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}} + K_S^{-1}\}\omega(u)\end{aligned}$$

in terms of $\tau_0(u)^2 = (\frac{\partial\mu(u)}{\partial\beta})^{\mathrm{t}} J_{00}^{-1}\frac{\partial\mu(u)}{\partial\beta}$ and the variance matrix

$$\operatorname{Var} G_S D = G_S K G_S^{\mathrm{t}} = \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} \pi_S^{\mathrm{t}} K_S \pi_S = \pi_S^{\mathrm{t}} K_S \pi_S.$$

This leads to the expression

$$w\text{-risk}(S) = \int \tau_0(u)^2\,\mathrm{d}W(u) + \mathrm{trace}\{(I_q - G_S)\delta\delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}}R\} + \mathrm{trace}(\pi_S^{\mathrm{t}} K_S \pi_S R) \quad (12)$$

for the limit risk, where

$$R = \int \omega(u)\omega(u)^{\mathrm{t}}\,\mathrm{d}W(u).$$

The first term is immaterial since it does not depend on S, so our generalised wFIC naturally becomes

$$\begin{aligned}w\mathrm{FIC}(S) &= \widehat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S)\\ &= \max[\mathrm{trace}\{(I_q - \widehat{G}_{n,S})(\widehat{\delta}\widehat{\delta}^{\mathrm{t}} - \widehat{K}_n)(I_q - \widehat{G}_{n,S})^{\mathrm{t}}\widehat{R}\}, 0] + \mathrm{trace}(\pi_S^{\mathrm{t}}\widehat{K}_{n,S}\pi_S\widehat{R}).\end{aligned}$$

Here $\widehat{R}$ is a sample-based estimate of the R matrix.

This more general wFIC can be applied when one wishes to consider average risk across both covariates and quantiles, for example.

6 Illustrations and applications

6.1 Diabetic retinopathy data

The Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984) provides information to study diabetic retinopathy as a function of several other measurements. The dataset consists of patient information for 348 men and 343 women. The binary outcome variable $Y = 0$ indicates whether there is no or only mild nonproliferate retinopathy on both of the eyes. A value $Y = 1$ is obtained when there is moderate to severe nonproliferate retinopathy, or proliferate retinopathy for at least one of the eyes. Variables measured are: duration of diabetes in years (x_1), presence of macular edema in at least one eye (z_1), percentage of glycosylated hemoglobin (z_2), body mass index (z_3), pulse rate in beats per 30 seconds (z_4), sex (z_5, with 1 for male and 0 for female), presence of urine protein (z_6), and area of residence (urban or rural, z_7).

A logistic regression model is used for the analysis. Since in earlier analysis of this dataset it is found that duration of diabetes is an important variable (see for example

Table 1: Values of the weighted focussed information criterion, where weights are indicator values for men (1st column) and for women (3rd column), together with the selected variables. All logistic regression models contain an intercept term, as well as the variable x_1, duration of diabetes.

Men		Women	
$w\text{FIC}_M$	z-variables	$w\text{FIC}_F$	z-variables
21.201	1,4,6	24.828	1,4,6
23.270	1,3,4,6	24.974	1,4,6,7
24.230	1,2,4,6	26.155	1,2,4,6
24.393	1,4,6,7	26.160	1,3,4,6
27.590	1,2,3,4,6	28.072	1,2,4,6,7
27.767	1,3,4,6,7	28.616	1,3,4,6,7
28.680	1,2,4,6,7	28.673	1,2,3,4,6
29.982	1,4,5,6	31.096	1,4,5,6
31.186	1,3,4,5,6	31.636	1,4,5,6,7
31.345	1,2,4,5,6	32.945	1,3,4,5,6

Claeskens, Croux and Van Kerckhoven, 2006), we include this in all of the models we consider, as well as an intercept term. We perform model selection amongst the other seven variables and allow for all possible subsets of the full model, leading to 128 possible models. As a model selection criterion we take first $w\text{FIC}_M(S)$ with weight vector $(1/n_M)I(\text{male})$ and next $w\text{FIC}_F(S)$ with weight vector $(1/n_F)I(\text{female})$ where the weights are indicator variables for men (in case 1) and for women (in case 2), and n_M (resp. n_F) denotes the number of men (resp. women) in the dataset. Note that the values of $w\text{FIC}(S)$are computed using the complete dataset, we are not splitting the dataset for model selection.

Table 6.1 gives for both criteria the ten smallest values, together with the variables in the corresponding models. The best model is in both cases the model containing the binary variables z_1: presence of macular edema in at least one of the eyes and z_6: indicator for urine protein, as well as z_4: pulse rate. The subgroups differ in the ranking of the next best models, for men the body mass index (variable z_3) is an important variable, while for women the area of residence (z_7) is more important, and z_3 only shows in the 4th best ranked model. Thus we learn that body mass index influences Y in possibly different ways, for men and for women, which may be taken into account for building a final model.

As a comparison we also used the overall model selection criteria AIC and BIC. The AIC picks the same model as the FIC does, namely the model with variables z_1, z_4 and z_6, while the BIC omits from this model the variable z_4, pulse rate in beats per second.

These criteria are computed using the complete dataset. When we would split the data into the results for men and for women separately and then run the AIC selection method separately for both data parts, AIC selects the variables z_1, z_2, z_6 for the subset of men, and variables z_1, z_4, and z_6 for the subset of women. No artificial data splitting is needed for the computation of the wFIC.

6.2 CH_4 concentrations

This example consists of CH_4 data, which are atmospheric CH_4 concentrations (ppbv) derived from flask samples collected at the Shetland Islands of Scotland (Steele, Krummel and Langenfelds, 2002). Monthly values are expressed in parts per billion by volume (ppbv). In total there are 110 monthly measurements, starting in December 1992 and ending December 2001. The regression variable u = time is rescaled to the $(0, 1)$ interval, and the response variable Y is the CH_4 concentration. We use a cosine series estimator based on the model

$$\mu(u) = \mathrm{E}(Y \,|\, U = u) = \beta_0 + \sum_{j=1}^{m} \gamma_j \cos(\pi j u),$$

where we will vary the value of m, which is the truncation point of the series. This defines a sequence of nested models, which fits into the regression context of the previous sections when defining $z_j = \cos(\pi j u)$. We wish to select the best order m. In our modelling efforts, we let m be any number between 1 and 15 (the wide model). A scatter plot of the data is shown in Figure 1(a). We applied the wFIC(S) method (9) with equal weights $w_i = 1$, and found that the best model is for $m = 2$; see the figure.

As a comparison we also computed FIC values for each of the individual 110 measurement months. This means that we take as a focus parameter the mean CH_4 concentration at that particular month (without averaging), which leads to a set of 15 FIC values, one for each order of $m = 1, \ldots, 15$, and this for each of the 110 months. The results are summarised in the following frequency table for the individually chosen model by FIC. For example, model order $m = 1$ was chosen 75 times in the 110 model selection applications, model order $m = 2$ was chosen 7 times, etc.

m	1	2	3	5	6	7	10	11
frequency	75	7	1	1	3	5	3	15

The overall chosen model with $m = 2$ is in this case not the model which was most frequently selected by the individual searches. Remarkably, the model with order 11 is chosen 15 times in the individual search. A possible explanation for this is that a high frequency model of order 11 is reflecting the random variability in the data cloud.

Another set of weights which makes sense for variables measured in time, is that which gives more weight to more recent measurements. As an example we used the weighting

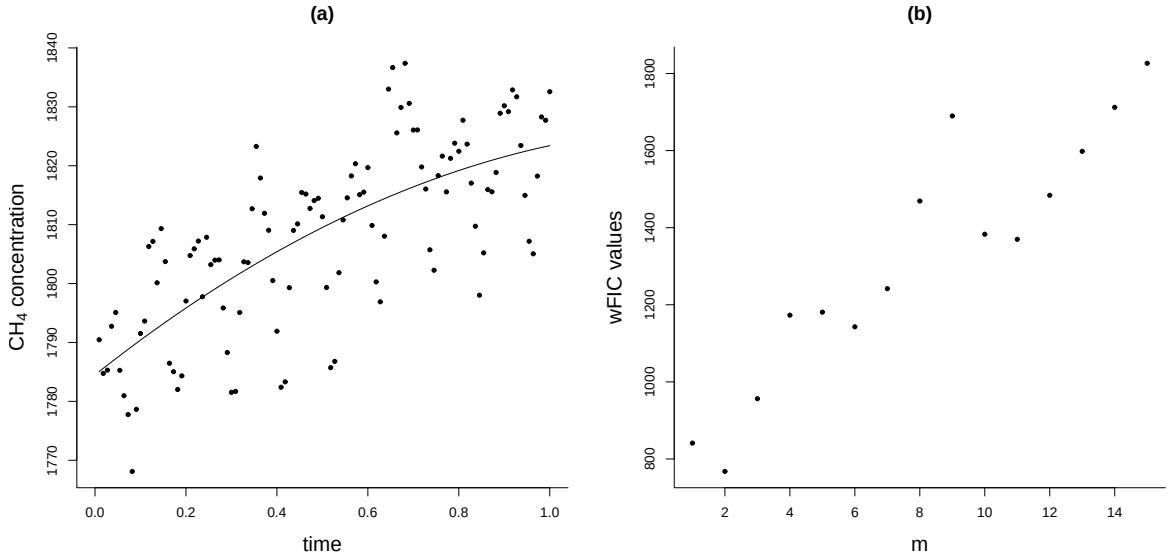

Figure 1: *(a) Scatterplot of the CH4 data, along with the estimated mean curve for the* $m = 2$ *model. (b) The* $w\mathrm{FIC}(S)$ *values with equal weights* $w_i = 1$.

scheme i/n ($i = 1, \ldots, n = 110$), and found in this particular case the same FIC selected model, namely the model with truncation point $m = 2$.

6.3 Highway data

To illustrate robust downweighting, we use Hoffstedt's highway data, see also Weisberg (2005, Section 7.2). This dataset is used to explain the 1973 accident rate per million vehicle miles, as a function of several variables. There are 39 observations made. In every model we include an intercept term and x_1, the length of the highway segment in miles. Variables to choose from are average daily traffic count in thousands (z_1), truck volume as a percent of the total volume (z_2), total number of lanes of traffic (z_3), number of access points per mile (z_4), number of signalised interchanges per mile (z_5), number of freeway-type interchanges per mile (z_6), speed limit in 1973 (z_7), lane width, in feet (z_8), width of the outer shoulder on the roadway (in feet) (z_9), and finally an indicator of the type of roadway or the source of funding for the road (z_{10}).

Based on robust C_p model selection, Ronchetti and Staudte (1994) support the model which includes, in addition to x_1, the variables z_5, z_6, z_7 and z_{10}, and also the model with additional variables z_2, z_3, z_4 and z_9. Here we construct weights $w_1, \ldots, w_{39}$ based on the robustification method outlined in Section 4.3; these rely on intially used robust estimators for the regression coefficients obtained in the full model. Five of the observations receive a weight which is smaller than one, all other observations get weight one. The weights that

differ from one are 0.903, 0.568, 0.436, 0.577, 0.811. These weights are then used in the wFIC(S) construction, and lead to preferring the model with the two variables x_1 and z_7. For this particular example, the model selected by FIC is more parsimonious than the model suggested by Mallow's C_p.

7 Concluding comments

Our paper has provided a fair middle ground between the extremes 'blind model selection' (where a model is found via say AIC or BIC, without any particular regard to the actual use of the model after selection) and 'fully focussed selection' (where a model is selected to perform optimally for a given estimand). We have seen that special versions of the wFIC correspond to the AIC for generalised linear models, so our methods may be seen as suitable generalisations of the AIC for use in situations where the context of one's modelling and analysis dictates more specific weighting schemes than the default ones. Below we give some concluding remarks, pertaining to themes related to but outside the main scope of the present paper.

1. In this paper we chose to concentrate on the generalised linear models framework, where methods have a particularly clear structure, but it is clear that the methods can be generalised to many other regression structures. Thus parametric regression models for multidimensional data, and for hazard rates with censored data, can be handled with essentially the same methods. Extensions to the semiparametric Cox model are less immediate, but can be accomplished via methods of Hjort and Claeskens (2006).

2. Our paper has developed strategies that for each model rely on maximum likelihood (or asymptotically equivalent) estimators. There is a need for generalising methods and results to more robust strategies, for example involving M-estimators. This is entirely possible, but requires more work and will result in algebraically and structurally somewhat less elegant methods.

3. Our $w\text{FIC}(S) = \widehat{\text{I}}(S) + \widehat{\text{II}}(S)$ is not algebraically equivalent to the simpler one of averaging individual FIC(S) scores. The two methods are the same only in cases where there are no modifications of setting negative estimates of squared bias to zero. The wFIC method, as outlined in Sections 3 and 5, performs the squared bias modification only once, at the end, as opposed to performing this operation for each individual application.

4. Our methods stem from precise limit distribution results that involve quantities like J and K, along with further relatives like K_S and G_S. Our wFIC(S) formulae involve estimates of these quantities, say $\widehat{J}_{n,S}$, $\widehat{K}_{n,S}$, and so on. The theory behind the methods ensure that they work well as long as estimates are used that are consistent, under the local misspecification framework $(\beta, \gamma) = (\beta_0, \delta/\sqrt{n})$. Among various possibilities we have chosen

to use the 'wide model perspective' in our implementations, starting from estimators $\widehat{J}_n$ that use estimates $(\widehat{\beta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$.

5. We have derived methods for selecting a model, but have not discussed the consequences of having selected the model in this fashion. Methods of Hjort and Claeskens (2003a) and Claeskens and Hjort (2003) make it however possible to analyse the performance of estimator-after-selection, also with the wFIC methods developed in the present article.

6. Though we have been specifically concerned with model selection, methods of Hjort and Claeskens (2003a) can be applied to provide model average procedures, say of the type

$$\widehat{\mu} = \sum_S c(w{\rm FIC}(S))\widehat{\mu}_S,$$

a data-dependent average across the estimators of the individual models, with

$$c(w{\rm FIC}(S)) = \exp\{-\tfrac{1}{2}\kappa\, w{\rm FIC}(S)\}\Big/\sum_{S'}\exp\{-\tfrac{1}{2}\kappa\, w{\rm FIC}(S')\}.$$

Here κ is an algorithmic parameter, with small κ corresponding to near uniform weighting while larger κ means giving nearly full weight to the model that wins the wFIC competition. Methods of that paper also make it possible to study performances of such average estimators, compared to the more usual estimators-post-selection.

7. Our methods have been developed inside a framework of 'first order asymptotics' for general parametric models. It might be important to supplement such methods by suitable second order corrections to make them work more precisely for moderate or smaller sample sizes. The content of Remark 2 of Section 3 is that the wFIC(S) expression is exactly correct for each finite $n > p+q$, when used in the linear model with non-random weights. This is an indication that even the first-order approximations to risks and their estimates are adequate also in other generalised linear models.

The situation is a bit more complicated when the weights themselves have a random component, as with the robust type $w_i = h({\rm res}_i)$ discussed in Section 4.3; here approximations stemming from the first-order asymptotics might need adjustments to be more accurate in practice. In other words, even though demonstrably $L_n(S) \to_d L(S)$ and w-risk$_n(S) \to w$-risk(S), one can expect the real variance of $L_n(S)$ to be bigger than that of $L(S)$, in cases with complicated data-dependent weights.

References

Akaike, H. (1974). A new look at statistical model identification, *I.E.E.E. Transactions on Automatic Control* **19**, 716–723.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics*, to appear.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion]. *Journal of the American Statistical Association* **98**, 900–916.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models.* Chapman and Hall, London.

Hjort, N. L. and Claeskens, G. (2003a). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association* **98**, 879–899.

Hjort, N. L. and Claeskens, G. (2003b). Rejoinder to 'The Focussed Information Criterion' and 'Frequentist model average estimators'. *Journal of the American Statistical Association* **98**, 938–945.

Hjort, N. L. and Claeskens, G. (2006). Focussed information criteria and model averaging for Cox's hazard regression model. *Journal of the American Statistical Association*, to appear.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* **102**, 520–526.

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

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica* **2**, 327–338.

Ronchetti, E. and Staudte, R. G, (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association* **89**, 550–559.

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

Steele, L. P., Krummel, P. B. and Langenfelds, R. L. (2002). Atmospheric CH4 concentrations from sites in the CSIRO Atmospheric Research GASLAB air sampling network (October 2002 version). In *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN, U.S.A.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edition. Wiley, New York.